ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 14, Issue, 09, pp.22195-22198, September, 2022
DOI: https://doi.org/10.24941/ijcr.43527.09.2022

INTERNATIONAL JOURNAL
OF CURRENT RESEARCH

RESEARCH ARTICLE

INTELLIGENTLY SPEECH RECOGNITION AND CONVERSION WITH NOISE REDUCTION

S.M Najrul Howlader[1], Md. Mukter Hossain[1], Romana Perveen[2], Kaniz Fatema[1] and M. Mesbahuddin Sarker[1]

[1]IIT, Jahangirnagar University, Dhaka, Bangladesh
[2]Department of Mathematics, Islamic University, Kushtia, Bangladesh

ARTICLE INFO

Article History:
Received 29th June, 2022
Received in revised form
17th July, 2022
Accepted 14th August, 2022
Published online 28th September, 2022

Key words:

Open and Distance Learning, Inclusive Education.

**Corresponding Author:*
S.M Najrul Howlader

ABSTRACT

Speech to Text conversion is an advanced artificial intelligence system where the machine can recognize the user's voice then, interpreted it into text. The work of the speech recognition process started before the 70s, now, it is more developed and works in a more efficient way, also can handle complex expressions. These systems can help people who want to write convert texts into different languages quickly. Though, existing systems lack more features like multiple languages (9) in a single domain. We have implemented a system that overcomes this problem. Speech recognition and synthesis technology are one of the im-portant and fastest-growing Information technologies. Nowadays, Many applications have many potential benefits in different areas. Among the whole world population, About 25% are suffering from various disabilities; There are many people among them who are blind or cannot able to use their hands perfectly. This system gives them unexpected and evaluated support in this field so that they can write and express information with humans by managing computers, tablets, mobile, or any kind of information through voice input. This project has been planned, designed, and developed using Visual studio with JavaScript and C# programming based on MFCC and HMM.

Citation: ***S.M Najrul Howlader, Md. Mukter Hossain, Romana Perveen, Kaniz Fatema, M. Mesbahuddin Sarker.* 2022.** "Intelligently speech recognition and conversion with noise reduction". *International Journal of Current Research*, 14, (09), 22195-22198.

INTRODUCTION

From the beginning, human needs to interact with each other, and they do it in several ways. Among them, eye contact, facial expression, and gesture speech is the primary interaction part. But language is a big barrier also, we need to keep in mind, handicapped people. In the education field, we saw visually impaired students need external help (Shaw, 2013; Rigoll, 1994) to write in the exam. So, from these motivations, researchers implemented the Speech to Text (STT) (Bijl, 2001) system. In STT (Witt, 1999; Bourlard, 2012), speech can be converted into texts, also the text can be converted (Lee, 1988) into other languages. Researchers have implemented different STT systems (Ballinger, 2011) which can do the basic stuff of STT with some improvements like spell correction, and grammatical correction using different machine learning techniques (Lerner, 1989). Such as decision trees, Support Vector Machine, Hidden Markov Model (Ballinger, 2016) etc. Also, Google has a policy where people can convert their speech to text but tracks the speech for a few seconds, and if you pause for some time it will close the speech tracing (Pinson, 2013). Again, these applications do not have multi-language support. The system is trained with deep neural networks to obtain high accuracy speech to text (Reddy, 1976) conversion for serving the examination and other purposes. It is. The technology we have used to achieve this system is explained below. For removing this limitation, we have developed a multi- language We call it Muskitron. We have implemented it using JavaScript and C# dot Net. Our system consists of all the existing features of an STT but we have added multi-language support (Radha, 2012) which is the contribution of this work. The remaining sections of this paper are organized here. In section III describes related works. We have introduced our methodology in section IV. In section VI, we have discussed our implementation and section VII describes our result and finally section VIII concludes our paper.

MOTIVATION: From the analysis of the world speech recognition Journal and research, I am coming to this decision that Speech Recognition (Jelinek, 1997) is the most important and unique issue that drives the researcher towards the field. As architecture and application are going large day by day, Speech Recognition has become an important challenge for the present world. Interest in this field is proportionally (Rampey, 2006) increasing with the Heterogeneity (Dietz, 2002) of this network. In a future world, the World may totally depend on AI and be controlled by speech everything.

RELATED WORKS: Speech recognition and speech to the text have seen a lot of research. But they tried to develop systems that contain only two or three languages (Dietz, 2002). But they did not try to implement more than three languages at a time.

