Error Analysis of Single Speaker Urdu Speech Recognition System

Saad Irtza, Dr. Sarmad Hussain
Center for Language Engineering, Al-Khawarizmi
Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan
Saad.lrtaza@kics.edu.pk,Sarmad.Hussain@kics.edu.pk

Abstract

Speaker independent, spontaneous and continuous speech recognition system (ASR) can be integrated to other technologies like mobile to create an interface between technology and illiterate people so that they can use modern technologies. One of the major hurdles in such ASR is unacceptable word error rate. The paper explores the possibility of analyzing the Urdu speech corpus based on recognition results to improve word error rate (WER).

1. Introduction

This paper describes the implementation of single speaker ASR system and process employed in the analysis of speech corpus based on recognition results. The speech corpus has been recorded based on earlier corpus design [1]. It consists of mixture of read and spontaneous speech and divided into two portions for training and testing [2]. Based on the test results confusion matrix has been generated indicating the correctly matched and confused phonemes. Speech corpus has been updated based on the results. The next section describes the previously work done on Urdu ASR and of other languages.

2. Literature Review

Many ASR systems has been developed using different methods on different languages. Brief survey of ASR in Urdu and different languages will be presented in this section.

Performance of English speech recognition has been evaluated in noisy condition by using HTK toolkit [3]. Data of fifty two male and female speakers, 8440 utterance and connected digits has been recorded at different places having different SNR's. Average word accuracy is 87.81%.

Performance of English ASR has been analyzed based on word recognition error rate on subset of Malach [4] corpus. Noise compensation technique has been implemented that results in 1.1% reduction of WER [5].

Speaker adaptive training [6] (SAT) and discriminative training with minimum phone frame error (MPFE) criterion has been used to decrease the errors in Finish Morph based continuous recognition system [7]. Error analysis based on acoustic model has been performed in continuous Chinese speech recognition system Easy talk [8][9].

Speech recognition system using subspace Gaussian mixture model approach has been developed by having sixteen Gaussian per state. The system has been trained and tested on English, German and Spanish languages of 15.1, 16.5, 14.7 and 1.8, 2.0, 3.7 hours of data respectively. CallHome [10], corpora has been used for evaluation of training and decoding of recognition performance. Phoneme error rate for English, German and Spanish language has been reduced from 54.9, 46.2, 56.3% to 51.7, 44.0, 53.4%.

Hindi (Swaranjali) speech recognizer has been developed for two male speakers [11]. Rcognition vocabulary consists of isolated hindi digits from zero to nine and trained with twenty utterance of a word for each speaker. Recognition result for two speakers has been found to be 84.49% and 84.27%. Hindi speech recognition system by using HTK toolkit has been developed for eight speakers. Recognizer is based on acoustic word model. Rcognition vocabulary consists of thirty isolated Hindi words. Word accuracy has been found to be 94.63% [12].

Robust Urdu speech recognition system by using Sphinx 3 toolkit has been developed in which three language models have been developed incrementally, one model consist of data from 40 female speakers only, one from 41 male speakers only, and one with both male and female speakers (81 speakers). The error rate was 60.2% [2]. An Urdu SR system using by using Pattern-Matching and Acoustic Modelling approaches

to SR for Urdu language has been proposed with a 55-60% accuracy rate [13]. They have used ANN (Artificial Neural Network) to convert a set of frames into phonetic based categories. They used Viterbi search algorithm to search the best sequence path for the given word to be recognized. A single speaker SR system for isolated Urdu digits by using ANN approach has been developed [14]. A mono-speaker database system for Urdu digits by using ANN approach in Multilayer Perceptron (MLP) has been proposed [15]. This system is implemented by using Matlab toolkit. Urdu ASR system has been developed by using sphinx4. This system is based on small vocabulary (fifty two isolated spoken Urdu words). Training set consists of speech data from ten speakers having total of 5200 utterances. The mean word error rate was 5.33% [16]. An Urdu ASR system of single speaker medium vocabulary, 800 utterances consisted of read and spontaneous speech data are mixed together in various ratios, has been developed and the system is tested using spontaneous speech data only [17].

3. Methodology

Two experiments have been developed. 1) Experiment-1 (Baseline) 2) Experiment-2 (Revised). Training and testing data for baseline experiment is described in Table-1.

