MINIMALLY BALANCED CORPUS FOR SPEECH RECOGNITION

Saad Irtza
Electrical Engineering Department
University of Engineering & Technology
Lahore, Pakistan
saad.irtaza@kics.edu.pk

Dr Sarmad Hussain
Center for Language Engineering
University of Engineering & Technology
Lahore, Pakistan
sarmad.hussain@kics.edu.pk

Abstract— *This paper reports the method of collecting minimally balanced corpus for speech recognition. Generally balanced corpora are used for training speech recognition systems. However, these balanced corpora are not optimal. The current paper demonstrates that these corpora can be reduced to a varied degree for various phonemes for developing a minimally balanced corpus. The experiments have been developed on ten speakers' speech data. Recognition accuracy and amount of training data of phonemes have been analyzed. The result for these speakers shows that different phonemes require a different amount of training data for optimal training.*

Keywords-component; *Minimally balanced corpus; Urdu speech corpus; speech recognition*

I. INTRODUCTION

Generally phonetically rich and balanced corpora have been used for training of Automatic Speech Recognition system (ASR) systems. Large amount of training data is required to develop the acoustic models that cover all phoneme contexts in sufficient amount. This speech data has to be recorded from different speakers to capture speaker and pronunciation variations. Tagging and transcription of such large corpora is a challenging task. This effort can be reduced by developing optimal corpora.

This paper describes the development of ASR system for ten speakers using Urdu speech corpus developed in [1]. The corpus contains both the read and spontaneous Urdu speech and is divided in two portions for training and testing purposes as described in [2]. Phoneme error analysis has been performed to explore the recognition issues. The current work further presents data which indicates that the balanced corpus being used for training the ASR can be further optimized across various phonemes.

The next section reviews the work done on Urdu ASR and error analysis techniques in other languages. The following sections give details of the experiments on the ASR development of Urdu, finally concluding with the experiment which indicates how phoneme analysis can further optimize training data. The paper finally discusses the implications and presents the conclusion.

II. LITERATURE REVIEW

Many speech corpora have been developed for different speech applications such as TTS [5] and ASR systems [3] [4] in different languages. These corpora have been collected in different contexts, e.g. isolated word [8], continuous speech [2] [4] [6] [7] [9], etc. Different algorithms have been developed to collect phonetically rich and balanced corpora in different languages e.g. Russian [9], Spanish [10], etc. These include greedy algorithms for maximal phonetic coverage in minimal data set [4] [5].

A variety of accuracy results have been reported for ASR systems for various languages. For example, distributed speech recognition system for English language over telecommunication channel has been analyzed with specified range of signal to noise ratio using HTK toolkit [12]. The training corpus consists of 8440 utterances from fifty two male and female speakers. Word accuracy has been found to be 87.81%. Another English ASR system developed on a subset of Malach corpus [13] shows that by improving signal to noise ratio an absolute improvement of 1.1% has been achieved [14].

Recent work on limited vocabulary, isolated word, speaker dependent ASR systems on Hindi [15] [16] also show promising results. Isolated words (0-9) Hindi (Swaranjali) speech recognition system has been developed for two speakers in [15]. Acoustic model has been trained from twenty utterances of a word for each speaker. Word accuracy for two speakers comes to be 84.49% and 84.27%. Speech recognition system on Hindi language has been developed in room environment for eight speakers on thirty isolated Hindi words. HTK toolkit has been used to train the acoustic word model. Overall word accuracy has been found to be 94.63% [16].

A few ASR systems have been developed for Urdu language on different corpora, e.g. [2] [17] [18] [19] [20]. Urdu speech recognition system has been developed for 81 speakers as reported in [2]. Acoustic model has been prepared on incremental basis in three stages by addition of data of two speakers. Three acoustic models have been tested on forty female, forty one male and eighty one combined speakers by using open source CMU sphinx toolkit. Word error rate for combined system has been found to be 60.2% [2]. In another study,

978-1-4673-2821-0/13/$31.00 ©2013 IEEE

Urdu speech recognition system has been developed based on artificial neural network, pattern matching and acoustic modeling approaches [17]. Viterbi algorithm has been used for decoding the model. Single speaker isolated digit recognition system has been developed for Urdu language by using back propagation neural network approach [18]. Small vocabulary automatic speech recognition system has been developed for Urdu language by using Sphinx4. Acoustic model has been prepared from 5200 utterances of speech data and fifty two isolated spoken Urdu words from ten speakers. The average word error rate comes to be 5.33% [19]. Automatic speech recognition system has also been developed for Urdu on single speaker medium vocabulary [20]. The acoustic model has been prepared from 800 utterances of read and spontaneous speech corpus combined in various ratios. Sphinx3 toolkit has been used to train and decode the model.