Khumbarana et al. (2006) created a speech pattern recognitionsystem that used both the Hidden Markov Model (HMM) and Digital Signal Processing (DSP). Khilari et al. (2002) tried to develop an Automatic SpeechRecognition (ASR) system. It uses system API for converting into only one language. They also used the Hidden Markov Model as the learning algorithm. Simon et al. (2008) worked for phonation, fluency, intonation,pitch variance, and voices. They used the Linear Prediction Coefficients (LPC) feature extraction process model and sta- tistical view and the Hidden Markov Model model for system design. Also, they used the MFCC method for accuracy and noise reduction. Wagner et al. also work with Real-time intra-lingual speech- to-text-conversion. He worked on grammatical mistakes and, he showed that the main challenges are time, message transfer,real-time presentation of the written text. Nasib et al. worked on Text Conversion for the Bengali language. They converted speech by using SNR (Signal-to- noise-ration) and digital workstation (DAW) and achieved77.7% accuracy. Nafis et al. tried to create a speech-to-text conversion in real-time. They implemented their work on MATLAB, and they focused on phonetics and accuracy.

METHODOLOGY

Speech-to-text conversion systems follow at least two mod- els: an acoustic model and a language model. The massive vocabulary systems use a pronunciation model. It is essential to understand that there is no such thing as a public discourse Identifier to get the best copy value, all these Models can be specialized for a given language, dialect, application domain, type of speech, and communication channel. The result of the speech accuracy is mostly dependent on the speaker's voice, the style of speech, and the environmental conditions mainly, but in this system, this problem may be reduced. How good or bad your pronunciation, is not theacceptance level of this system.

Speech detection actually completes its task in three points:

- End point Detection
- Mel Frequency Cepsral Coefficient (MFCC) (Ittichaichareon, 2012), and
- Hidden Markov model recognizer (21)

End point Detection-EPD: In the Speech detection process, there are two sounds provided- voiced and unvoiced. In a noisy environment, most of the time, speech is containing unwanted signals and back- ground noise, which is removed by endpoint detection. EPD method is work based on STLE- the short-term log energyand STZCR- the short-term zero-crossing rate. The following equation calculates STLE and STZCR

Mel Frequency Cepsral Coefficient (MFCC): Reducing the signal size of the proceed speech before the pattern is done by MFCC. The steps of MFCC- Mel frequencyCepstral Coefficients calculation are

- Framing,
- Windowing,
- Discrete Fourier Transform,
- Mel frequency filtering,
- Logarithmic function, and
- Discrete Cosine Transform

The block diagram of the MFCC process is given in the following figure.

HIDDEN MARKOV MODEL RECOGNIZER (HMM) Though there are many methods for speech recognition, HMM (Rabiner, 1986) is the most usable method widely. As writing structure in speech and HMM-Hidden Marcove MOdel follow the left to right sequences. The different states of HMM models describe the word or phonetics of speech recognition. The Hidden Markov model affects the accuracy of the speech recognition (Nilsson, 2002) system. This is the most popular and usable processing tool for designing extensive time-series data.

DTW based speech recognition process: A method which was actually used historically for recog- nized speech but today it is been replaced more successfully by the HM approach. for observing similarity between two sequences which is very speed using Dynamic Time Warping algorithm. by this algorithm, walking patterns similarities are detected, even it should be applied to the video, audio, and graphics where anyone makes a video with walking quickly, sang quickly or illustrate with a heavy pixel. so, linear datacan be analyzed by DTW (Weintraub, 1989).

$$p(W|A) = \frac{p(W)p(A|W)}{p(A)} \tag{1}$$

$$W' = argmaxp(W)p(A|W) \tag{2}$$

$$sgn[p(W)] = \begin{cases} +1, & s(n) \geq 0 \\ +2, & s(n) < 0 \end{cases} \tag{3}$$

Where E_{log} is the $STZCR$ and $s(n)$ is the signal of thespeech, ZCR is the ST(Short-Term) zero-crossing rate.