Table1- Baseline Data

No. of training utterances	620
Duration of Data	56 minutes
No. of test utterances	45
Read speech utterances	351
Spontaneous speech utterances	269

The speech corpus on Urdu language for the testing and training has been developed described in [2]. Training and testing data is non-overlapping. The transcription of speech files will be done manually and orthographically in Urdu script. The transcription rules will be based on [19]. In the transcription of speech files non-speech areas in the segments will be represented by the Silence, Vocalization and Breath tags manually. All the files will be converted to the Sphinx format using the Sphinx Files Compiler described in [17].

Based on the recognition issues data of revised experiment-2 has been modified as described in Table-2.

Table2- Revised Data

No. of training utterances	671
Duration of Data	65 minutes
No. of test utterances	60
Read speech utterances	400
Spontaneous speech utterances	269

Data has been added on incremental basis such that amount of training data does not remain a significant factor in decreasing word error rate. This additional data has been selected to increase the amount of training phoneme. This data has been added in form of full read sentences. Words have been chosen that consists of these phonemes. Sentences have been selected such that it contains maximum of these words. The aim is to analyze the effect of increasing training data on phoneme accuracy.

3.1. Toolkit

Speech data has been recorded on laptop in wav format at 16 kHz sampling frequency. Praat [18] has been used to record the speech files over the microphone channel. Sphinx, speech recognition toolkit has been used to develop and test the acoustic model. The Latest nightly release of Sphinx train Sphinx3 has been used to develop the ASR system.

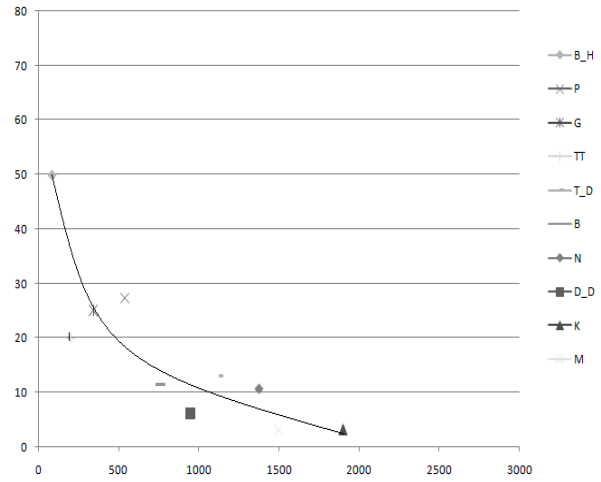
4. Experiment-1 Recognition Results

Baseline recognition results are described here. To perform error analysis on above recognition results, algorithm has been developed to find the frequency of training, matched and confused phonemes and tabulated in the form of matrix. The following graphs show the relationship between the percentage error rate and amount of training data for every phoneme.

Table3- Baseline Recognition Results

No. of tied states	100
Beam width	1e-120
Language weight	23
Word error rate	18%

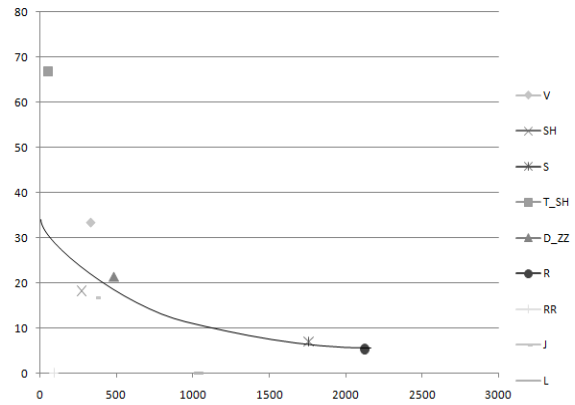
Percentage error rate



Amount of training data

Figure1- Graph for Stops

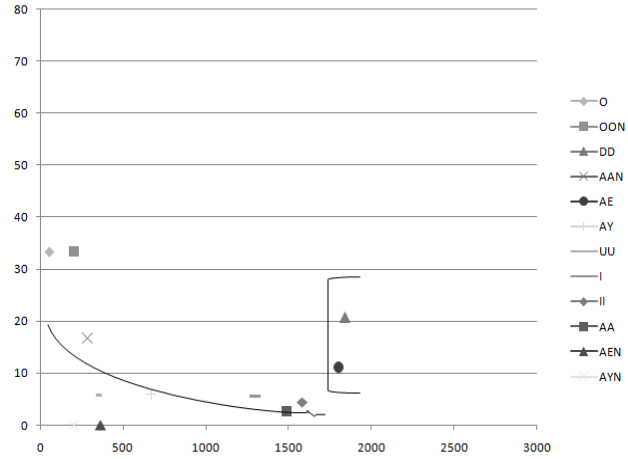
Percentage error rate



Amount of training data

Figure2- Graph for Fricatives, Trills, Flap, Approximants

Percentage error rate



Amount of training data

Figure3- Graph for Vowels

The following Table shows the original phoneme, the confused ones and confusion frequency.

