Pronunciation Variant and Substitutional error analysis for Improving Telugu Language Lexical performance in ASR system Accuracy

M. Nagamani, P.N. Girija

Abstract— In this paper we describe the error analysis in Automatic Speech Recognition system results. Substitutional errors will cause the ASR system performance degrade when pronunciation variants will occur in decoding process by substituting different phonemes in place of correct phonemes. This will increase the Word Error Rate(WER). When ASR systems are defined for specific languages, and phone set will be independent of language then any phone set which will cover the target language phonemes will be adapted. In this work Telugu language data is considered to train and test the ASR system. Sphinx Speech recognition engine will use the default CMU phone set for any language ASR system development. The phone set for CMU lexicon defined based on the American English. The same phone set is not sufficient to represent the Telugu language. The Telugu language is not stress timed but it is a syllable timed language. It required super set of CMU phone set. To achieve goal a new phone set derived to represent the Telugu Language sounds(phonemes). The Substitutional error analysis is done by comparing these two phone set for same data samples collected from Telugu language simple isolated words. The confusion matrix are considered for vowel and consonants separately to verify the more Substitutional phones in recognition process in different pronunciation variations occurred during the data sample collection. Applying data driven rules to the new derived phone set which is known as UOH phone set to decreasing the Substitutional errors.

Index Terms - Automatic Speech Recognition, Word Error Rate, UOH lexicon, Phonemes, Phone set, Substitutional errors, confusion matrix.

1 INTRODUCTION

Speech is a process used to communicate from a speaker to listener. Pronunciation relates to speech, and humans have an intuitive feel for pronunciation. For instance, people chuckle when words are mispronounced and notice when foreign accent colors a speaker's pronunciations (Strik J H, 1999).

The ultimate aim of ASR research is to allow a computer to recognize with 100% accuracy all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics and accent, or channel conditions.

Despite several decades of research in this area accuracy greater than 90% is only attained when the task is constrained in some way. Depending on how the task is constrained, different levels of performance can be attained; for example, recognition of continuous digits over a microphone channel (small vocabulary, no noise) can be greater than 99%. If the system is trained to learn an individual speaker's voice, then much larger vocabularies are possible, although accuracy drops to somewhere between 90% and 95%. For large-vocabulary speech recognition of different speakers over different channels, accuracy is no greater than 87%, and processing can take hundreds of times real-time [1].

A large vocabulary speech recognition is usually accomplished by classifying the speech signal into small sound units (or sub-word units), and then combining them into words, and eventually phrases and utterances. The glue that binds words to their corresponding sound units is the pronunciation model. The pronunciation model of a recognizer is usually specified as a pronunciation dictionary (also known as a pronouncing dictionary, or pronunciation lexicon),9 which is a list of words followed by acceptable pronunciations specified in terms of the phoneset of the recognizer Significant progress has been made towards identifying standards to achieve improvement of speech recognition accuracy goal. A wide variety of measures have been used, including measures like task success [8]. Other metrics evaluate user satisfaction in conjunction with task success. The research work propose a Usability Standard based on three factors

1) Accepting the speech signal in an optimal way
2) Assessing the Speech recognition task success
3) Assessing the user satisfaction in conjunction with the task success.

The research propose a procedure for accepting the speech signal based on Input Signal processing[5] which identifies spoken word for validating fast and slow .The research also measure task success and also user satisfaction in conjunction with task success. User satisfaction was calculated using questionnaires. Interaction between the user and the system was recorded to calculate the remaining two metrics. The research also proposes an efficient method to identify errors in recognition and repair procedures. In real time speech recognition application typically, where the confi-

