International Research Journal of Computer Science (IRJCS)
Volume 1 Issue 3 (November 2014)
ISSN: 2393-9842
www.irjcs.com

Isolated English Words Recognition Spoken by Non-Native Speakers

Mr.V.K.Kale[1], Dr.R.R.Deshmukh[2], Dr.G.B.Janvale[3], Mr.V.B.Waghmare[4], Mr.P.P.Shrishrimal[5]
[1,2,4,5] *Department of Computer Science and Information Technology,*
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, 431004 (MS) India
[3]*Symbiosis Centre for Information Technology,*
Symbiosis International University, Pune, 412115 (MS) India

Abstract: - The aim of this experiment is to recognize English Isolated spoken words by different male and female non-native English speakers from Marathwada region of the Maharashtra state. The linear Predictive Coding (LPC), and Mel Frequency Cepstral Coefficients (MFCC), of the audio signal, has been used as a features, which are subsequently used for testing the samples. Classification is done by using Confusion Matrix. The 61.40% and 95.75% recognition rate has been achieved by LPC and MFCCs features, when the proposed approach is tested using a dataset of 1000 speech samples.

Keywords: - **Speech Recognition, LPC, MFCC.**

I. INTRODUCTION

Speech recognition is a process to recognize utterances automatically in computer system. In this process, computer captures the spoken word of speech which is recorded through microphone [1]. Continuous speech recognition systems have been developed for many applications, often using commercial speech recognition software. However, high performance and robust isolated word recognition, particularly for the letters of the alphabet recognizer and for digits, is still useful for many applications such as recognizing telephone numbers, spelled names and address [2]. The variations in accent affect on recognition rate. The accents are varying due to age, gender, and different geographical areas. The speech recognition also face changes of recognition accuracy affected by varying speech signals depend on database of Male and Female speaker, Age group, speaking style , environmental. These are the problem to face in speech recognitions. We have tried to find out the recognitions rate in different Non –Native speakers of English. All these constrain can be eliminated through speech recognition by designing appropriate algorithms. For the recognition, research has used two feature extraction techniques i.e. Linear Predictive Coding (LPC) [3] and Mel Frequency Ceptstral Coefficients (MFCCs). A Confusion Matrix is used to recognize respective words and finding accuracy of Non native speaker's .Then compare both technique results which is having higher recognitions accuracy.

II. ADVANCES IN ISOLATED SPOKEN WORDS RECOGNITIONS

In speech recognitions, there are so many different methodologies which have been proposed for isolated word speech recognition. These methods can usually be grouped in two classes: speaker-dependents and speaker-independents. Speaker dependent methods usually involve training a system to recognize each of the vocabulary words uttered single or multiple times by a specific set of speakers while for speaker independent systems such training methods are generally not applicable and words are recognized by analyzer their inherent acoustical property[4].

Various feature extracting techniques have been used singly or in combination with others i.e. Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCCs) and a combination of several features as formant frequency and Zero Crossing Rate (ZCR) Discrete Wavelet Transform (DWT). [5]

Lot of research work completed for accepting of foren language English accent. In 1988, Fledge investigates the factor affecting on accents in English language. From these studies, it was determined that each person develops a speaking style up to the age of 12, which consists of phoneme production, articulation, tongue movement and other physiological phenomena related to the vocal tract [6]. On –Native speakers learn this speaking style when learning a second language and therefore alternate phonemes from their native language when they come across a new phoneme in the second language [7]. It has been shown that is a challenging problem for speaker learning a second language in another country. During this period, there is fast pronunciation improvement. Fledge, showed that pronunciation score for non-native speakers living in the United States from one to five years were not significantly special [8]. In the remainder of this section, a series of clarification regarding accent and phonemic content will be considered. The intention here is to show that accent affects prosodic structure as well as aspects of phoneme. Christ scientist in 1964 investigates the sounds that highlight on-native speaker accent during English speech production. For example the "AE" sound' as does not exist in most other languages. In American English, there are twelve principal vowels [9]. Phoneticians frequently recognize a thirteenth vowel called the "schwa" vowel. It is sometimes called a "degenerate vowel" since other vowels drop towards this neutral vowel when spoken fast in the course of continuous speech. Since it is substitute so liberally, the non-native speaker finds the schwa to be one of the most difficult sounds of American English [10]**.** The schwa appears in the initial, medial and final position of many word units in American English [11]. Many secondary stressed syllables have neutralized vowels which approach the "schwa" position.

