International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 12 (2020), pp. 5129-5134

Finding Accuracy of Utterance of Language Model

Nadeem Ahmed. Kanasro[1], Najma Imtiaz Ali[2*], Ghulam Muhammad[3], Mujeeb U Rehman Maree[4], A.G Memon[5]

IMCS & SUCL, University of Sindh, Jamshoro, Pakistan

IMCS, University of Sindh, Jamshoro, Pakistan & IIUM, Malaysia

Bahria University Karachi Campus, Pakistan

IMCS, University of Sindh, Jamshoro, Pakistan

IMCS, University of Sindh, Jamshoro, Pakistan

Abstract

In computer world the research over speech recognition technology is carried out by many scientists, engineers and researchers. Since the birth of recognition of speech & voice many number of applications have been created where researchers trying to get more accurate results. For this purpose variety of algorithms are designed which are providing services like ADC (Analog to Digital Conversion), feature extraction, generate speech fingerprints and compare & select most probable words. Utterance is any input which may be speech stream, word, phrase and sentence. This research article presents number of hit (Word Accuracy Rate) and un-hit utterances (Word Error Rate) of Language Model namely (HTML) taken during the detection process of speech in our developed ASR (Automatic Speech Recognition) System namely (Text Editor). Experiment on recognition of 63 words / phrases are done.

Keywords: WER (Word Error Rate), WAcc (Word Accuracy), Language Model

1. INTRODUCTION

Automatic Speech Recognition (ASR) system accurately translates spoken utterances into text. Speech to Text recognition is the ability of a machine to recognize the human speech and convert it into text sequence. For speech analysis and synthesis the first researcher namely Homer Dudley proposed system model that mimics human behavior particularly capable of recognizing speech inputs and responds properly, researchers and scientists intrigued in his proposed model. Research over speech recognition technology is carried out by many scientists, engineers and researchers. Since the birth of recognition of speech & voice many number of applications have been created where researchers trying to get more accurate results. The accuracy is the main reasons behind designing of new algorithms, for this purpose variety of algorithms are created for characteristic origin:

- Linear predictive cepstral coefficients (LPCC)
- Linear predictive analysis (LPC)
- Mel-frequency cepstral coefficients (MFCC)
- perceptual linear predictive coefficients (PLP)
- Mel scale cepstral analysis (MEL)
- Power spectral analysis (FFT)
- Relative spectra filtering of log domain coefficients (RASTA)
- Gammatone Cepstral Coefficients (GTCC)
- First order derivative (DELTA) [5]

And for comparing & selecting best probable match for word and sounds Models/Techniques/Algorithms are Dynamic Time Warping (DTW), Hidden Morkov Model (HMM), Gaussian Mixture Model (GMM), Neural Network, Deep Neural Network (DNN)[5,6].

A. Utterance?

When a user uses the microphone and speaks something between two periods of silence is known as utterance. Utterance is any speech stream. It is a word, phrase or even complete sentence its processing is done by speech recognition engine. In recognition process silence is essential criteria because it delineates the beginning and ending of utterance. Figure #01shows the concept of an utterance. The beginning of utterances is signaled when engine detects input. Similarly the end of utterance is signaled when engine detects certain amount of silence following the input audio [8].

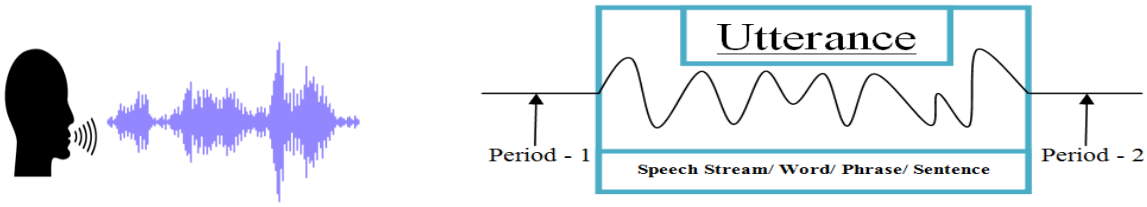

Figure #01 Concept of Utterance

B. The Speech Recognition Engine working mechanism:

Taking a help of a machine speech recognition technology can acknowledge spoken words and phrases. After recognizing words and phrases it can further be used to generate text. The working mechanism of speech recognition system depends upon two modeling techniques namely Language Model (LM) and Acoustic Model (AC). Language Model is main component of an automatic speech recognition (ASR) system [7]. The roll of a language model is to generate the probability that a certain sequence of words can occur in natural speech. This probability is helpful when the acoustic model gives a number of hypotheses with similar probabilities and additional information is needed in order to decide the resulting sequence of recognized words.

