An Efficient Algorithm To Collect Minimal Speech Corpora

Saad Irtza[1], and Sarmad Hussain[2]

1. Electrical Engineering Department, University of Engineering & Technology, Lahore, Pakistan,
2. Center for Language Engineering, KICS, University of Engineering & Technology, Lahore, Pakistan

* **Corresponding Author:** E-mail: saad.irtaza@kics.edu.pk

Abstract

Generally phonetically rich and balanced corpora are popular for training speech recognition system but these corpora are costly to develop. Different greedy algorithms have been develop to collect such corpora. A significant effort is required to record and transcribe such speech corpora. Therefore there is motivation to further reduce their size. This paper demonstrates such an algorithm. Earlier work shows that different amount of training data is required to train different phonemes. The current work further develops these findings to reduce phonetically rich training data. Experiments show that this algorithm reduces the size of an Urdu speech corpus by 56.49% without degradation in accuracy.

Key Words: Component; Reduced speech corpus; Urdu speech corpus; Automatic speech recognition

1. Introduction

Speaker independent automatic speech recognizer (ASR) can be used to develop speech interface for different applications. To develop such speech interfaces, domain specific phonetically rich and balanced speech corpora has to be recorded from large population. Speaker accents, recording channel, age, gender and noise are the important variables that effect the performance of ASR system. Preprocessing of speech data increases, if the size of speech corpus is significantly large. A significant effort and resources are required to record and transcribe the speech corpora. This effort can be reduced if one uses reduced corpora instead of balanced one because less number of sentences will eventually reduce the recording and labeling effort.

This paper describes an effort to further reduce the phonetically rich speech corpora. The developed algorithm has been tested on Urdu speech corpus as developed in [1]. It consists of both the read and spontaneous Urdu speech data. Phoneme error analysis technique [2][3] has been performed to analyze the effect on phoneme accuracy with increase of amount of training data. The current effort is a continuation work of [3] which indicate that the balanced corpus being used for training of ASR can be further optimized across various phonemes.

2. Literature Review

There has been significant progress on development of domain specific speech corpus to be used in applications as text to speech (TTS) [7] and ASR systems [5, 6]. These corpora are available in different contexts, e.g. isolated words [10], continuous (read and spontaneous) speech [4, 6, 8, 9, 11, 12], etc. These corpora have been collected using different greedy algorithms to cover maximal phonetic coverage.

Different phonetically rich corpora have been developed in multiple languages. Russian speech corpus, TeCoRus [13], has been developed to have phonetically rich data from interview sessions and some spoken material to train phone model.

Chinese speech corpus has been developed to analyze phonetic variations, phoneme duration reduction in read and spontaneous speech [14]. The university class lectures and public meetings resources have been used to develop the corpus. Phonetically rich read speech corpus based on maximal syllables coverage in Ethiopia language has been developed [15]. Read speech data has been collected from newspaper and magazine articles. In first phase, 100,000 phonetically rich sentences has been selected. In second phase, sentences with maximal phonetic coverage and finally sentences with maximal syllable coverage and having rare syllables have been selected. SALA-II American English speech corpus has been developed over the mobile channel for speech recognition systems [16]. The speech data from the Harvard and TIMIT corpora have been used in SALA-II to increase phonetic coverage. Biphone and triphone phonetically rich corpora have been developed on Taiwanese language [17]. The effectiveness of these corpora have been analyzed by developing ASR systems. Syllable recognition accuracy has been found to be better for biphone phonetically rich corpus.

Another effort has been made to develop minimally phonetically rich corpus from website and newspaper sources [18] in, Tamil, Marathi and Telugu, local Hindi languages. Optimal text selection algorithm has been used to cover the phonetic variations in Tamil, Marathi and Telugu languages. Hindi speech corpus has also been selected using the same optimal text selection algorithm [19]. The baseline data has been collected from articles, magazines and online content available. A large vocabulary Urdu speech corpus has been developed on read and spontaneous data. To collect spontaneous speech data questions sets from hobbies, daily routines, interests and past experience have been designed. For read speech data, 725 phonetically rich sentences and six paragraphs have been developed from 18 million Urdu words.