The four characteristics of defective expression are evident for this sound among on-native English speakers; the schwa sound is heard as an addition, distortion, omission or a substitution when foreign accent is present. Among the second language learner, sound substitution is the most common problem for the schwa. In American English, for words that begin with "a" or "un" such as above and unknown, the tendency of the non-native speaker is to substitute the "AA" sound for the most important vowels [12]. The medial substitution in words such as uniform, disappear, disappoint, disability is another source of problem. Other words which include the schwa sound are laboratory, president and communication. The word initial and word final additions of the schwa are also common among on-native speakers. For example, Spanish speakers add a schwa sound to the beginning of words that begin with the "S"sound [13]. On the other hand, when attempt to produce what the second language hears as the final sound in such words as bag the speaker adds a voiced release to the end. The importance of these initial and final additions becomes a vivid indicator of foreign accent [14].

A. English Language Phonetics:

A fundamental characteristic unit of a language is a phoneme. Different languages contain different phoneme sets. Syllables contain one or more phonemes, while words are formed with one or more syllables, concatenated to form phrases and sentences [15]. One broad phoneme classification for English is in terms of vowels, consonants, diphthongs, and semi vowels. Phonetics is the scientific study of speech. The central concerns in phonetics are the discovery of how speech sounds are produced, how they are used in spoken language, how we can record speech sounds with written symbols and how we hear and recognize different sound [16].

B. Accents :

Accent is one of the most important characteristic of speakers. It is manner of pronunciation of a language. In case of the speakers are non native, carry the intonation phonological process and pronounces rules from their mother tongue in their English speech [17]. The word is used in two different senses. The way, accent refers to importance given to a language unit, usually by the use of pitch. For example, in the word 'potato' the middle syllable is the most salient; if you say the word on its own you will probably produce a fall in pitch on the middle language unit, making that syllable accented [18]. In this sense, accent is distinguished from the more general term stress, which is more often used to refer to all sorts of importance or to refer to the effort made by the speaker in producing a stressed language unit. And second, accent also refers to a particular way of pronouncing: for example, you might find a number of English speakers who all share the same grammar and vocabulary, but pronounce what they say with different accents such as European country [19].

III. DEVELOPMENT OF SPEECH CORPUS FROM NON-NATIVE SPEAKERS OF ENGLISH

For development of a Speech database, the basic requirement is the grammatically correct Text corpus which would be recorded from various speakers. The text corpus should be correct in terms of composition and grammar.

A. Speaker Selection

All the speakers were the college students, including 10 males and 10 females between the age groups of 18 to 25 from Aurangabad city of Maharashtra state. They were given some English words to pronounce. Before that, they were trained and comfortable with speaking the English language with datasets.

B. Data Collection

The isolated English words were recorded through normal microphone. The isolated words were selected as a dataset which are frequently used in communication. The 'PRAAT' software was used for recording database with the sampling rate of 44 kHz and 16 bit. Each sample was recorded ten times. The distance between mouth and microphone was adjusted nearly 30 cm. The recorded data were stored in WAV file format,

Database Corpus **:**

The development of database corpus was developed. Firstly, we have set up the words corpus like, Hello, Please, Help, Ok, Come, By, Yes, Sad, Happy, and New. These utterances were recorded through PARRAT software. The 25 Males and 25 Females of Non-Native English speakers participated in the experiments. The total number of occurrences of each word and each speaker is 10, and the total number of words is also 5000. This corpus was used for features extractions.