III. METHODOLOGY

Phonetically rich speech corpus developed in [2] has been tested by developing an ASR system [1] using 45 hours of read and spontaneous data recorded from 82 speakers. Error rate of this system is quite high. The aim of this study is to investigate the issues and address them to improve the accuracy results. Developing and collecting large amount of speech data from different speakers is a difficult task. As described in [11], the amount of training data of each phoneme at which maximum accuracy can be achieved is different for every phoneme. Therefore, the possibility to minimize the amount of training data for training the ASR system is explored.

In first experiment, baseline ASR system has been developed on ten male speakers. In second experiment, the testing data cleaning and enhancement based on error analysis technique proposed in [11] on single speaker data has been applied on the baseline ASR system. Phonemes and silence regions have been re-aligned with audio speech data based on phoneme error analysis in data cleaning process. In third experiment, one speaker data has been replaced with the one speaker data originally used in Experiment-2 of [11] to investigate impact of data cleaning in training process.

Finally, the fourth experiment investigates how data used for training the system may be reduced for various phonemes. Six different phonemes have been selected from the existing training corpus, which has been selected based on their frequency, ranging from low occurrence to high occurrence but with high accuracy of recognition. In addition, different variety of sounds has also been chosen including manner, place, voicing and nasalization. Original training data of each phoneme has been divided in six incremental steps for training, by removing the speech files from the corpus. Acoustic model has been trained for each of these six cases for one phoneme and recognition results of that phoneme have been analyzed. In this way saturation limits of training data have been determined for these six phonemes.

Amount of training and testing data for Experiment-1 is described in Table 1. This baseline speech data of ten speakers has been selected from Large Vocabulary Continuous Speech Recognition system for 81 speakers as developed in [2].

Table 1- Experiment-1 Data

No. of training utterances	1946
Recording time	250 minutes
Duration of Data	160 minutes
No. of test utterances	119
Read speech utterances	873
Spontaneous speech utterances	1073

In Experiment-2 data cleaning has been performed based on phoneme error analysis. Amount of data is presented in Table 2.

Table 2- Experiment-2 Data

No. of training utterances	1946
Recording time	250 minutes
Duration of Data	160 minutes
No. of test utterances	119
Read speech utterances	873
Spontaneous speech utterances	1073

Additionally, in this experiment transcription and tagging errors have been removed from testing data. Language model has been prepared from 81 speaker corpus [21]. Moreover, the non-speech areas have been identified in the segments automatically.

In Experiment-3, one speaker data from Experiment-1 has been replaced with the single speaker data described in [11]. Amount of data is presented in Table 3.

Table 3- Experiment-3 Data

No. of training utterances	1999
Recording time	262 minutes
Duration of Data	169 minutes
No. of test utterances	123
Read speech utterances	883
Spontaneous speech utterances	1116

On the basis of these experiments minimal corpus selection criteria is discussed in Experiment-4.

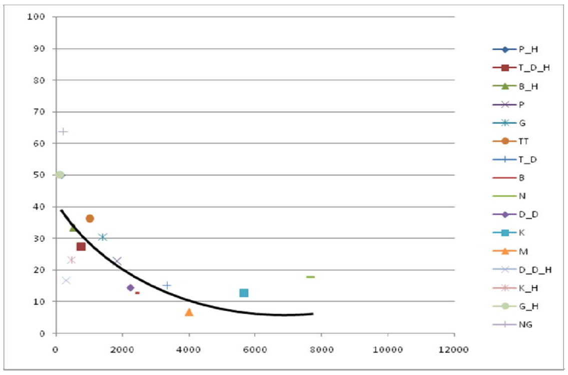
IV. EXPERIMENT-1 RECOGNITION RESULTS

Baseline recognition results are described in Table 4. These are based on the configuration parameters given in this table. Error analysis has been performed on the recognition results as described in [11]. In the following graphs percentage error rate have been plotted versus the amount of training data to analyze the phoneme error.