Earlier, corpora have been preferred to develop covering balanced phonetic coverage that results in large size of speech corpora. It requires significant effort in recording and labeling process. Then greedy algorithms have been developed to reduce this effort and to have balanced phonetic coverage in smaller size of speech corpus. Greedy algorithms have been developed to collect phonetically rich and balanced corpora from different sources [20, 21, 22]. A greedy algorithm has been developed to collect Turkish speech corpus [20]. In first phase, depending on diaphone coverage, cost has been assigned to each sentence. In second phase, maximal cost sentences have been selected. Finally sentences having unique diphones have also been included in corpus. Out of 11500 sentences in baseline corpus 2500 sentences have been selected in final corpus. Irish speech corpus has also been developed by using slight modification in above greedy algorithm [21] i.e. in second phase, sentences have been selected to have maximal unit coverage instead of diphone. A more robust greedy algorithm has been developed to collect phonetically balanced and distributed sentences using iterative method for Thai language [22]. ORCHID standard corpus has been used as baseline corpus. In first phase, initial score has been assigned to sentences based on phoneme frequency in a sentence. Sentences have selected from low to high frequency phoneme sentences. The dot product is computed to find similarity between the distribution of units in baseline corpus and final corpus. The final corpus consists of 398 phonetically balanced and 802 phonetically distributed sentences out of 27,634 sentences.

The work presented in [3] describes that one might not need the balanced phonetic coverage in speech corpus to have better recognition results. It still has to be explored that if one collects the corpus by determining the relationship between phoneme training data and accuracy, will it further reduce the corpus size or not.

3. Methodology

The effort presented in [3] to collect minimally balanced corpus has been further extended to develop an algorithm to collect optimal speech corpus. In [3] six phonemes has been selected to analyze the effect of increasing training data on phoneme accuracy. It has been concluded that training data for each phoneme saturates at some point and does not further improve phoneme accuracy. The saturation limit for each phoneme is different from the other. In this paper this concept has been extended on all the phonemes in corpus by developing an algorithm that collects optimal training data for each phoneme. In first phase (Experiment-1), phonetically rich speech corpus [1] has been used to develop large vocabulary continuous and read ASR system. In Experiment-1 speech corpus recorded from 82speaker's has been used to develop ASR system. The contents of phonetically rich speech corpus [1] and Experiment-1 data has been described in Table-1 & Table-2 respectively.

Table 1. Contents of Phonetically Rich Urdu Speech Corpus

	Category	Questions
Spontaneous speech	Bio data	10
	Past experience Questions	22
	Hobbies	32
	Miscellaneous	157
Read sentences	Phonetically rich sentences covering 18 million Urdu words	725
Read passages	Open Urdu content	6

Table 2. Experiment-1 Data

No. of training utterances	30,983
No. of Speakers	40 male, 42 female
Recording time	30 hours
Recording Environment	Laptop/Office
Sampling rate	16KHz
No. of test utterances	6,190
Read speech utterances	12,393
Spontaneous speech utterances	18,590

In second phase (Experiment-2), the developed algorithm has been applied on speech corpus to collect optimal corpus. Another ASR system has been developed on optimal corpus. Word and phoneme accuracy has been compared of both systems to analyze the effect of optimal corpus on minimally balanced corpus. Experiment-2 data has been described in Table-3.

Table 3. Experiment-2 Data

No. of training utterances	18,590
No. of Speakers	40 male, 42 female
Recording time	17 hours
Recording Environment	Laptop/Office
Sampling rate	16KHz
No. of test utterances	6,190
Read speech utterances	8,135
Spontaneous speech utterances	10,455

In third phase, the algorithm presented in [22] has been applied on speech corpus to collect reduced phonetically rich corpus. Experiment-3 data has been described in Table-4.