The most commonly used language model is an *n*-gram model that predicts the probabilities of words based on their $n-1$ predecessors in the text. Over the years there have been many attempts of perfecting general language models and their adaptation for different domains of use [1, 2, 3]. Actual statistical relationship between phonemes and audio signals is represented in acoustic modeling.

This very first step when a user uses MIC (microphone) utterances may be input as list of words which are recognized then engine loads audio of those utterances and analyzes audio for distinct sounds and characteristics and then compare sounds to internal acoustic model using list of words, finally return probable matches. Figure # 02 shows Speech Recognition Engine Process.

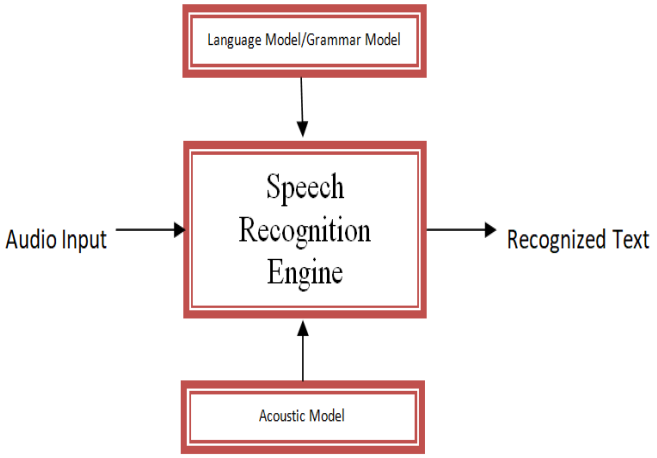

Figure # 02 Speech Recognition Engine

C. Evolution of Speech Recognition Systems

In Vienna the first mechanical machine which was capable of handling acoustic was created by a scholar. Then in 1879 the first machine which could handle dictation was invented by Thomas Edison. Then in 1952 at Bell Laboratories speech recognition system was developed which was capable of recognizing numbers it was speaker dependent just understand voice of inventor.

Continuing with the chain, a Harpy System was developed by a scholar in 1970 which could recognize 1000 words and was capable of recognizing different pronunciations and some phrases [8, 9]. In 1980s A mathematical based approach namely HMM (Hidden Morkov Model) was introduced which could analyze sound waves efficiently. In 1986 IBM Tangora using HMM made a technique of prediction which were recognizing upcoming phonemes in speech. Segmentation of keywords in recorded speech for speech recognition system was started by Nation Security Agency (NSA).

Then there arrived a boom of speech recognition technology when the world's top IT companies comprising of Google, Facebook, Microsoft, Amazon and Apple started offering this functionality in various devices through services like Amazon Echo, Google Home, Apple Siri and many more [12]. The goal of these top tech companies is to make voice assistants response and reply with more accuracy.

Today's world most popular speech recognition systems are:

1. YouTube close captioning
2. Voicemail Transcription
3. Dictation System
4. Google Voice
5. Amazon Echo
6. Apple Siri

D. Speech Recognition Areas

More number of areas exists which are using speech recognition technology like:

- Voice user interfaces
- Simple data entry
- Speech-to-text processing
- Home appliance control
- Air Craft Systems
- Medical Systems
- Car Systems

E. Speech Recognition Computing Environments

Various electronic devices like PC, Laptop, mobile phone, traditional phone, IPod are user-end devices as shown in figure # 03 are being used now as main source of communication users could even control home appliances using Speech Recognition Technology in addition to GPS system, Wi-Fi, Bluetooth, and Traditional telephone system different computing environment[10] exists like:

- Traditional Computing
- Web based Computing
- Cloud Computing
- Pervasive Computing

- Embedded Computing

Figure #03 Computing Environment

F. Metric for finding Accuracy & Performance

Two parameters namely (Accuracy and Speed) are used for evaluating the performance of speech recognition technology. The metric (i) is used for finding WER (Word Error Rate) where the metric (ii) is used for finding Word Accuracy Rate (WAcc)

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \qquad (i)$$

In equation number (i) I is (Insertions), S is (Substitutions), D is (Deletions), C is (Correct) number of words and N=(S+D+C)

$$WAcc = 1 - WER = \frac{N-S-D-I}{N} = \frac{H-1}{N} \qquad (ii)$$

In equation number (ii) H is N-(S+D) is the number of correctly recognized words. IF I=0 then WAcc will be equivalent to Recall (Information Retrieval) a ratio of correctly recognized words 'H' to Total number of words in reference 'N'.