Table 4- Experiment-1 Recognition Results

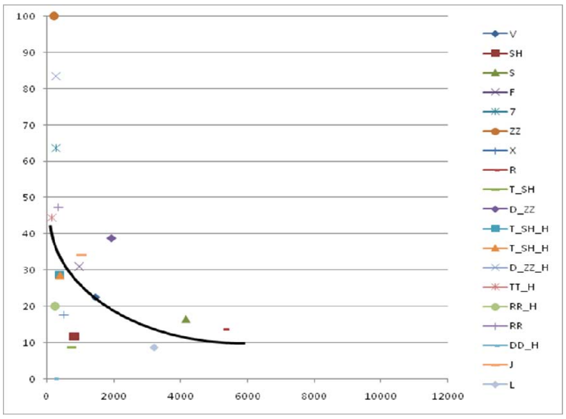
Beam width	1e-120
Language weight	20
Word error rate	63.58%

Phonemes have been divided in three categories on the basis of opening of vocal tract i.e. Stops (closed vocal tract), Vowels (Completely open vocal tract), Fricatives, Trills, Flap and Approximants (Partially open vocal tract).



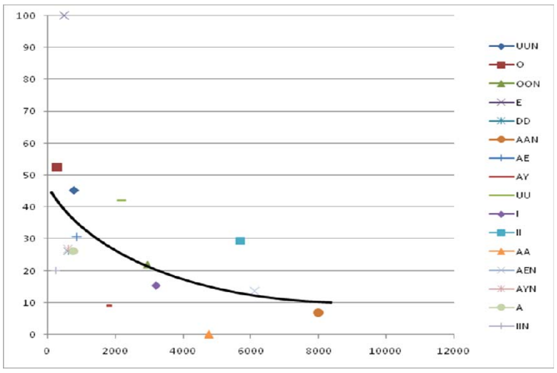
Amount of training data

Figure 1- Percentage error rate for stops



Amount of training data

Figure 2- Percentage error rate for other consonants



Amount of training data

Figure 3- Percentage error rate for vowels

A. Experiment-1 discussion

Figures 1, 2 and 3 show the distribution of phonemes (x-axis) against error scale (y-axis). This distribution region can be categorized in four regions 1) low training data, low error 2) low training data, high error 3) high training data, low error 4) high training data, and high error. Category 2 and 4 are challenging. Training data of phonemes that fall in second category may be increased to analyze the effect on accuracy. Speech data and transcription of phonemes that fall in fourth category may be analyzed to find the recognition issues.

A general trend has been observed from the three figures that phoneme accuracy eventually saturates as training data is increased. In addition, we can also see that some phonemes have very accurate recognition even with smaller amount of training data. Further, there are some phonemes which have low accuracy even with large amount of training data. Both these cases can be exploited for reducing training data.

V. EXPERIMENT-2 RECOGNITION RESULTS

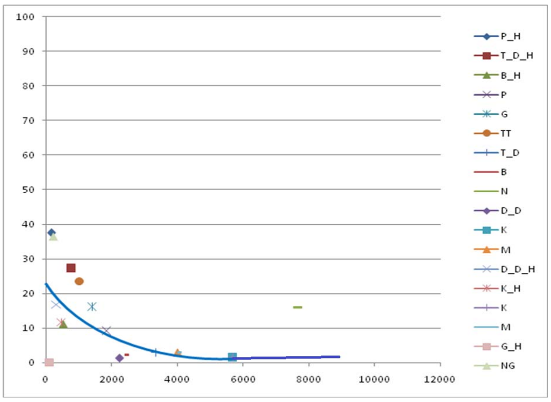
Improved recognition results are described in Table 5. These are based on the configuration parameters given in this table.

Table 5- Experiment-2 Recognition Results

Beam width	1e-120
Language weight	20
Word error rate	25.88%

A. Experiment-2 discussion

Figures 4, 5 and 6 show the improvement in phoneme error rates. Phoneme training data remains the same but error rate decreases due to cleaning of testing data. Figures 5 and 6 indicate that error rates of many phonemes have been decreased to 0%, e.g. F and ZZ in Figure 5, A and O in Figure 6. From Figure 4, error rate of phoneme G_H has been reduced to 0%. This phoneme has fewer amount of training data and nearly 50% error rate in Figure 1.



