KJAR
Kurdistan Journal of Applied Research

Performance Analysis: AI-based VIST Audio Player by Microsoft Speech API

Ribwar Bakhtyar Ibrahim
Database Technology
Technical College of Informatics
Sulaimani Polytechnic University
Sulaimani, Iraq
ribwar.ibrahim@spu.edu.iq

Article Info

Volume 6 – Issue 1- June 2021

DOI:
10.24017/science.2021.1.2

Article history:

Received:04/04/2021
Accepted:28/04/2021

Keywords:
Speech Recognition, Microsoft Speech API, Subtitles, Speech to Text, speech-to-text recognition, Artificial Intelligence. A Voice Interactive Speech to Text (VIST), Microsoft Speech API.

ABSTRACT

Speech recognition has gained much attention from researchers for almost last two decades. Isolated words, connected words, and continuous speech are the main focused areas of speech recognition. Researchers have adopted many techniques to solve speech recognition challenges under the umbrella of Artificial Intelligence (AI), Pattern Recognition and Acoustic Phonetic approaches. Variation in pronunciation of words, individual accents, unwanted ambient noise, speech context, and quality of input devices are some of these challenges in speech recognition. Many Application Programming Interface (API)s are developed to overcome the issue of accuracy in a speech-to-text conversion such as Microsoft Speech API and Google Speech API. In this paper, the performance of Microsoft Speech API is analyzed against other Speech APIs mentioned in the literature on the special dataset (without background noise) prepared. A Voice Interactive Speech to Text (VIST) audio player was developed for the analysis of Microsoft Speech API. VIST audio player creates runtime subtitles of the audio files running on it; the player is responsible for speech to text conversion in real-time. Microsoft Speech API was incorporated in the application to validate and make the performance of API measurable. The experiments proved the Microsoft Speech API more accurate with respect to other APIs in the context of the prepared dataset for the VIST audio player. The accuracy rate according to the precision-recall is 96% for Microsoft Speech API, which is better than previous ones as mentioned in the literature.

1. INTRODUCTION

Speech recognition is the conversion a speech signal to words sequence through a computer program or an algorithm [1]. Human communication is mostly in the form of Speech. With the

help of speech recognition computer understands human voice commands [1]. The goal of speech recognition is to develop techniques that a human and machine may interact with each other and understand human voice commands [2].
The progress of automatic speech recognition starts from fifty years ago. In the duration of 1950-1960, the researcher had tried on human input voice signals, and most of the speech recognition system investigated used spectral resonance during the vowel region of each utterance which was extracted from the output signal analog filter bank of the logic circuit.
In the 1970s, speech recognition research was achieved, the area of isolated and discrete utterance became a usable technology. Velichko and Zagorruyko in Russia advanced the use of pattern recognition model ideas in speech recognition [3].

In the 1980s, The Hidden Markavo Model (HMM) technique was introduced in the same way for speech recognition. In the 1990s, the focus of the researcher was on robust speech recognition. Matsui et al[4] compared the VQ-based method with the discrete/continuous HMM-based method, particularly from the viewpoint of robustness against utterance variations. A DARPA program (2000): Effective Affordable Reusable Speech to text (EARS) was developed as speech-to-text technology with the aim of achieving accuracy in speed to text conversion. The program was focusing on unconstrained human-human and natural speech from foreign conversational speech and broadcasts in multiple languages, to make machines do better work for detecting, extracting, and summarizing, and translating information [5].
In 2001, computer speech recognition had reached 80% of accuracy and not progressed to 2010. When Google developed a voice search application for iPhone, Google added personalized recognition to voice search on Android phones. Google developed an application that produced the most accurate speech model and added voice search to chrome browser in mid-2011 [6].
There are a number of audio and video players are available in the market, which has different characteristics and features. If we want to watch a movie or scene or want to hear audio, and we have no speaker system or we do not know the language of the movies. Then in that case subtitles of the movie are very helpful to understand the scene or movie. But the problem is that these subtitles are added with a movie scene by editing movie scene and it is embedded in every frame of the movie. Also know that subtitles are not added to every movie or video, because it is a very time-consuming process.
The challenges of converting an audio soundtrack to accurate text are multiple which might have different sound effects, voices, background music, unexpected vocabulary, and different accents switching between them, and punctuation. A clear audio from a single speaker can be slower while processing and the results might be unpredictable. Transcribing through a re-speaker would be quicker over automatic soundtrack transcription. Though a number of problems exists in speech recognition, this this paper focuses on the synchronization of audio speech with text displayed and to find the accuracy of Microsoft speech API, under certain limitations or conditions.