- *NagaMani Molakatala is currently working as Asst. Professor in University of Hyderabad, India, PH-09966727247. E-mail: molakatala@gmail.com*
- *P.N.Girija, is currently working as Professor in University of Hyderabad, India, PH-04023134018. E-mail:pngsc@gmail.com*

dence level is low, systems will reject the recognition, and reprompt. If the system has a hypothesis, but is unsure as to its correctness, a confirmatory question is asked. Both strategies can be very frustrating to the user if they are used repeatedly. More sophisticated systems might proceed with an implicit confirmation, as an Example. In these cases, the system also has to allow for the user's protest when recognizing the next response, and to negotiate an appropriate correction of the error

The Causes of errors: Speech engines produce hypotheses by seeking to find the best word sequence for a given speech input that maximizes the words given some language model. [2] Speech recognition errors occur because the sounds in the utterance heard by the computer are dissimilar to its acoustic model and/or the language employed is not contained in the language model being used. Combined probability of the sounds being from the proposed words and the words being. The literature shows that there are many potential sources of differences in the acoustic domain. These include hyper articulation, pronunciation variation, cold speech, dysarthric speech, children's speech and noise in the signal. It seems errors revealed by implicit confirmation take longer to repair than those handled by explicit confirmation [3] and, in reality, very few real systems exhibit such sophistication; error handling and repair strategies adopted are generally quite simplistic, and sometimes poorly designed. The focus of our work is handling misrecognitions by solving two Problems: Error Recognition. – To classify hypotheses as correct or not, with a very high level of accuracy. Error Repair. – To repair such errors in a manner that does not frustrate or baffle the user.

2. TELUGU LANGUAGE

The evolution of Telugu [1] can be traced through centuries in terms of its form as well as its function. Although culturally Telugu is close to its southern neighbors -- Tamil and Kannada -- genetically, it is closer to its northern neighbors -- Gondi, Konda, Kui, Kuvi, Pengo and Manda. There is evidence to show that these languages were freely borrowed from Telugu even from the prehistoric period whereas borrowing between Telugu and Tamil and Kannada has been mostly during the historic period, i.e., post-5th century B.C. Its vocabulary is very much influenced by Sanskrit. In the course of time, some Sanskrit expressions used in Telugu got so naturalized that people regarded them as pure Telugu words. Some Kannada and Tamil words were also taken into Telugu language. The sounds of Telugu are represented by a visual symbol to each of these 57 where presently confined to 52 sounds. These 52 syllabic sounds represent vowels and consonants. But vowels do not always occur by themselves; they combine with consonants to give the different nuances of the consonant (e.g. ta, too, tee etc.). In such cases we generally add a vowel sign to the consonant. These vowel signs are 16 in number. The categories are vowels, vowel signs, consonants, semivowels, sibilants, and aspirates. Telugu is syllabic in nature - the basic units of writing are syllables. Since the number of possible syllables is very large, syllables are composed of more basic units such as vowels ("achchu" or "swar") and consonants ("hallu" or "vyanjan"). The first 16 of Telugu alphabet are commonly called 'Acchulu'which can also be referred to as'Praanaaksharamulu' or 'Swaramulu'. Consonants or 'hallulu' in consonant clusters take shapes which are very different from the shapes they take elsewhere. Consonants are presumed to be pure consonants, that is, without any acchu (vowel sound) in them. However, it is traditional to write and read hallulu (consonants) with an implied 'a' vowel 'acchulu' sound. When 'hallulu' combine with other 'acchulu', the vowel (acchu) part is indicated orthographically using signs known as 'Gunintaalu' or 'maatras'. The shapes of 'Gunintaalu' are also very different from the shapes of the corresponding vowels.

2.1 Pronunciation Variants

The amount of pronunciation variation present in the speech under study has gradually increased. Pronunciation variation will deteriorate the performance of an ASR system if it is not well accounted for[3]. If the words were always pronounced in the same way, automatic speech recognition(ASR) would be relatively easy. However, for various reasons words are almost always pronounced differently.