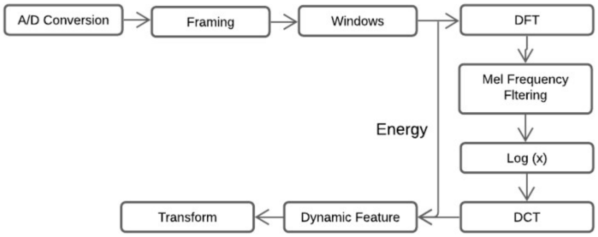


Fig. 1. Block Diagram of MFCC

Mel Frequency Cepsral Coefficient (MFCC): Reducing the signal size of the proceed speech before the pattern is done by MFCC. The steps of MFCC- Mel frequencyCepstral Coefficients calculation are

- Framing,
- Windowing,
- Discrete Fourier Transform,
- Mel frequency filtering,
- Logarithmic function, and
- Discrete Cosine Transform

The block diagram of the MFCC process is given in the following figure.

HIDDEN MARKOV MODEL RECOGNIZER (HMM) Though there are many methods for speech recognition, HMM (Rabiner, 1986) is the most usable method widely. As writing structure in speech and HMM-Hidden Marcove MOdel follow the left to right sequences. The different states of HMM models describe the word or phonetics of speech recognition. The Hidden Markov model affects the accuracy of the speech recognition (Nilsson, 2002) system. This is the most popular and usable processing tool for designing extensive time-series data.

DTW based speech recognition process: A method which was actually used historically for recog- nized speech but today it is been replaced more successfully by the HM approach. for observing similarity between two sequences which is very speed using Dynamic Time Warping algorithm. by this algorithm, walking patterns similarities are detected, even it should be applied to the video, audio, and graphics where anyone makes a video with walking quickly, sang quickly or illustrate with a heavy pixel. so, linear datacan be analyzed by DTW (Weintraub, 1989).

Neural networks: Since the 1980s, neural networks have been defined asan interesting lexical method of modeling in ASR. Since then, it used for many purposes of speech recognition, such as classification of phoneme. It is also used in phoneme classification using an evolutionary algorithm, isolated word recognition. Neural networks are also used for audiovisual speech recognition and speaker recognition and adaptation. Most of the time, there have some future statistics about speech is created by neural networks explicitly, and all these qualities make HMM a very important and effective recogni- tion model for speech recognition. When we used the HMM to assume some speech probabilities of the future segment,it allows discontinue natural training in an efficient manner. Though In short-time classifying units there has been effective- ness like Distinctive sounds and isolated sounds, the last neural networks were seldom successful or continued recognition due to the slight ability to model temporary dependencies. Pre-processing is a limitation of neural networks, including feature conversion, a step before speech recognition making with HMM. However, in recent times, LSTm and RNN of related recurrent and TDNN-time-delay neural networks have changed this situation and improved situation.

Deep feedforward and recurrent neural networks: DNN-Deep Neural Network is under investigation. Demo-nizing Auto Encoder is also today Under investigation. An Artificial Neural Network also contains a deep feed-forwardneural network- DNN with many hidden layers of units be-tween two layers like input and outputs. Like Shallow NeuralNetworks, DNNs Model can Non-linear complex relationships. Compositional models are generated by DNN Architecture,where excessive layer enables compositions of features froma lower layer which is giving many learning capacities and by these the potential of modeling speech data complex pattern. In 2010 there was a breakthrough for DNN in recognition of large vocabulary discourse by some industry researchers,in collaboration with many academic researchers, where largeoutput layers of DNN are based on context-dependent HMM conditions created by a decision tree.

As of October 2014, in the recent Springer book of Microsoft research, we can seea comprehensive review of this development and the state ofthe industry. We can also look at various machine learning examples and the background to the impact of automated speech recognition-related work, especially recent overview articles on deep learning and artificial intelligence. One of the basic principles of deep learning is the elimi- nation of raw features through hand-built feature engineering. The principle was first successfully explored in autoencoders in the ”raw” linear filter-bank feature or spectrogram to demonstrate its superiority over the mail-spectral properties and to contain certain stages of transformation from spec- tral to specific. In predicting and creating large-scale speech identification results, the ”raw” characteristics of speech and waveform have been shown.