2. PROPOSED METHODOLOGY

As we have proposed three Language Models one Grammar model as shown in Figure #04. Speeches to text processing models are (Dictionary, HTML, and Special Characters). These language models would be used as Dictation purpose while IDE (Integrated Development Model) is Grammar Model would be used as Command and Control Purpose [4]. Their Context Free Grammar (CFG) is implemented using Automatic Speech Recognition Engine as shown in Figure #05.

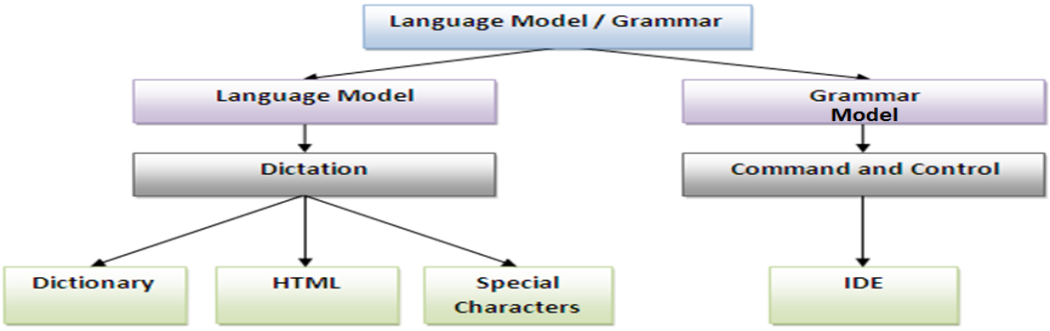

Figure #04 Classifications of Proposed Model [4]

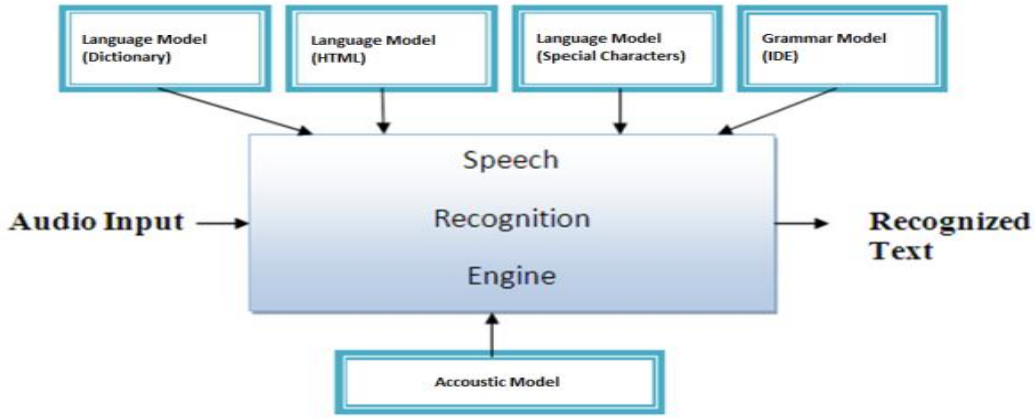

Figure #05 Implementation of Proposed Language and Grammar Model [4]

3. DESIGNING, IMPLEMENTATION AND TESTING

A user friendly application is designed in Visual Basic.net using RAD (Rapid Application Development) Model as in Figure #06. A simple script of basic html tags is executed / dictated as shown in Figure #06.