1.2 Types of Speech Recognition:

There are different classes of speech recognition, such as:
Isolate words: In isolated word recognition, it accepts a single word at a time. This depends on listening or not listen to the state of the speaker, the speaker has to wait for a response from the machine [1].

Connected words: It is similar to Isolated Words Recognition System but allows separating words to run together with a minimal pause between these words [1].

Continuous speech: This type of speech allows the user to speak continuously, but for computers, the detection of separation of words becomes difficult in this case [1].

1.3 Approaches to speech recognition:

Acoustic Phonetics, Pattern Recognition, Artificial Intelligence are the three approaches to speech recognition.

Acoustic Phonetic Approach: Generally, speech recognition deals with speech variability. In Acoustic Phonetic Approach, the system tries to decode the speech signal in a sequential manner which is based on the observed acoustic features of the speech waveform and the known relations between an acoustic feature and phonetic symbols.

Pattern Recognition Approach:

In Pattern Recognition Approach, the speech pattern is used directly without explicit feature determination and segmentation. This method has two steps, first is step training of speech pattern and the second is pattern recognition [2].

Artificial Intelligence Approach*:* The third approach is the Artificial Intelligence approach or knowledge base approach. It is a hybrid approach of the acoustic-phonetic approach and pattern recognition approach. Some speech researchers developed a recognition system that used phonetic knowledge to develop classification rules for speech sounds [2].

1.4 Issues in Speech to text recognition:

Software that is using speech recognition still has some big problems, which prevents it from being used exclusively. Some of the speech recognition problems that are difficult to solve are variation in the pronunciation of words, individual accents, and unwanted ambient noises. The context of the words being spoken is also a problem, which leads to an incorrect word or incorrect spelling [3].

The quality of the input device creates problems in speech recognition. If the microphone is not sensitive enough or is highly sensitive, then it can create audio information that is difficult for the software to convert accurately. The main recognition issues are Accuracy, User Responsiveness, performance, reliability, and fault tolerance. This paper is focused on the analysis of speech recognition using Microsoft Speech API in software developed software VIST Audio Player, this paper also checked synchronization of audio speech to the converted text displayed by the VIST Audio Player in the form of subtitles.

1.5 VIST Audio Player

The VIST (Voice Interactive Subtitle) Audio Player is a desktop application that is used to play audio speech of different formats such as mp3, wave, etc. and is capable to display subtitles (text) of only wave format in the form of a label, human voice interactive commands control different operations in the player, for example, stop, play, pause, change volume and control playlist. Speech-to-text conversion and speech recognition are the main parts of the VIST Audio player.

The rest of the paper is organized as section 2 is composed of Related Work, section 3 discusses the methodology, section 4 shows the experiments, section 5 is based on the results and analysis, in Section 6 conclusion, limitations and the future work are discussed.

2. RELATED WORK

In the literature review, published research papers and other material related to the research topic being studied, and data extracted from different research papers. A number of researches have been done since Speech API is developed; most of the researches were done by the Microsoft speech group itself.

Two factors affect the accuracy of speech recognition i.e. Speech Engine and Language Model. Microsoft group introduced a new technology called whisper, which explains different approaches for accurate, efficient, and more usable speech recognition [9]. With the use of Context-Free Grammar (CFG) and applying a set of rules, during normalization produce differentiate noise from speech and reduce the error rate up to 50% [9]. Another paper focuses on the fast talker (rate of speaking words), the utterance of speech is stretching in such a way that it matches the acoustic model, it reduces the error rate up to 13% [7]. In another research article semantic language model, Integrated Development Environment (IDE) was used in the speech language model [8]. Another interesting paper is discussing name recognition; name recognition is the same as command recognition. This idea was used in the VIST audio player

developed, for input commands, and on matching input command with pre-defined text it performs appreciate operation in the VIST audio player. In this pattern coronation model, the accuracy is 99%.
There is another technique for accuracy is the training technique, experiment shows that this technique is fast and accurate and 5.75% word error rate [6]. The research paper was done by the Microsoft research group is about voice search, this paper first describes the recombination process and then proposes a statistical language model that automatically converts listing into query form, which works as training for voice [10].
Human being authentication by offline handwritten signature biometric research has been increasing, especially in the last decade. Verification process of an offline handwritten signature is not trivial task [11]

All the concepts discussed in the above papers have focused on the efficiency and accuracy of speech recognition. The researchers are trying to develop such a technique that increases the efficiency of speech recognition APIs.