IV. FEATURE EXTRACTION TECHNIQUES

A. Preprocessing:

The preprocessing step we record the isolated word speech sample using PARAT software and stored in WAV file format. Then recorded sample process for remove noise using cool editor software. Last stage sends process of data sample for feature extraction experiment. The isolated database of Non-Native was used in all experiments.

B. Linear Predictive Coding (LPC)

Linear prediction is a good tool for analysis of speech signals. Linear prediction models the human vocal tract as an infinite impulse response system that produces the speech signal.

In speech coding, the success of LPC have been explained by the fact that an all pole model is a reasonable approximation for the transfer function of the vocal tract [20]. All pole models are also appropriate in terms of human hearing, because the ear is more sensitive to spectral peaks than spectral valley [21]. Hence, an all pole model is useful not only because it may be a physical model for a signal, but because it is a perceptually meaningful parametric representation for a signal [22]. Advantages of LPC as below

- LPC provides good model of speech signal.
- The way in which LPC is applied to the analysis of speech signal leads to a reasonable source-vocal tract separation.
- LPC is analytically tractable model. The method of LPC is mathematically precise and straight forward to implement in either software in hardware.

Disadvantages of LPC the Linear Prediction models the input signal with constant weighting for the whole frequency range. However, human perception does not have constant frequency perception in the whole frequency range [23]. For example, in general, low frequencies are perceived with higher accuracy than high frequencies Therefore, since LP treats all frequencies equally, effort is wasted on high frequencies while important information in the low frequencies is discarded. Among others, linear predictive models known as WLP, FWLP, DAP, PLP, LPES, LPLE and SLP.

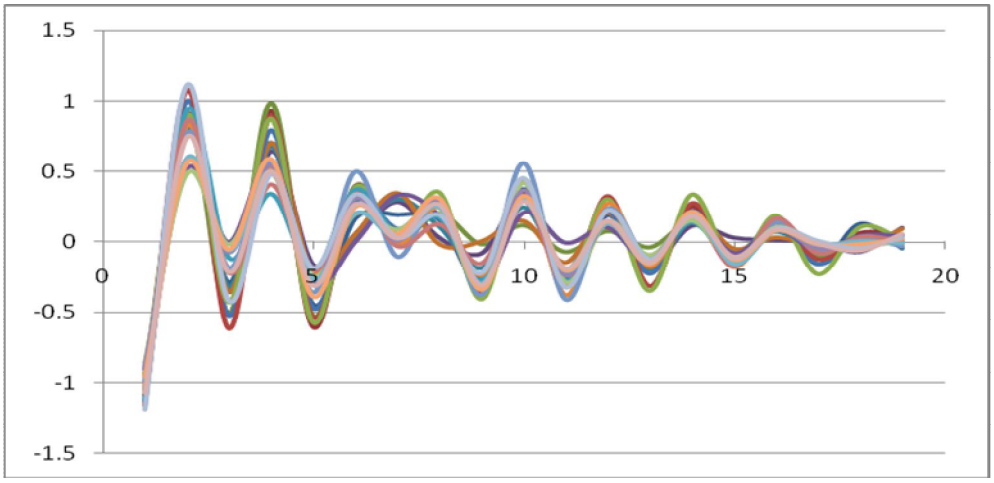

Fig. 1: Average LPC features of a spoken word 'By'

A. Mel Frequency Cepstrals Coefficients (MFCC)

The speech signal is often assumed to be the output of an LTI system; i.e., it is the convolution of the input and the impulse response. As are characterizing the signal in terms of the parameters of such a model, we must separate source and the model filter [24]. In ASR the source fundamental frequency and details of glottal pulse are not important for distinguishing different phones. Instead, the most useful information for phone detection is the filter, i.e. the exact position and shape of the vocal tract.