Amount of training data

Figure 4- Percentage error rate for stops

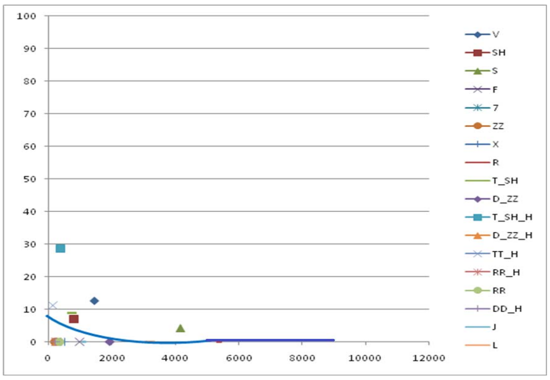


Figure 5- Percentage error rate for other consonants

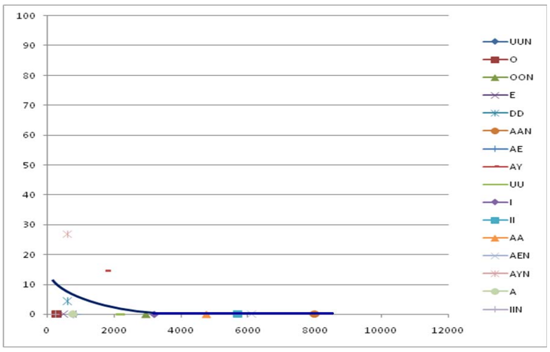


Figure 6- Percentage error rate for vowels

During cleaning of test speech data, noise and pronunciation problems have been found. Pronunciation problems have been found largely in spontaneous speech data, e.g. میں ("I") has been extended in duration but in transcription the extended period has been mapped to silence. Moreover, the phoneme error rate of N from 17.82% to 12.7%, because of wrong transcription of phoneme N, e.g. in the word ٹینجیبل ("tangible"). Such issues have been resolved. Moreover silence marker has been adjusted automatically using force alignment algorithm [22].

Similar improvement in training data could also improve accuracy of the recognition system.

VI. EXPERIMENT-3 RECOGNITION RESULTS

In this experiment, data of a speaker has been replaced with the cleaned training data of a speaker from Experiment-2 of [11]. Improved recognition results are described in Table 6. These are based on the configuration parameters given in this table.

Table 6- Experiment-3 Recognition Results

Beam width	1e-120
Language weight	20
Word error rate	23.21%

A. Experiment-3 discussion

Figures 7, 8 and 9 show the improvement in phoneme error rate. Effect of addition of cleaned, balanced training data can be analyzed from these figures. Training data of different phonemes have been increased, e.g. in Figure 7 training data of phoneme S has been increased from 4000 to 5500 and error rate reduced to 0%. Error rates of all phonemes in Figures 8 and 9 have been reduced to 0% except TT_H and DD respectively. Therefore, training corpus seems to be critical for better performance of ASR system.