3. METHODOLOGY

The developed a desktop application "VIST Audio Player". This application is used for the audio speech translation into text form in real-time, and displays that text in the form of subtitles, just like a video player. As speech translation is a different and separate process from playback audio, this research analysis is trying to match these two processes (audio speech and text in the form of subtitles) means synchronize and find the accuracy of word recognition. In figure 3.1, a data flow diagram of the VIST audio player is shown which helps in understanding the working of the software.
Using a Microsoft API, this player was developed. In the first portion, this player was made voice interactive, which can understand human voice command. The Voice recognition pattern method was used for voice recognition. The voice as input is matched with the pre-defined and stored text data in the form of strings. On the success of recognition, the specified action is taken as defined.
Microsoft API is also used for the second time for converting audio speech into text form. But this time it is based on word recognition, and events are developed for the detection of words as well as characters. On completion of the whole world, it correctly recognized words. As mentioned above, the audio playback and speech to text conversion are separate processes, there are two threads created for each process to run both the process parallel and synchronize speech to the text displayed. These treads can be stopped, started, and paused simultaneously. Besides this, it is proved experimentally that it gives 96% of accurate results.

4. EXPERIMENT

In order to check the performance of developed software (VIST Audio Player), and the performance of Microsoft Speech API under certain limitations, experimental work is done. Twenty-five audio speech samples were checked using the player for the performance checking of VIST audio player software, and for the efficiency of Microsoft platform Speech API the number of words of each audio file was counted in the first step; similarly, playback duration of these twenty-five samples was collected. Here experimentally checked every audio file (from the selected list of twenty-five files) separately which displayed text on the player.
Through careful observation, the number of accurate words and incorrect words was detected. The accuracy of each audio file was measured in percentage. In the last step, we also find the average percent accuracy of twenty-five files. Along with the accuracy checking, synchronization of audio speech was also checked.

5. RESULT AND PERFORMANCE ANALYSIS

Table 5.1 shows the results of twenty-five audio sample files of wave format used in the experiment. This table contains numbers of the accurate word, a number of incorrect words and percent accuracy of each file separately and total average percent accuracy, which is 96%, and was found a better result as compare to other methodologies as mentions in the studied in the previous research works. Generally, as mentioned in the above literature, in one place the efficiency of speech API or ASR (automatic speech recognition) is shown 99%, but it must be remembered that it is in the case of a pre-defined pattern recognition model in real-time voice input as a microphone. Here, speech API was used for pre-recorded audio files, which were directly converted into text form. From the studied literature it was found an average of 70 to 80% (in the case of pre-recorded files or real-time speech recognition).
Secondly, synchronization of audio speech to the displayed text in the form of labels is a big issue, which was working across threading, was solved in VIST Audio player, to a great extent. The average delay in most cases is zero or one or two seconds as we can see in Table 5.1, but in some cases, it is three seconds. In the worst case, it is seven seconds.