The one-pronunciation-per-word model, however, is often too rigid to capture the variation in pronunciations seen in speech data. Often, phones are changed from the canonical ideal in continuous speech; this means that the acoustic realization of phones will not match the acoustic models corresponding to the individual HMM states well. The most important sources of pronunciation variation will be based on Intraspeaker variation and interspeaker variations.

Inter speaker variation refer to the fact that the same speaker can pronounce the same word in different ways depending on various factors. The same speaker can pronounce the word in different way in isolated and connected or continuous speech. Because in connected speech all sort of interactions may take place between words, which will result in the application of various phonological process such as assimilation, co-articulation, reduction, deletion and insertion. The degree to which these phenomena occur will vary depending on the style of speaking the speaker.

The speech can be varied by many factors like during its recording, the format used for recording, system which accept the format, the microphone, operating system in which the tools we are using. Before giving to the training system record speech need to refine. The present work we used three ways to record the data. Praat tool, windows recorder and

Linux command to record the speech samples which are used in this developed system. The format used for recording in all three is 16000Hz, 16bit mono PCM coded wav format. In Linux system fixed time for all the recorded samples where as praat and windows recorder based on sample recorded length defined[6]. If more silence is pre and post included in Linux recorded samples we remove silence by using Praat tool. With this silence and noise removal most of the deletion and insertion errors are rectified. Even few substituted errors also corrected by simple processing the speech record data. Sometimes if still the train data is not convergent then re-recording the speech samples to minimize the error rate in ASR system.

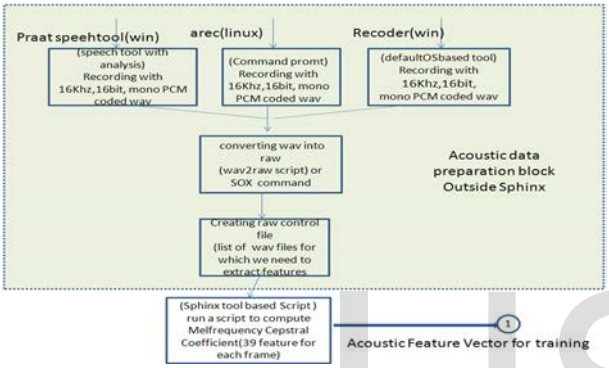

Fig1: Procedure for Acoustic Level Pronunciation adaptation.

2.2 Pronunciation adaptation at Lexical level

During human evolution, the vocal organs adapted themselves in such a way that producing speech sounds became possible(which was not the original function of the vocal organs)[7]. Simultaneously, the system adapts itself in order to be able to process those speech sounds. Adapting automatic speech recognizers in order to improve their processing of those speech sounds that humans learned to produce and understand throughout a long period of evolution. The adaptation types are speaker adaptation, lexicon/pronunciation adaptation, language model adaptation, database/environment adaptation, oise/channel compensation. The acoustic models and language models are generally the output of an optimization procedure, whereas in case of lexicon it is not. The lexicon, together with a corpus, is usually the input, and not the output, of a training procedure. Furthermore, the lexicon is the interface between the words and the acoustic. The lexicon defines the acoustic-phonetic units used during recognition, which are usually phones. The pronunciations present in the lexicon are transcriptions in terms of these acoustic-phonetic units. The lexicon can be adapted by adding new words to the lexicon, in order to reduce the out-of-vocabulary(OOV) rate. This will certainly lower the word error rate. Other way to deal with the kind of lexicon adaptation that is necessary to model pronunciation variation i.e. pronunciation adaptation at the lexical level. The need for modeling pronunciation variation in ASR originates from the simple fact that the words of the language are pronounced in many different ways as a result of variations in speaking style [8], degree of formality[9] ,[10]

2.3 Experimental setup

The Speech Recognition system in these experiment is used as Sphinx III continuous speech recognition system in single machine tar mode. Where training and decoding and front end modules are combined in a single package. Before experiment start we need to prepare the date for setting up training system and decoding system. There are five essential files need to create to setup the ASR system training. Speech corpus(audio files), its transcription, dictionary file and phonelist file. A control file needs to create which contain the audio feature extracted file list. The audio file is in wav format with 16 kHz and 16bit mono format of speech samples are collected for different experiments with speaker and gender variant factors. There are around 20K speeches samples are used for the experiments. The male and female age group between 20-40 ages. The speech samples are recorded by using head mounted microphone in Lab environment where system noise and room noise is common for all the speech samples. The Isolated words are uttered by the speakers. The same data is used for the training and testing of ASR system which is running on the Linux operating system Environment.