If we knew the shape of the vocal tract, we would know which phone was being produced. To separate the source and filter vocal tract parameters efficient mathematical way is cepstrum.The cestrum is defined as the inverse DFT of the log magnitude of DFT of the signal [25]. The cepstral coefficients have the extremely useful property that the variance of the different coefficients tends to be uncorrelated. This is not true for the spectrum, where spectral coefficients at different frequency bands are correlated [26]. The fact that cepstral features are uncorrelated means that the Gaussian acoustic model doesn't have to represent the covariance between all the MFCC features, which hugely reduces the number of parameters [27].

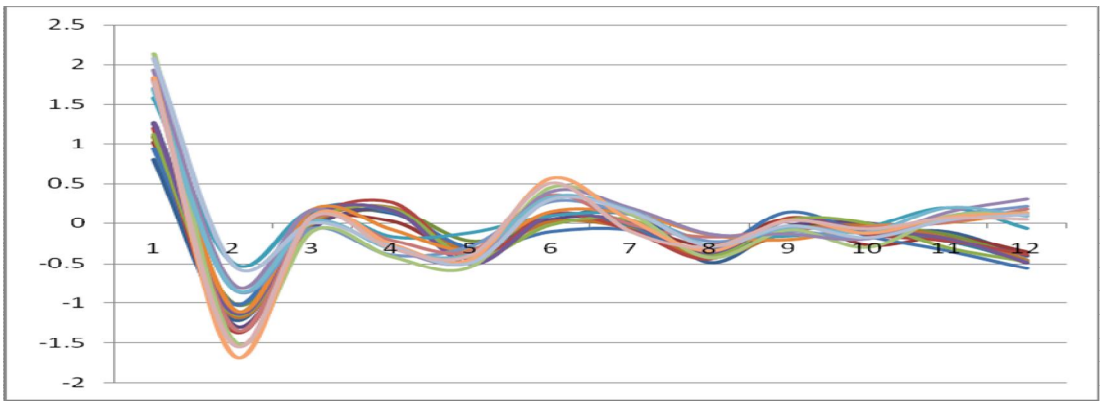

Fig. 2: Average MFCC features of a spoken word 'Come'

Since the MFCC is the most popular feature extraction technique for ASR, the steps involved in extraction of MFCC is explained below in detail.We have tested our dataset with LPC and MFCCs algorithms and have got 18 parameters for LPC and 12 coefficients for MFCCs as shown in figure 1 and 2.

TABLE I CONFUSION MATRIX OF LPC FEATURES

	'Hello'	'Ok'	'By'	'New'	'Come'	'Help'	'Happy'	'Sad'	'Please'	'Yes'	Total Test Sample	Recognition In %
'Hello'	16	0	1	0	0	1	2	1	0	3	24	66.66
'Ok'	1	16	0	5	0	2	0	0	0	0	24	66.66
'By'	2	0	18	0	2	0	0	0	1	7	30	60.00
'New'	4	5	1	17	0	0	2	0	1	1	31	54.83
'Come'	0	0	9	0	15	0	0	1	0	1	26	57.69
'Help'	3	0	0	0	0	10	2	0	2	3	20	50.00
'Happy'	0	0	0	0	0	2	13	1	6	0	22	59.09
'Sad'	1	0	4	0	0	0	0	14	0	7	26	59.09
'Please'	4	0	0	0	0	4	0	0	14	1	23	60.86
'Yes'	1	0	1	0	1	0	1	1	0	19	24	79.16
Total No of Samples											250	61.40

TABLE II CONFUSION MATRIX OF MFCCS FEATURES

	'Hello'	'Ok'	'By'	'New'	'Come'	'Help'	'Happy'	'Sad'	'Please'	'Yes'	Total Test Sample	Recognition In %
'Hello'	12	0	0	0	0	0	0	0	0	0	12	100
'Ok'	0	10	0	0	0	0	0	0	0	0	10	100
'By'	0	0	13	0	0	0	0	0	0	1	14	92.85
'New'	1	0	0	13	0	0	0	0	0	0	14	92.85
'Come'	0	0	0	0	9	0	0	0	0	0	9	100
'Help'	0	0	0	0	0	9	0	0	1	1	11	81.81
'Happy'	0	0	0	0	0	0	9	0	0	0	9	100
'Sad'	0	0	0	0	0	0	0	9	0	0	9	100
'Please'	0	0	0	0	1	0	0	0	9	0	10	90
'Yes'	0	0	0	0	0	0	0	0	0	2	2	100
Total No of Samples											100	95.75325