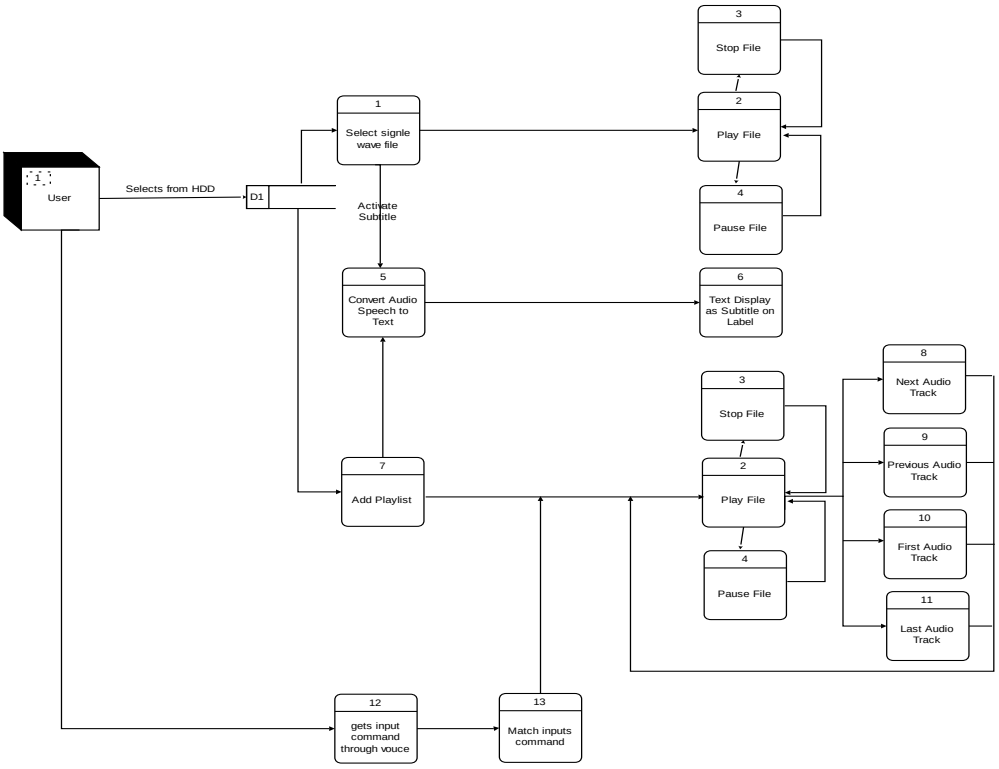

Figure 1: Flow Diagram of the AI-based VIST Audio Player

Table 1: Results

S.No	File Name	File Format	Playback duration	Total Words	Incorrect Words	Correct words	Not Retrieved words	Precision	Recall	Synchronization delay
1	01.wav	Wave	1:06	158	6	152	0	96	100	0 sec
2	02.wav	Wave	1:45	255	7	248	0	97	100	0 sec
3	03.wav	Wave	0:34	66	4	62	0	94	100	1 sec
4	04.wav	Wave	1:08	143	2	141	0	99	100	1 sec
5	05.wav	Wave	0:34	76	9	67	0	88	100	0 sec
6	06.wav	Wave	1:15	168	12	156	0	93	100	0 sec
7	07.wav	Wave	0:40	116	7	109	0	94	100	1 sec
8	08.wav	Wave	1:49	187	6	181	0	97	100	2 sec
9	09.wav	Wave	0:31	67	2	65	0	97	100	0 sec
10	10.wav	Wave	1:05	147	4	143	0	97	100	0 sec
11	11.wav	Wave	1:28	181	13	168	0	93	100	0 sec
12	12.wav	Wave	0:55	98	5	93	0	95	100	0 sec
13	13.wav	Wave	1:52	207	10	197	0	95	100	3 sec
14	14.wav	Wave	1:25	191	9	182	0	95	100	3 sec
15	15.wav	Wave	1:52	251	12	239	0	95	100	3 sec
16	16.wav	Wave	1:52	337	36	301	0	89	100	7 sec
17	17.wav	Wave	1:21	232	5	227	0	98	100	0 sec
18	18.wav	Wave	0:54	93	1	92	0	99	100	0 sec
19	19.wav	Wave	0:37	64	3	61	0	95	100	0 sec
20	20.wav	Wave	0:27	52	0	52	0	100	100	2 sec
21	21.wav	Wave	0:38	91	2	89	0	98	100	0 sec
22	22.wav	Wave	0:20	53	1	52	0	98	100	0 sec
23	23.wav	Wave	0:51	134	2	132	0	99	100	0 sec
24	24.wav	Wave	1:03	126	3	123	0	98	100	0 sec
25	25.wav	Wave	1:32	190	5	185	0	97	100	3 sec
								96%	100%	

The analysis is done using slandered Precision and recall method, as result it gave 96% Precision and 100% Recall Using the formula of precision and recall respectively.