Table4- Confusion Matrix

Phone	Confusion	Frequency
P	Sil	3
TT	Sil	1
T_D	Sil	3
T_D	D_D	1
N	Sil	3
K	Sil	2
K	P	1
K	B	1
M	Noise	1
V	R	1
Z	D_D	2
Z	R	1
Z	Noise	1
F	Sil	2
SH	K	1
SH	H	1
S	Noise	1
H	Sil	1
T_SH	AA	1
D_ZZ	Z	1
D_ZZ	Noise	2
R	Noise	2

J	Noise	2
O	OON	1
OO	O	2
OO	AE	1
AE	Sil	1
U	AA	1
U	Sil	2
I	II	1
I	Sil	2
AA	OO	2
AA	Sil	2
AA	Noise	4

4.1. Experiment-1 Discussion

Following are the conclusions extracted from the Figure-1, 2, 3 and Table4. Large y-axis value and small x-axis value gives large error rate due to the reason that training data was not sufficient. Large y-axis value and large x-axis value gives large error rate. As the training data was sufficient so there might be two reasons for it:

i) Tagging error
ii) Phoneme is problematic

As described in section-3, the Silence, Vocalization and Breath tags will be defined manually to represent non-speech areas in the segments. From Table-4, there are a lot of confusion between phonemes and silence. Following four solutions have been used to overcome the above problem:

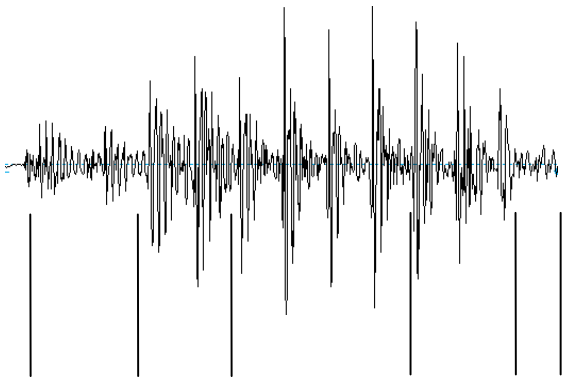
1- Carefully analyze the transcription to check tagging error.
2- Add more data such that phonemes are balanced.
3- The non-speech areas in the segment should be identified automatically.

There are some phonemes whose training data is sufficient but error rate is high. As described in section-3, the segmented speech files have been transcribed orthographically in Urdu script manually. These phonemes have been analyzed in the transcription and original sound files and following are some issues are:

1- Stops phoneme (T_D) was not completely uttered by the speaker.
2- Vowels (DD, AE) were not correctly transcribed at some places.

These issues are solved according to original wave files. There are some phonemes whose training data was not sufficient. Based on results from figure-1, 2, 3, data has been added to previous speech corpus. The non-speech areas in the segment have been identified automatically by using force alignment algorithm. It aligns the transcribed data with the speech data [21].

Original Transcript: <s> <sil> NORMAL KAE HAALAAT_D HAYN <sil> </s>
Force-aligned Transcript: <s> NORMAL KAE <sil> HAALAAT_D HAYN </s>



sil NORMAL KAE HAALAAT_D HAYN sil

5. Experiment-2 Results

Revised recognition results are described here. Following graphs show the improved results of above approaches.

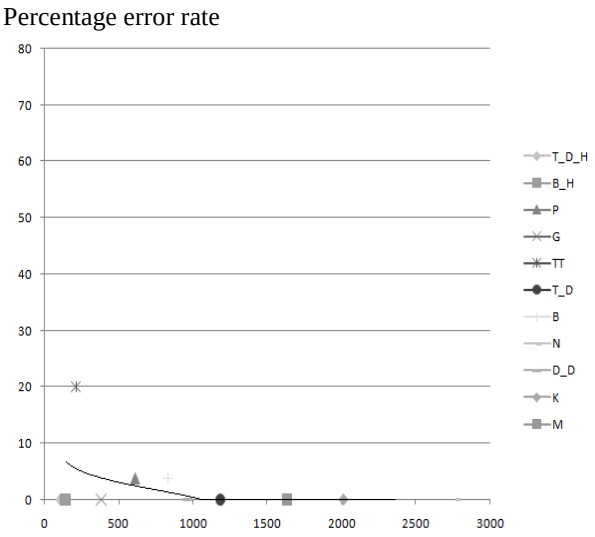
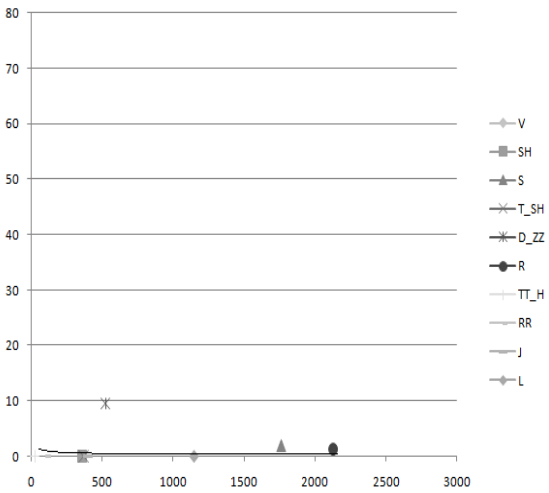