V. CONCLUSION

The isolated spoken English utterances are collected from ten non native speakers of English. The MFCCs and LPC are computed for these dataset. We have got 95.75 % recognition for MFCCs and 61.40 % for LPC through confusion matrix. . Results obtained by MFCCs show that this techniques straightforward, efficient.

ACKNOWLEDGMENT

This work is supported by University Grants Commission (UGC), New Delhi as a major research project. The authors would like to thank the Dr.Babasaheb Ambedkar Marathwada University Aurangabad, Authorities and Department of Computer Science and IT, for providing the infrastructure to carry out the research and participant who involved in the experiments.

REFERENCES

[1] Tristan Kleinschmidt, Michael Tason, Eddie and Sridha Sridharan, "The Australian English Speech Corpus for In-car Speech Processing", IEEE ICASSP 2009, 978-1-4244-2354-5/2009
[2] Yen – Minkhaw and Tien-Ping Tan, "Pronunciation Modeling for Malaysian English", International Conference on Asian Language Processing, DOI 10.1109/IALP 2012.72, 978-0-7695-4886-9/12. 2012.
[3] Liu Xiao-Feng, Zhang Xue-ying, and Duan Ji-Kang , "Speech Recognition Based on Support Vector Machine and Error Correcting Output Codes", IEEE Computer Society, 978-0-7695-4180-8/10, (2010) DOI 10.1109./PCSSPA 2010.
[4] Sheguo Wang, Xuxiong Ling, Fuliang Zhang, and Jianing Tong, "Speech Emotion Recognition Based on Principal Component Analysis and Back Propagation Neural Network", IEEE Computer Society, 978-0-7695-3962-1/10, DOI 10.1109/ICMTMA 2010.
[5] Ren Wenxia, Zhaug Huili, and Lv Wenzhe, "Realization of Isolated – Words Speech Recognition System", IEEE Computer Society, 978-0-7695-3614-9/09, 2009, DOI 10.1109/PACCS 2009..
[6] Albert Croll Baugh and Thomas Cable, "A History of the English Language", Rowledge, , ISBN 0415093791, 9780415093798, 1993.
[7] Joshi M., Iyer M. and Gupta N., "Effect of Accent on Speech Intelligibility in Multiple Speaker Environment with Sound specialization", IEEE Computer Society, ISBN 978-0-7695-3984-3/10 2010.
[8] Shamalee Deshpande, Sharat Chikkerur, and Venu Govindaraju, "Accents Classification in Speech", IEEE Computer Society, 0-7695-2475-3/05 2005.
[9] Elizabeth Hume and Keith Johnson, "The Impact of Partial Phonological Contrast on Speech Perception", Proceeding in the 15th International Congress of Phonetic Science 2003.
[10] Chitralekha Bhat, K.L. Srinivas, and Preeti Rao, "Pronunciation Scoring for Indian English Learners using a phone recognition system", ACM 978-1-4503-0408-5/10/12, 2012.
[11] Agenes Stepheson, Hong Jiao and Nathan Wall, "A Performance Comparison of Native and Non – Native Speakers of English on an English language Proficiency Test", Pearson Technical Report, August 2004.
[12] Yemi Olagbaju, Buket D. Barkana, Navarun Gupta, "English Vowel Production by Native Mandarin and Hindi Speakers", IEEE Computer Society, ISBN 978-0-7695-3984-3, 2010
[13] Laura Tomokiyo, "Linguistic Properties of Non- Native Speech", ISBN 0-7803- 6293- 4, 1335-1338, Vol. 3, June 2000.
[14] Qingqing Zhang, Ta Li, Jielin Pan and Yonghong Yan, "Non native Speech Recognition based on Stste – Level Bikingual Model Modification", IEEE Computer Society, ISBN 970-07695-3407-7/08, 2008
[15] Bishnu Prasad Das, Ranjan Parekh," Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers". International Journal of Modern Engineering Research, Vol.2, Issue.3, pp-854-858 ISSN: 2249-6645, 20.
[16] Jhing-Fa Wang and Shi-Huang Chen, "Wavelet Transforms for Speech Signal Processing", Journal of the Chinese Institute of Engineers Vol.22 No. 5, PP. 549-560.
[17] Liu Wai and Pascule Fung, "Fast Accent Identification and accented Speech Recognition", 0-7803-5041-3/99. IEEE 1999
[18] Pierre-Yves Oudeyer, "The Production and Recognition of emotion in speech: Features and algorithms", International Journal of human computer studies Elsevier Science doi 10.1016/S1071-5819(02)00141-6.\ 59(2003) 157-189, 2002
[19] Xia Mao, Lijiang Chen and Bing Zhang, "Mandarin Speech Emotion Recognition based on a hybrid of HMM/ANN", International Journal of Computers. Issue 4, Volume 1, 2007.
[20] Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, and Nikos Fakotakis, "Comparison of Speech Features on the Speech Recognition Task", Journal of Computer Science 3 (8): 608-616, , ISSN 1549-3636, 2007
[21] Hossan M.A., "A Novel Approach for MFCC Feature Extraction", 4th International Conference on Signal Processing and Communication System pp 1-5, December, 2010.
[22] Ganesh B. Janvale, Vishal B. Waghmore, Vijay Kale, and Ajit S. Ghodke, "Recognition of Marathi Isolated Spoken Words Using Interpolation and DTW techniques", ICT and critical Infrastructure: Proceeding of the 48th Annual of