$$Precision = \frac{A}{(A+C)} * 100\% - - - \rightarrow (i)$$

$$Recall = \frac{A}{(A+B)} * 100\% - - - - \rightarrow (ii)$$

Where,

A= No of relevant words that are retrieved

B=No of relevant words not retrieved

C= No of irreverent words retrieved

6. CONCLUSION, LIMITATIONS, AND FUTURE WORK

6.1 CONCLUSION

VIST Audio player is a desktop application that is human voice command interactive player, which converts audio speech into English text format. This application is using Microsoft Speech API, for speech recognition and voice-to-text conversions. This paper mainly focuses on the performance analysis of speech API which was experimentally proved by using twenty-five audio samples of. Wave format. The resultant accuracy was found 96% using VIST Audio Player, which is a better result in the case of pre-recorded audio files against the previous techniques discussed in the literature. The VIST software application is helpful for those peoples who cannot hear an audio speech due to some reasons such as external noise in the environment, or VIST Audio Player software helps those peoples who want to communicate and control (such as play pause stop, etc.) with voice commands.

6.2 LIMITATIONS

There are a few limitations or restrictions in this research works. Software that is using speech recognition still has some problems, which are difficult to solve such as variations in the pronunciation of words, individual assent and unwanted noise in large amount leads to inefficiency of applications using speech recognitions. Context of the words is also a separate problem that leads to incorrect spelling and sometimes completely incorrect words, which are difficult to recognize.

Similarly, the quality of input devices or sensitivity or less sensitivity of microphone also fails speech recognitions. The main recognition issues are Accuracy, which was solved here under certain conditions. The resultant accuracy is 96% if this software VIST audio player is using recommended context and suggested pronunciation by Microsoft. Here five recommended contexts are used for the twenty-four files.

Another limitation is related to the application developed for analysis. The look makes the speech-to-text conversion in real life the process synchronization is done using threads. The main issue with real-time conversion is, it requires more processing time and matching pattern, due to more processing time the speech and converted text being displayed, it cannot be synchronized especially in case of rapid conversion and less processing speed or slow conversion and more processor speed. Secondly for getting accurate result pattern matching is also necessary, which is difficult to provide for every word or speech.

Third, this software is inefficient in a speech-to-text conversion when other languages except English are used. Similarly, in the case of music audio files, this software is also inefficient in-text conversions. This software only supports ".wave" audio format for speech-to-text conversions.

6.3 FUTURE WORK

In the future, the researchers are required to mainly focus on speech API, and it is possible to develop such a speech API, which is efficient with each respect. And this speech API having the capability to recognized every natural speech of humans, with this development automatic speech recognition will be used in a number of software applications. This will be a greater achievement in human voice interactive software applications, through which humans will be able to control different operations of different hardware or machinery.

References

[1] S. Atma Prakash, R. Nath, and S. Kumar. "A Survey: Speech Recognition Approaches and Techniques", 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE, 2018.

[2] F. Sadaoki. "50 years of progress in speech and speaker recognition research." ECTI Transactions on Computer and Information Technology (ECTI-CIT) 1.2 (2005): 64-74.

[3] V.M. Velichko and N. G Zagoruyko: Automatic recognition of 200 Word, Int. J. Man- Machine Studies.2 pp, 223,1970

[4] M. Tomoko, and S. Furui. "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's." IEEE Transactions on speech and audio processing 2.3 (1994): 456-459.

[5] H. Miss, S. Kaur, and V. Chaudhary. "Literature Survey on Automatic Speech Recognition System‖." International Journal of Emerging Technology and Advanced Engineering (2014).

[6] D. Jasha, and A. Acero. "Joint discriminative front end and back end training for improved speech recognition accuracy" 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Vol. 1. IEEE, 2006.

[7] R. Matthew, et al. "Improvements on speech recognition for fast talkers." Sixth European Conference on Speech Communication and Technology. 1999.

[8] W. Yongzhen, X. Liu, and Z. Gao. "Neural related work summarization with a joint context-driven attention mechanism." arXiv preprint arXiv:1901.09492 (2019).

[9] H. Xuedong, et al. "Microsoft Windows highly intelligent speech recognizer: Whisper." 1995 International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. IEEE, 1995.

[10] L. Xiao, et al. "Language modeling for voice search: a machine translation approach" 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008.

[11] N. Abbas, K. Yasen, K. Faraj, L. A Razak, F. Malallah, "Offline handwritten signature recognition using histogram orientation gradient and support vector machine" Journal of Theoretical and Applied Information Technology 2018.

[12] H. Yanzhang, et al. "Streaming end-to-end speech recognition for mobile devices." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.