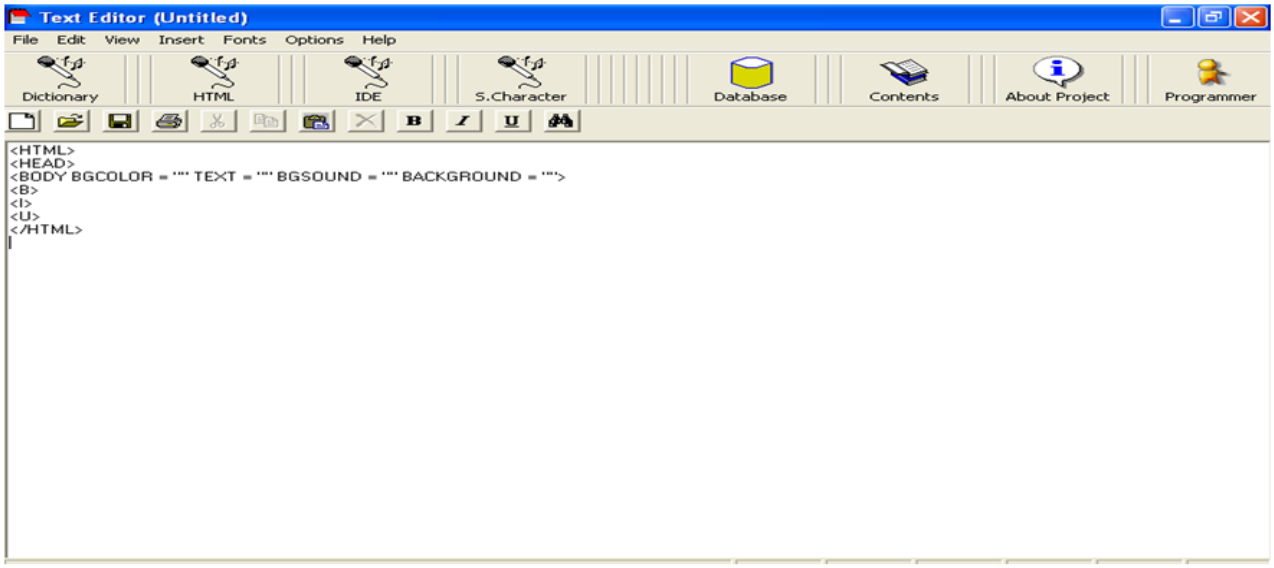

Figure #06 Speech Recognition of designed Language Model (HTML)

4. RESULTS

Initially testing done over one language model namely HTML. This model contains about 63 different words and phrases which are shown in table# 01, each single word or phrase was uttered 100 times, the excellent accuracy rate was obtained and less error rate was found. Normal Person's voice sample was taken out Using Microsoft Speech Recognition in addition to Google transcribe and our developed module. The average accuracy rate was found 91.11% and the average error rate was found 8.88% as shown in chart #01.

Table#01 List of Words / Phrases of Language Model (HTML)

S.No:	List of Words	Output	Out of 100 inputs	
			Accuracy Rate	Error Rate
1	HTML	<HTML>	91	9
2	Head	<Head>	96	4
3	Title	<Title>	93	7
4	Body	<Body>	94	6
5	Image	<Img>	92	8
6	B	<B>	99	1
7	I	<I>	99	1
8	U	<U>	99	1
9	Center	<Center>	92	8
10	Font	<Font>	90	10
11	HR	<Hr>	89	11
12	BR	 	88	12
13	P	<P>	91	9
14	Table	<Table>	92	8
15	TH	<TH>	94	6
16	TR	<TR>	93	7
17	TD	<TD>	94	6
18	H1	<H1>	90	10
19	H2	<H2>	89	11
20	H3	<H3>	89	11
21	H4	<H4>	89	11
22	H5	<H5>	90	10
23	H6	<H6>	87	13
24	Sub	<sub>	91	9
25	Sup	<sup>	90	10
26	Marquee	<Marquee>	89	11
27	Frame	<Frame>	89	11
28	Frameset	<Frameset>	88	12
29	Form	<Form>	89	11
30	Input	<Input>	91	9
31	Select	<Select>	96	4
32	Option	<Option>	93	7
33	Text Area	<Textarea>	94	6

S.No:	List of Words	Output	Out of 100 inputs	
			Accuracy Rate	Error Rate
34	Close HTML	</HTML>	92	8
35	Close HEAD	</HEAD>	99	1
36	Close TITLE	</TITLE>	99	1
37	Close Body	</Body>	99	1
38	Close B	</B>	90	10
39	Close I	</I>	89	11
40	Close U	</U>	88	12
41	Close Center	</Center>	91	9
42	Close Font	</Font>	92	8
43	Close HR	</HR>	94	6
44	Close BR	</BR>	93	7
45	Close P	</P>	94	6
46	Close Table	</Table>	90	10
47	Close TH	</TH>	89	11
48	Close TR	</TR>	89	11
49	Close TD	</TD>	89	11
50	Close H1	</H1>	90	10
51	Close H2	</H2>	87	13
52	Close H3	</H3>	91	9
53	Close H4	</H4>	90	10
54	Close H5	</H5>	89	11
55	Close H6	</H6>	89	11
56	Close Sub	</Sub>	88	12
57	Close Sup	</Sup>	89	11
58	Close Marquee	</Marquee>	87	13
59	Close Frame	</Frame>	88	12
60	Close Frameset	</Frameset>	87	13
61	Close Form	</Form>	86	14
62	Close Select	</Select>	87	13
63	Close Text Area	</Textarea>	85	15