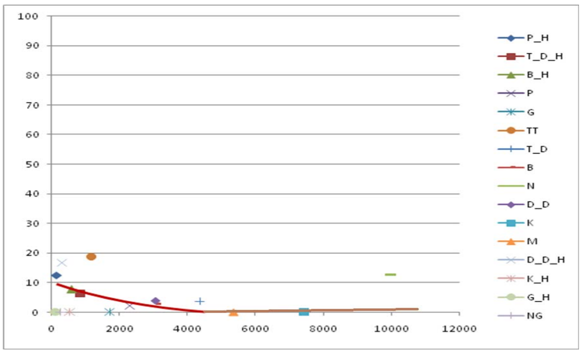


Figure 7- Percentage error rate for stops

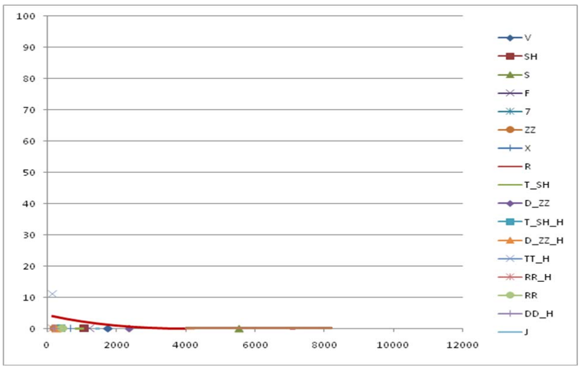


Figure 8- Percentage error rate for other consonants

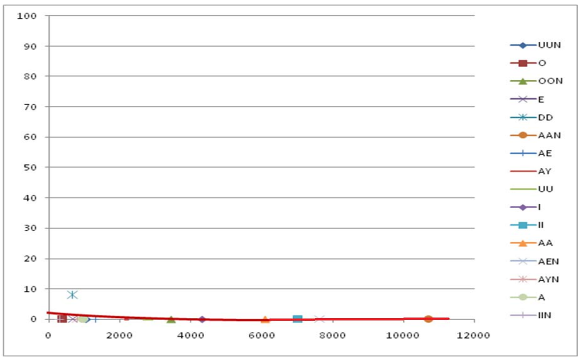


Figure 9- Percentage error rate for vowels

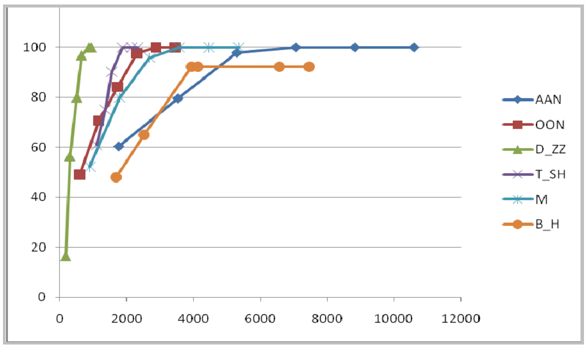
Figures 7, 8 and 9 show that the corpus has unequal amount of training data for different phonemes. It is unclear whether this amount of training data is optimal or not, as there is too much data for some phonemes and less data for other phonemes. For example, AAN phoneme has 10600 where as D_ZZ has 935 samples in the training data.

VII. EXPERIMENT-4

Phoneme accuracy has been determined on different amount of training data to reduce size of training corpus. Six phonemes have been selected, including stops, fricative, affricates and vowels, to analyze how the amount of the training data impacts the recognition accuracy of various sounds.

Table 7- Phoneme accuracy with varied training data

Phoneme	Original training data	Reduced training data	Accuracy (%)	Reduction in Corpus Size (%)
AAN	10600	7064	100	33.4
OON	3443	2870	100	16.6
D_ZZ	935	870	100	6.9
T_SH	2364	1867	100	21
M	5354	3568	100	33.4
B_H	7426	3938	92.31	46.9
Average	**5020**	**3362**	**98.71**	**26.3**



Amount of training data

Figure 10- Phoneme accuracy versus training data

A. Experiment-4 discussion

It can be seen from the Figures 7, 8 and 9 the amount of training data for each phoneme at which 100% accuracy can be achieved is different for each phoneme. In Figure 9 the amount of training data for AAN is 10600 and for OON is 3700. There might be a possibility that amount of training data has been saturated before 3700 for OON phoneme. This possibility has been explored and Table 7 shows the results.

Figure 10 indicates that accuracy of AAN, OON, D_ZZ, T_SH, M and B_H phonemes saturates at amount of training data 7000, 2880, 850, 1800, 3500 and 3900 respectively. Table 7 shows the default and reduced amount of training data of these phonemes in phonetically rich corpus. Same accuracy for these phonemes has been achieved with lesser amount of training data.

Figure 10 shows no effect on accuracy if the training data has been increased beyond the saturation limits e.g. accuracy for B_H phoneme saturates at 92.31%, and increasing the amount of training data has no effect on the accuracy. The accuracy of this phoneme has been found to be 100% in [11]. Same speech corpora have been recorded in the development of both ASR systems so there might be transcription issues as reported in [11]. This problem may be solved by cleaning the training data of this phoneme to improve the acoustic model.

It can be seen that accuracy of phonemes depends largely on training of ASR system. By analyzing the slopes of curves in Figure 10, affricates, D_ZZ and T_SH, phonemes give 100% accuracy on less amount of training data. The slopes of curves may reflect inherent differences in the phonemes. The saturation value is also different for all these phonemes.

VIII. CONCLUSION

The current work shows that accuracy of recognition improves with more balanced and clean data, both for training and for testing. This is perhaps expected. However, the final experiment shows that balanced data does not necessarily mean an equal amount of data to train all the phonemes. The experimental results show that some phonemes may require very little data for their recognition accuracy to saturate to 100%, where as some others may require more data. If the data is incrementally tagged, the slope of accuracy improvement is indicative of the requirement of the additional data. Further, some phonemes may saturate to less than 100% accuracy. These factors can be taken into consideration for incrementally developing a speech corpus and reducing the amount of corpus in a more optimal fashion.

IX. REFERENCES

[1] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, R. Parveen, "Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System", *in proceeding of O-COCOSDA,* Kathmandu, Nepal, 2010.