computer society of India Vol I. Advances in Intelligent system 3-319-03107-3-3, Print ISBN 978-3-319-031066, Online ISBN 978-3-319-03107-1.,2013

[23] Vishal B. Waghmare, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, and Ganesh B. Janvale, "Emotion Recognition System from Artificial Marathi Speech using MFCC and LDA Techniques", Proceeding of International Conference on Advances in Communication, Network, and Computing.

[24] Mihalis Siafarikas, Iosif Mporas, Todor Ganchev and Nikos Fakotakis, "Speech Recognition using Wavelet Packet", Journal of Wavelet Theory and Applications, ISSN 0973-6336 Volume 2 No.1, 2008

[25] Felix Weninger, Jarek Krajewski, Anton Batliner, and Bjorn Schuller, "The Voice of Leadership: Models and Performances of Automatic Analysis in Online Speeches", IEEE Transactions on Affective Computing, Vol. 3, No. 4, October –Dumber 2012.

[26] James K. Tamgno, Etienne Barnard, Claude Lishou, and Morgan Richomme, "Wolof Speech Recognition Model of Digits and Limited-Vocabulary Based on HMM and ToolKit", IEEE Computer Society, 978-0-7695-4682-7/12, DOI 10.1109/ 2009.

[27] Zhao Lishuang and Han Zhiyan, "Speech Recognition System Based on Integrating feature and HMM", IEEE Computer Society, 978-0-7695-3962-1/10, DOI 10.1109/2009

[28] Rupayan Chakraborty and Utpal Garain, "Role of Synthetically Generated Samples on Speech Recognition in a Resource-Scarce Language", IEEE Computer Society, 1051-4651/10, DOI 10.1109/2009

[29] Clarence Goh Kok Leon, "Robust Computer Voice Recognition Using Improved MFCC Algorithm", IEEE Computer Society, 978-0-7695-3687-3/09, DOI 10.1109/2009