2.4 Pronunciation model adaptation

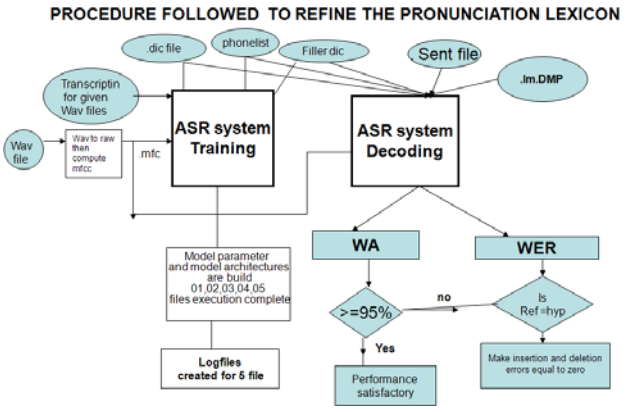

Fig 2: Lexicon refinement procedure through ASR system

The above Figure describes the adaptation techniques applied in different levels of ASR system. The colored blocks can be modified by user and remaining are system modules. The training and decoding process is continued with modification still the ASR system reaches the 95% and above Word Accura-

cy i.e. the word error rate is 5% and below level. The procedure will start by analysis the Errors(Insertion, Deletion and Substitution) of result. The Hypothesis, reference and aligned words also taken count. If Deletion and Insertion errors rectified all three i.e. Hypothesis, reference and aligned words are equal in result of ASR decoding process. Substitution errors are rectified by looking into Lexicon, transcription and acoustic signal i.e. wave files.

2.5 Results analysis of Isolated words

The ASR system recognition is performed by taking the Vowel sounds, consonant sounds and Isolated words of different size of letters present in it. The experiments are carried out for different variation factors like speakers, gender with comparison of newly proposed UOH lexicon which is handcrafted and the CMU lexicon which we found in online by using American accent pronunciation of English phonemes adapted for the new language. The comparison table shows the word error rate and word accuracy with different color codes.

TABLE 1:
SPEAKER AND LEXICAL MODEL VARIATION OF ASR WORD ACCURACY(WA) AND WORD ERROR RATE(WER)

	665w	%WA		%WER	
	Speaker	UOH	CMU	UOH	CMU
1	SPK1-F	87.1	71.1	12.9	28.9
2	SPK2-F	88.1	63.2	13.1	38.6
3	SPK3-F	76.5	48.7	24.4	54.3
4	SPK4-F	86.9	65	14.1	36.7
5	SPK1-M	19.9	6.9	81.6	94.6
6	SPK2-M	84.4	64.2	15.6	35.8
7	SPK3-M	80.5	60.8	20.2	41.8

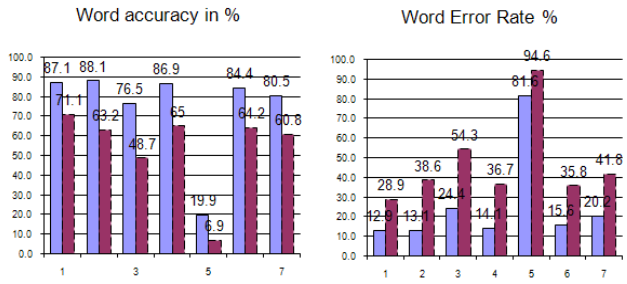

Fig 3: WA and WER comparison for 7 speaker's data with CMU and UOH Lexicon.