Figure4- Graph for Stops

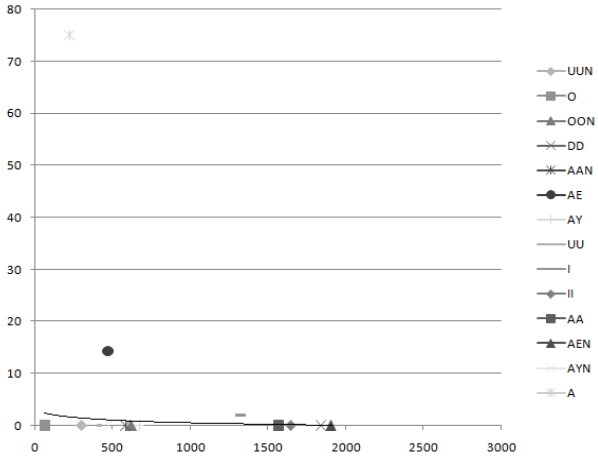
Percentage error rate



Amount of training data

Figure5- Graph for Graph for Fricatives, Trills, Flap, Approximants

Percentage error rate



Amount of training data

Figure6- Graph for Vowels

Analysis of transcription improves the phoneme accuracy as described in Table5.

Table5- Improved Phoneme Accuracy

Phon e-me	Training Data	Previous Error rate (%)	Improved Error rate (%)	Percentage Improvement (%)
T_D	1127	13.04	6.52	50
DD	1842	20.69	6.89	66.67
AE	1804	11.11	6.67	39.69

Effect of increasing training data improves the phoneme accuracy as described in Table6.

Table6- Improved Phoneme Accuracy

Phoneme	Original training data	Increased training data	Improved accuracy from (%)-to (%)
B_H	82	142	50-0
P	540	608	27.3-3.7
G	342	415	25-0
SH	276	360	18.1-0
T_SH	55	515	66.3-0
D_ZZ	485	524	21.4-9.5
O	25	101	33.3-0
OON	203	621	33.3-0
AAN	285	585	16.6-0
AY	572	675	5.3-0
TT	290	974	20-20

Improved Recognition results are described in Table7.

Table7- Revised recognition Results

No. of tied states	100
Beam width	1e-120
Language weight	23
Word error rate	3.9%
Percentage Improvment	78.3%

5.1. Experiment-2 Discussion

Analysis described in section-3 gives the information that how much times each phoneme appears in the training data. Percentage error rate (PER) has been found by using the formula

$$PER = [(f2-f1) / f2]*100$$

Where
f2= # of times phoneme appear in test data
f1= # of times phoneme correctly decoded

The phoneme ‘T_D_H’ did not appear in figure-1 but in figure-4 because it did not appear in test data. Same is the case with phoneme ‘TT_H’ in figure-2,

'UUN' and 'A' in figure-3. Test data has been increased to add the above phonemes.

Training data has been increased based on the technique described in section-3. Table-6 shows the original training data, increased training data and improved accuracy. From Figure-1, 2 and 3, phonemes having relatively low training data and large error rate have been short listed in Table-6. The amount of increased training data is different for every phoneme because it has been increased in form of full sentences. Training data of phonemes other than those listed in Table-6 has also been increased because of adding full sentences e.g. training data of phoneme 'K' has been increased from approximately 1800 (Figure-1) to 2000 (Figue-4). Its error rate has also been decreased to 0%. Some phoneme has no effect on increasing training data and their error rate is also not alarming. From Table-6 error rate of some phoneme has been reduced to 0% and for others to some numerical value. It depends on the context in which phonemes appear in training data. The last phoneme 'TT' in Table-6 shows that there is no effect of increasing training data because utterance/pronunciation of this phoneme was not correct in original wav file. The saturation value of training data for each phoneme is different for single speaker. It may be different for different speaker (male/female) or number of speakers (male/female).