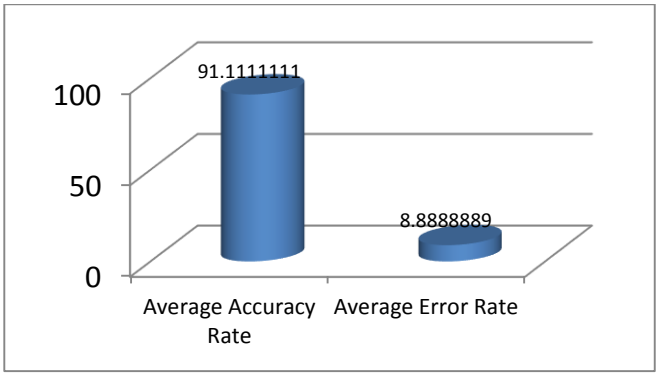

Chart #01 Average Accuracy Rate & Error Rate of Language Model (HTML)

5. CONCLUSIONS

In this research work we have portrayed a concept that a web programmer could create simple web pages using speech recognition technology. Speaker Independent Automatic System Recognition is designed and developed, One of the most successful and automatically trained speech recognition algorithm namely Hidden Morkov Model (HMM) was used for Speech Processing which only recognized isolated and discontinuous speech. We initially acquired testing over only one language model namely (HTML) using dataset of (63) words/phrases. The tested outcome satisfied the certain work and needs more extension in datasets as well as in proposed models/algorithms, found excellent Accuracy Rate and Less Error Rate of Each word and phrase. For language modeling n-gram language model was used.

REFERENCES:

[1] M. F. Brown, P. V. de Souza, R. L. Mercer, V. J. Della Pietra, J. C. Lai, "Class-Based *n*-gram Models of Natural Language," *Computational Linguistics*, vol. 18, December 1992.

[2] V. Lavrenko, W. B. Croft, "Relevance-Based Language Models," *Special interest Group on Informational Retrieval*, pp. 120-127, New York, USA, 2001.

[3] M. Larson, "Sub-Word-Based Language Models for Speech Recognition: Implications for Spoken Document Retrieval," 2001.

[4] NADEEM AHMED. KANASRO, H.U. ABBASI, M.R. MAREE, A.G. MEMON, "Speech Recognition based web scripting from predefined Context Free Grammar (Language Model & Grammar) programmed in Visual Programming and Text Editor," Sindh Univ. Res. Jour. (Sci. Ser.) Vol.45 (3) 634-639, September- 2013

[5] Mary Harper, "The Automatic Speech recognition In Reverberant Environments (ASpIRE) Challenge," in Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2015.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement based on Deep Neural Networks," IEEE Signal processing letters, vol. 21, no. 1, Jan 2014.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Global Variance Equalization for Improving Deep Neural Network based Speech Enhancement," in IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), 2014, pp. 71–75.

[8] Anuroop Sriram, Heewoo Jun, Yashesh Gaur, and Sanjeev Satheesh, "Robust speech recognition using generative adversarial networks," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5639–5643.

[9] Tian Tan, Yanmin Qian, Hu Hu, Ying Zhou, Wen Ding, and Kai Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 8, pp. 1393–1405, 2018.

[10] M.R JAMALI, A.G. MEMON, M.R. MAREE, "Security issues in data at rest in a non-relational Document Database," SindhUniv. Res. Jour. (Sci. Ser.) Vol. 52 (03) 279-284 (2020) http://doi.org/10.26692/sujo/2020.09.41

[11] Natsue Yoshimura, Atsushi Nishimoto, Abdelkader Nasreddine Belkacem, Duk Shin, Hiroyuki Kambara, Takashi Hanakawa, and Yasuharu Koike, "Decoding of covert vowel articulation using electroencephalography cortical currents," Frontiers in neuroscience, vol. 10, pp. 175, 2016.

[12] Pradeep Kumar, Rajkumar Saini, Partha Pratim Roy, Pawan Kumar Sahu, and Debi Prosad Dogra, "Envisioned speech recognition using eeg sensors," Personal and Ubiquitous Computing, vol. 22, no. 1, pp. 185–199, 2018.