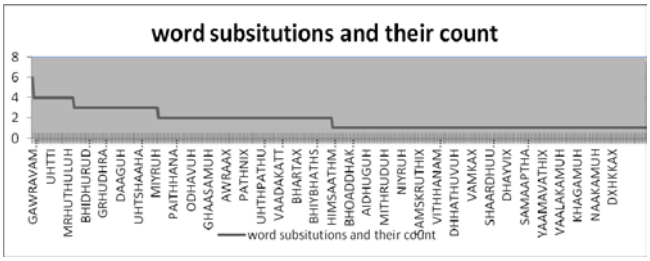

Fig4. of Substitution words in confusion pair.

The figure 4: (substitution words in confusion pair) shows the list of words and with the number of time their substitution in error file. These are compared with CMU phonelist based lexicon to the UOH phonelist based lexicon used words. The two figures gives the analogy that the substitution errors are more in CMU phonelist based than that of UOH phonelist based lexicon. It also show not only number of time substitution but also more number of

words also confused in CMU based lexicon rather in UOH based lexicon. Our proposed phonelist and lexicon will give better performance as it taken language parameter to define the phone list. CMU lexicon is American accent English phone list directly applied for Telugu language. UOH phonelist is adaptive phone list for Telugu language phoneme by considering the language properties into account.

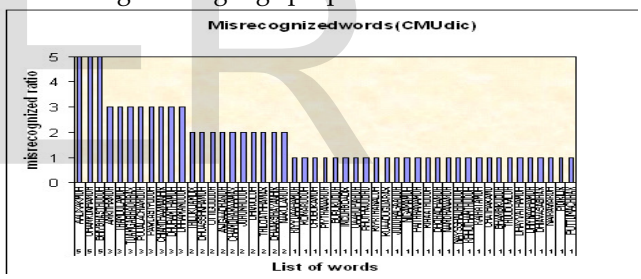

Fig 5. Mis recognized words list in CMU lexicon based ASR system error files

The figure 5 shows the misrecognized words when we are using the CMU lexicon. The confusion of phonemes causes the system to recognize wrong word in place of actual word which is given in reference word list. The misrecognized words are available in hypothesis file which is system generated after decoding the speech signal. Few list only shown along with how many times the misrecognized word is repeated.

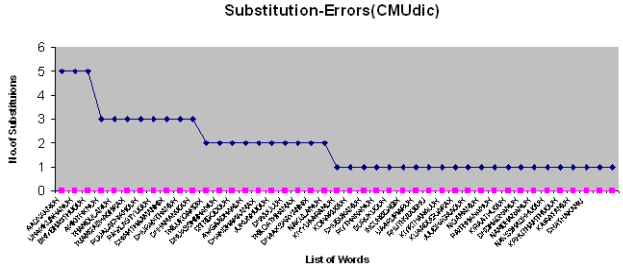

Fig6: No. of Substitution Errors in CMU lexicon based ASR

system error files.

The figure 6 shows the no. of times the wrong word is replaced with the right word. The maximum no. of times the wrong word inserted is 5 and minimum number is 1. The graph shows few list of words that are substituted because of the system recognize the wrong phoneme in place of right phoneme with the influence of pronunciation variations cause by various reasons mention in section 3.

The confusion pair of word will give the list of misrecognized because of substitution, insertion and deletion of phones in a given word shown in Figure12 and 13. The Sphinx system will give only word level error files. These words are mapped from the Lexicon to know the are the phones are matched and which are one lead to error which is shown in figure11.

The Figure.6 give the list of substituted words and their corresponding phone so that we can draw the confusion matrix for phoneme level of given word and know reason for error based one the phoneme classification of Telugu language. The Figure 7 & 8 shows the vowel substitutions and Figure. 9 & 10 shows the consonant substitutions.