End-to-end automatic speech recognition: In 2014, Navdeep Jaitley of the University of Toronto and Alex Graves of Google Deep Mind launched the first attempt at an end-to-end ASR with a connective temporal classification (CTC) based system. This model explains recurrent CTC levels and neural networks. Indeed, the pronunciation and lexical models are learned together by the RNN-CTC model. Like HMM, it is incapable of learning the language for certain conditional freedom assumptions. As a result, although CTC models can learn English characters directly to map speech acoustics, the model makes many common spelling mistakes and has to rely on a different language model to clear transcripts. It was then significant that the work was created by Baidu with very large datasets and created some commercial successin Chinese Mandarin and English. In 2016, Oxford Univer- sity introduced LipNet, the first end-to-end sentence-level lip reading model, using an RNN-CTC architecture and with spa- tiotemporal convolution to surpass human-level performance in a limited grammar dataset. In 2018, google DeepMind introduced a large-scale CNN-RNN-CTC architecture that performed 6 times better than human experts.

Accuracy

The accuracy of the speech identification may depend on the various factors that I explain previously, as follows:

- Increase error rate to increasing size of vocabulary: For example, 10 numbers from ”zero” to ”no” may be fully recognized, but may have an error rate of 4 %, 8 %, or 48 %, respectively, in the form of 400, 6000, or 150,000 vocabularies.
- If it cannot identify the voice of words as it is maleor female, the vocabulary is difficult to identify: As confusing words, it is very difficult to distinguish all the English alphabet(mainly, e-set: ”B, C, D, E, G, P, T, V, Z”); An 8% error rate is considered good for this vocabulary (Congalton, 1991)
- Freedom VS. Speaker dependence: A single speaker uses a speaker-dependent system. For any speaker (albeit moredifficult), the purpose of a speaker-independent system.
- Isolated, uninterrupted or uninterrupted speech: Single words are used in isolated speech, so speech is easy to identify. Complete sentences separated by silence are used in continuous speech, so it becomes easier to identify speech as well as isolated speech. Normally pronounced sentencesare used in continuous speech, so speech becomes difficult to identify, which is different from both isolated and isolated speech.
- Language and work limitations: Asking such questions can disprove the notion of ”orange blue.” Such limitations can be semantic; Rejected” Orange is angry.” Such as syntactic; Rejection is ”blue orange.”

IMPLEMENTATION

To convert the speech to text from input to output, wefollowed five steps. These steps are

- Vocabulary database,
- Pre-processing,
- Speech recognition,
- Noise removing (22) and,
- Convert to a specific language.

At first, we created a vocabulary dictionary in the database. Then we determine the language of the input. When selecting a particular language, speech recognition steps would start and convert to that language from the input voice. If there is any noise in the voice, the noise remover starts its function and carried out the text accuracy (Abdel-Hamid, 2012).

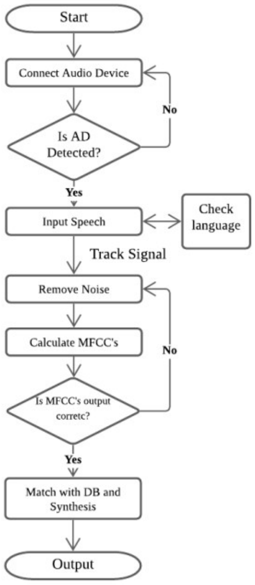


Fig. 2. Flowchart of the Speech to Text conversion process

For getting the full output,we are working with HTML5 and CSS3. For Platform, we used mainly Visual Studio because this project is done by C# programming language. By HTML and CSS, we create front end design and C# with core JS. We have made a backend processing task. After the process, the dictionary provides the text in the output box in a specific language among about 100 languages. In a visual studio platform, it needs to select the ASPX file. It also needs a host or local host. without internet orhosting, it can't be work.HTML5 and CSS3 actually designed the user interface but it possible to design by another front-end language. At the time of testing of this system, the place must have to noise-free. The flowcharts of the implementationprocess are given in figure 2.

RESULTS

To test the system, we need to first select a specific languagefrom the drop-down menu. There is some language that also has different regional states 2. After choosing the language, we have to tap the speech icon and it will allow your microphone to receive the voice input. And then it will show the result. Here we have given some input signals and output languages

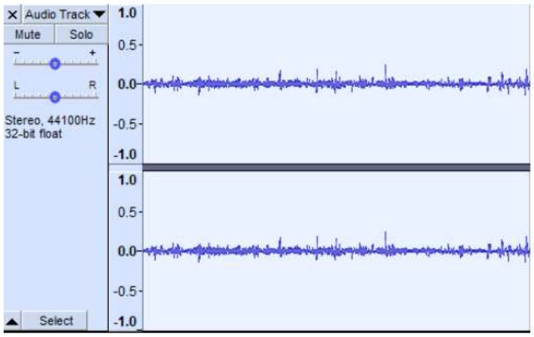
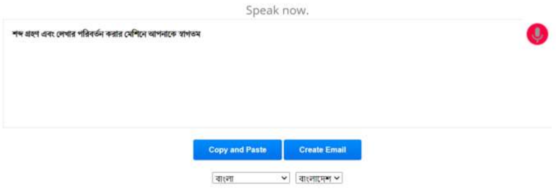


Fig. 3. Bengali (Bangladesh format) language conversion input


Speak now.