6. Conclusion

Section-5 shows that word error rate can be improved by refining training data, applying force alignment algorithm on transcription and increasing training data for selected phonemes. However these analysis methods will improve WER if and only if utterance/pronunciation of phonemes under observation is correct. For example the phoneme 'TT' had no effect of increasing training data (as shown in Table-6) because pronunciation was not correct in original data. The word that contains this phoneme has been difficult to pronounce.

7. References

[1] Huda Sarfraz, Sarmad Hussain, Riffat Bokhari, Agha Ali Raza, Inam Ullah, Zahid Sarfraz, Sophia Pervez, Asad Mustafa, Iqra Javed, Rahila Parveen, Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System *in proc. O-COCOSDA*, Kathmandu, Nepal, 2010.

[2] Huda Sarfraz, Sarmad Hussain, Riffat Bokhari, Agha Ali Raza, Inam Ullah, Zahid Sarfraz, Sophia Pervez, Asad Mustafa, Iqra Javed, Rahila 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] Pearce david, Hans-günter Hirsch, The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *ISCA ITRW*, September 18-20, 2000.

[4] 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.

[5] Olivier Siohan, Bhuvana Ramabhadran, Geoffrey 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.

[6] Teemu Hirsimaki and Mikko Kurimo, Analysing Recognition Errors in Unlimited-Vocabulary Speech Recognition, in *the proc. of NAACL HLT*, Boulder, Colorado, June, 2009.

[7] C. K. Raut, K. Yu and M. J. F. Gales Cambridge University Engineering Department, Adaptive Training using Discriminative Mapping Transforms, *ISCA 9th international conference of international speech communication*, Brisbane, Australia, September 22-26, 2008.

[8] Luo Chunhua, XU Mingxing, Zheng Fang, Center of Speech Technology, Acoustic Level Error Analysis in Continuous Speech Recognition, *ISCSLP*, Beijing, China, October 13-15, 2000.

[9] Fang Zheng, Zhangjiang Song, MingXing Xu, Jian Wu, Yinfei Huang, Wenhu Wu, Cheng Bi., EasyTalk, A Large-Vocabulary Speaker-Independent Chinese Dictation Machine, *EuroSpeech'99, Vol.2, pp.819-822*, Budapest, Hungary, Sept.1999.

[10] Canavan A, and G. Zipperlen, CALLHOME Spanish Speech, Linguistic Data Consortium, 1997, 1997 Contarra systems, 2001
http://www.contarra-systems.com/

[11] Pruthi T, 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.

[12] Kumar kuldeep, r. k. aggarwal, Hindi speech recognition system using htk, *Internation journal of computing and business ISSN(online) :2229-6166, volume 1*,May 2011.

[13] Akram M. U. 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.

[14] Azam S. M., 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.

[15] Ahad Abdul, Ahsan Fayyaz, Tariq Mehmood, Students Conference, Speech Recognition using Multilayer Perceptron, *ISCON apos'02. IEEE Volume 1*, 16-17 August, 2002.

[16] Ashraf Javed, Naveed Iqbal, Naveed Sarfraz Khattak, Ather Mohsin Zaidi, Speaker Independent Urdu Speech Recognition Using HMM, *INFOS IEEE*, Cairo, 28-30 March, 2010.

[17] Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah and Zahid Sarfraz, An ASR System for Spontaneous Urdu Speech, *In the Proc. of Oriental COCOSDA*, Kathmandu, Nepal. 24-25 November 2010.

[18]Praat, doing phonetics by computer, www.fon.hum.uva.nl/praat, accessed June 2010.

[19] S. Hussain , Letter to Sound Rules for Urdu Text to Speech Sytem, *in proc. of workshop on computational approaches to Arabic script-based languages*, Geneva, Switzerland, 2004.

[20] Chai Wutiwiwatchai, Patcharika Cotsomrong, Sinaporn Suebvisai, Supphanat Kanokphara, Information Research and Development Unit National Electronics and Computer Technology Center, Phonetically Distributed Continuous Speech Corpus for Thai Language, *COCOSDA*, 2003.

[21] Photina Jaeyun Jang and Alexander G.Hauptmann, Improving Acoustic Models with Captioned Multimedia Speech, Multimedia computing system, *IEEE*, Florence, Italy, July, 1999.