2.5.1 Confusion matrix for Vowels

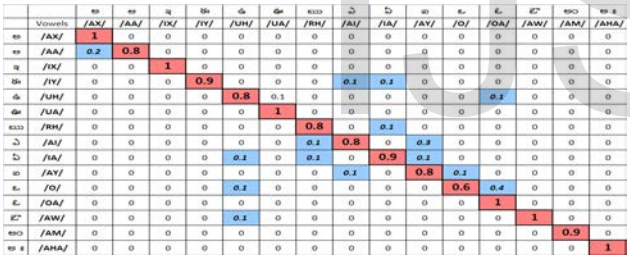

Fig7. Confusion matrix drawn for Vowels in 10 experiments

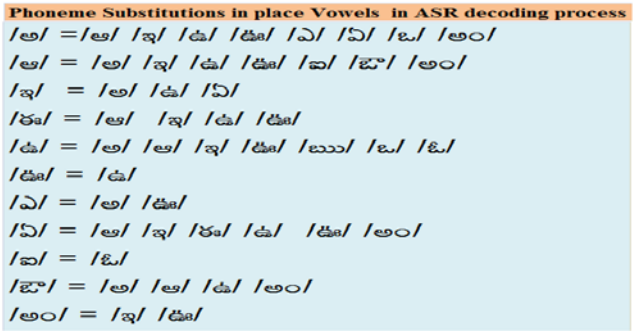

Fig8. Vowel and substitutions of wrong phoneme(vowels)

2.5.1.1 Confusion matrix for Consonants

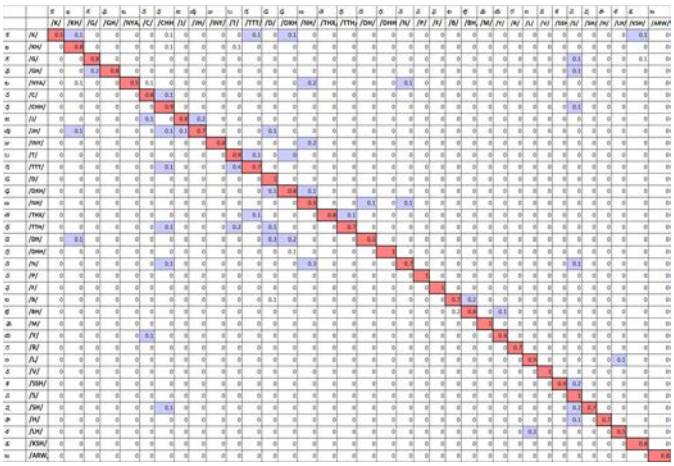

Fig9: Confusion matrix for Telugu Consonants.

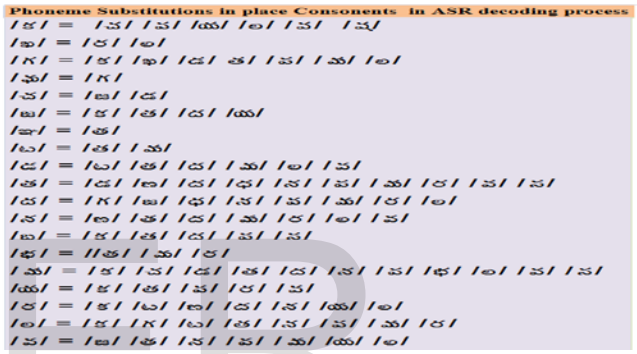

Fig10. Consonant and error substitution in terms of Consonant.