শব্দ গ্রহণ এবং লেখার পরিবর্তন করার মেশিনে আপনাকে স্বাগতম

Fig. 4. Bengali (Bangladesh format) language conversion output

CONCLUSION

Improving Speech recognizer and speech to text has been a hot research fields. Speech to text conversion process can really help a lot of people as it has spelling correction, grammatical correction and multi-language support. Disabled people will find it very easy to use and they will be benefited. We hope our Maskitron will have a great impact into this research field.

REFERENCES

Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Penn, G. 2012. Applying convo-lutional neural networks concepts to hybrid nn-hmm model for speech recognition. In: 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP), pp. 4277–4280. IEEE

Ballinger, B.M., Schalkwyk, J., Cohen, M.H., Allauzen, C.G.L. 2016. Lan- guage model selection for speech-to-text conversion. US Patent 9,495,127

Ballinger, B.M., Schalkwyk, J., Cohen, M.H., Allauzen, C.G.L., Riley,

M.D. 2011. Speech to text conversion. US Patent App. 12/976,972

Bijl, D., Hyde-Thomson, H.: Speech to text conversion (2001). US Patent 6,173,259

Bourlard, H.A., Morgan, N. 2012. Connectionist speech recognition: a hybrid approach, vol. 247. Springer Science & Business Media

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote sensing of environment **37**(1), 35–46.

Dietz, T.A.: Speech recognition text-based language conversion and text-to-speech in a client-server configuration to enable language translation devices (2002). US Patent 6,385,586

Ittichaichareon, C., Suksri, S., Yingthawornsuk, T. 2012. Speech recognition using mfcc. In: International Conference on Computer Graphics, Simulation and Modeling, pp. 135–138

Jelinek, F. 1997. Statistical methods for speech recognition. MIT press.

Lee, K.F. 1988. Automatic speech recognition: the development of the SPHINX system, vol. 62. Springer Science & Business Media

Lerner, N. 1989. Speech recognition bibliography. ACM SIGOIS Bulletin **10**(3), 1–13.

Nilsson, M., Ejnarsson, M. 2002. Speech recognition using hidden markov model

Pinson, M., Pinson, D., Flanagan, M., Makanvand, S., et al.: System and method for automatic speech to text conversion (2013). US Patent 8,566,088

Rabiner, L., Juang, B. 1986. An introduction to hidden markov models. ieee assp magazine **3**(1), 4–16.

Radha, V., Vimala, C. 2012. A review on speech recognition challenges and approaches. doaj. org **2**(1), 1–7.

Rampey, F.D., Macmillan, J.M. 2006. Speech to text conversion system. US Patent 7,130,401

Reddy, D.R. 1976. Speech recognition by machine: A review. Proceedings of the IEEE **64**(4), 501–531

Rigoll, G. 1994. Maximum mutual information neural networks for hybrid connectionist-hmm speech recognition systems. IEEE Transactions on Speech and Audio Processing **2**(1), 175–184

Shaw, V., Evora, R.Z. 2013. User profile based speech to text conversion for visual voice mail. US Patent 8,358,752

Simoneau, L., Soucy, P. 2008. Method to train the language model of a speechrecognition system to convert and index voicemails on a search engine. US Patent 7,415,409

Tiwari, V. 2010. Mfcc and its applications in speaker recognition. International journal on emerging technologies **1**(1), 19–22.

Wang, J.C., Wang, J.F., Weng, Y.S. 2002. Chip design of mfcc extraction for speech recognition. Integration **32**(1-2), 111–131.

Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Baldwin, G., Bell, D. 1989. Linguistic constraints in hidden markov model based speech recognition. In: International Conference on Acoustics, Speech, and Signal Processing,, pp. 699–702. IEEE.

Witt, S.M. 1999. Use of speech recognition in computer-assisted language learning.

****