[2] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, R. Parveen, "Large Vocabulary Continuous Speech Recognition for Urdu", in the *Proceedings of International Conference on Frontiers of Information Technology (FIT),* Islamabad, Pakistan, 21-23 December 2010.

[3] S. T. Abate, W. Menzel, and B. Tafira. "An Amharic speech corpus for large vocabulary continuous speech recognition". *In Proceedings of the 9th European Conference on Speech Communication and Technology*, Interspeech-2005, Lisbon, Portugal, 2005.

[4] A. Gopalakrishna, R. Chitturi, S. Joshi, R. Kumar, S. Singh, R.N.V Sitaram and S.P. Kishore, "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems", *Proceedings of International Conference on Speech and Computer (SPECOM)*, Patras, Greece, Oct 2005.

[5] B. Bozkurt , O. Ozturk , T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection", *in Proceedings of the Eurospeech*, 2003.

[6] V. Chourasia, K. Samudravijaya, and M. Chandwani, "Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database", *In the Proceeding of O-COCOSDA*, pp. 132–137, 2005.

[7] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V.Diakoloukas, "Large Vocabulary Continuous Speech Recognition in Greek: Corpus and

an Automatic Dictation System", *Eurospeech,* pp. 1565-1568, 2003.

[8] G. Raškinis, "Building medium-vocabulary isolated word Lithuanian HMM speech recognition system", *In Informatica Journal*, Volume 14, pp. 75-84, 2003

[9] A. L. Ronzhin, R. M. Yusupov, I. V. Li, and A. B. Leontieva, "Survey of Russian Speech Recognition Systems", *In Proc. of 11th International Conference SPECOM*, St. Petersburg, Anatoliya, pp. 54-60, 2006.

[10] L. V. Pineda, M. M. Gomez, D. Vaufreydaz, and J. F. Serignat, "Experiments on the Construction of a Phonetically Balanced Corpus from the Web, Lecture notes in computer science", *Conference on Intelligent Text Processing and Computational Linguistics Cicling,* pp. 416-419, 2004.

[11] S. Irtza, S. Hussain, "Error Analysis of Single Speaker Urdu Speech Recognition System", *CLT-12*, University of Engineering and Technology, Lahore, Pakistan, 2012.

[12] P. david, H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *ISCA ITRW*, September 18-20, 2000.

[13] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Haji c, D. Oard,M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, "Automatic recognition of spontaneous speech for access to multilingual oral history archives", *IEEE Transactions on Speech and Audio Processing*, 2004.

[14] O. Siohan, B. Ramabhadran, G. Zweig, "Speech Recognition Error Analysis on the English MALACH Corpus", *ICSLP 8th international conference on spoken language processing*, Jeju island, Korea, October 4-8, 2004.

[15] T. Pruthi, Saksena, S and Das, P. K. Swaranjali, "Isolated Word Recognition for Hindi Language using VQ and HMM", *International Conference on Multimedia Processing and Systems (ICMPS)*, IIT Madras, 2000.

[16] K. Kuldeep, R. K. Aggarwal, "Hindi speech recognition system using htk"*, International journal of computing and business ISSN(online)* :2229-6166, volume 1,May 2011.

[17] M. U. Akram and M. Arif, Design of an Urdu "Speech Recognizer based upon acoustic phonetic modelling approach", *IEEE INMIC* 2004, pp. 91-96, 24-26 December, 2004.

[18] S. M. Azam, Z.A. Mansoor, M. Shahzad Mughal, S. Mohsin, "Urdu Spoken Digits Recognition Using Classified MFCC and Backpropgation Neural Network", *IEEE Computer Graphics, Imaging and Visualisation CGIV*, Bangkok, 14-17 August, 2007.

[19] J. Ashraf, N. Iqbal, N. S. Khattak, A. M. Zaidi, "Speaker Independent Urdu Speech Recognition Using HMM", *INFOS, IEEE*, Cairo, 28-30 March, 2010.

[20] A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah and Z. Sarfraz, "An ASR System for Spontaneous Urdu Speech", *In the Proc. of Oriental COCOSDA*, Kathmandu, Nepal. 24-25 November 2010.

[21] T.Misu and T. Kawahara, "A bootstrapping approach for developing language model of new spoken dialogue systems be selecting web texts", *ICSLP*, 2006.

[22] P. J. Jang and A. G. H., "Improving Acoustic Models with Captioned Multimedia Speech", *Multimedia computing system, IEEE*, Florence, Italy, July, 1999.