2.5.1.2 Sub subsections

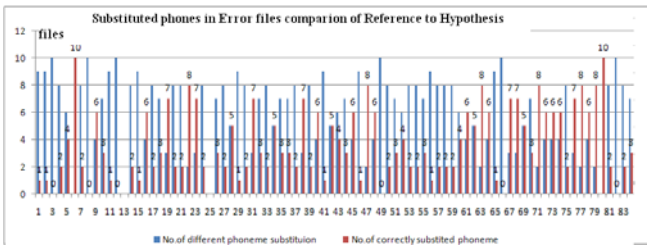

Fig11. Error phones in Hypothesis

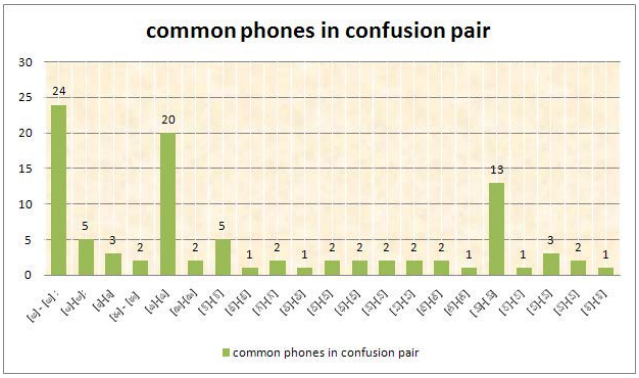

Fig12. The list of common phones in confusion pair list of UOH lexicon based ASR system.

The Figure 11 shows the Number of error files in hypothesis and Figure.12 shows the Common phones in confusion pairs which is compared from hypothesis and reference words and corresponding phones got from the lexicon.

3. CONCLUSIONS

In this paper the study of Indian language like Telugu language usage in ASR system and reasons for different phonemes and phonelist used for building lexicons discussed. The two languages phonelist used for building Telugu language ASR system described. The analysis is done how the Telugu phoneme based phonelist used in building Lexicon of TASR system improving the word Accuracy by reducing the confusion words which are causing the performance degradation of ASR system. The experiment result shows that by using UOH based lexicon used in Isolated word recognition system shows the improvement of 10% to 25 % increase in word Accuracy in comparison with the CMU based lexicon

4. ACKNOWLEDGMENTS

Our thanks to the entire speech (Voice) recording contributor for spending time and voices to record the speech corpus. The M. Tech and MCA students and other speakers. Who contributed for, Native male and Non- native male voice.

REFERENCES

[1] Telugu Vaaramandi, http://teluguvaramandi.net/home.html

[2] Strik, H and C. Cucchiarini, "Modeling pronunciation variations for ASR: A survey of the literature", Speech communication, 29,1999, pp.246-255.

[3] Benzeguiba. M, De Mori. R, Deroo. O, Dupont. S, Erbes. T, Jouvet. D, Fissore. L, Laface. P, Mertins. A, Ris. C, Rose. R, Tyagi. V, Wellekens. C, "Automatic Speech Recognition and Intrinsic speech variation", ICASSP 2006,pp. V1021-1024.

[4] Jacob Benesty, M. M. Sondhi, Yiteng Huang, "Springer Handbook of Speech Processing", Springer-Verlag, Berlin, Heidelberg, 2008.

[5] "Automatic Speech Understanding" http://ewh.ieee.org/r10/bombay//news6/AutoSpeechRecog/ASR.htm

[6] Ms.E.Chandra , Influence of Acoustics in Speech

[7] Recognition for Orienteal Language Accepted by IJCPOL, World Scientific Publishing, March 2006.

[8] Chandra, E., Ramaraj, E., "Speech Recognition Standard Procedures, Error Recognition and Repair Strategies" Communication Technology, 2006. ICCT '06. International Conference on 27-30 Nov. 2006,pp 1 – 9.

[9] Helmer Strik, "Pronunciation adaptation at the lexical level", Proceedings ISCA ITRW Workshop Adaptation Methods for Speech Recognition, Sophia Antipolis, 2001pp.123—130.

[10] Eskenazi, M., "Trends in speaking styles research" in Proceedings of Eurospeech-93, Berlin, 1999, pp.501-509.