Table 4. Experiment-3 Data

No. of training utterances	20,145
No. of Speakers	40 male, 42 female
Recording time	19.5 hours
Recording Environment	Laptop/Office
Sampling rate	16KHz
No. of test utterances	6,190
Read speech utterances	6,556
Spontaneous speech utterances	13,589

4. Optimal Corpus Selection Algorithm

Phonetically rich speech corpus [1] will be used as baseline (input) corpus for this algorithm. The algorithm has been divided in two phases. In first phase, low frequency phonemes will be selected to have good phonetic coverage. Phoneme list will be determined from baseline corpus and sorted on the basis of increase in frequency. In each iteration, non-overlapping k (e.g. five) sentences of a low frequency phoneme will be included in corpus CR. ASR system will be developed on CR and tested on testing corpus CT to analyze the phoneme accuracy. These sentences will be kept in CR if phoneme accuracy is greater than previous one. As the training data of ASR system will be very low in first phase so the value of baseline threshold T should be low (e.g. 25%). This iterative method will be repeated until phoneme accuracy greater threshold T (e.g. 50%) is achieved. The same process will be repeated for all phonemes. Phase-1 of this algorithm will gives a corpus in which all the phonemes will have accuracy greater than threshold T. In this process, some high frequency phonemes will also be included.

The pseudo code of phase-1 has been described below:

Phase – I

1. *Input phonetically rich corpus C_S containing sentences S1 till Si, C_R reduced output corpus and testing corpus C_T*

2. *Extract all unique phonemes from C_S and store unique phoneme list L1*
3. *Sort L1 in increasing order of phoneme frequency in C_S*
4. *For all phonemes in L1, starting with lowest frequency phoneme in L1*
   - *4.1. Set current accuracy threshold T to 25%*
   - *4.2. From different combinations of k (e.g. 5) sentences from corpus C_S which contain the current phoneme, select a combination which gives a phoneme accuracy greater than T*
   - *4.3. Move these k sentences from C_S to C_R*
   - *4.4. Set T for the current phoneme to its current accuracy value*
   - *4.5. Repeat from Step 4.2 until T=50%*

In second phase, frequency of phonemes in L1 has been updated from CR. Non overlapping five sentences of a low frequency phoneme will be included in CR. ASR system will be trained on CR and tested on CT. The selected sentences will be kept in CR if new phoneme accuracy is greater than previous one. Baseline threshold in phase-2 will be final threshold of phase-1 i.e. 50%. This iterative method will be continued until accuracy of that phoneme is greater than threshold T (e.g. 90%). The same process will be repeated for all phonemes. A phoneme will not be included in this iterative method if its accuracy is already greater than threshold T.

Phase – II

1. *Update frequency of phonemes in L1 from C_R*
2. *For all phonemes in L1, starting with lowest frequency phoneme in L1*
   - *2.1. Set current accuracy threshold T to 50%*
   - *2.2. From different combinations of k (e.g. 5) sentences from corpus C_S which contain the current phoneme, select a combination which gives a phoneme accuracy greater than T*
   - *2.3. Move these k sentences from C_S to C_R*
   - *2.4. Set T for the current phoneme to its current accuracy value*
   - *2.5. Repeat from Step 2.2 until at least T=85%*

5. Experimental Result

The recognition results of Experiment-1 has been described in Table-5.

Table 5. Experiment-1 Recognition Results

No. of tied states	1000
Language weight	11
Word error rate	57.3%

Phoneme error analysis has been performed on above ASR system. Phoneme error rate been determined and plotted versus amount of training in Figure-1. Detail of phoneme training data and error rate has been given in Appendix A.

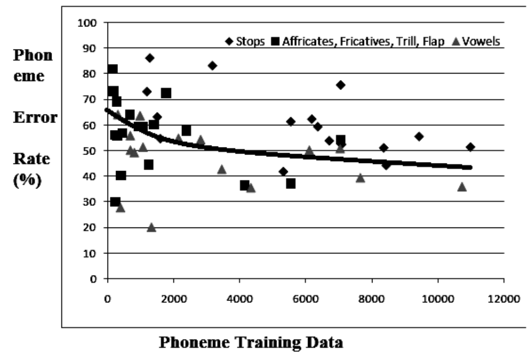

Fig. 1 Phoneme error rate of Experiment-1

The recognition results of Experiment-2 has been described in Table-6.

Table 6. Experiment-2 Recognition Results

No. of tied states	1000
Language weight	11
Word error rate	14.9%

Figure-2 shows the graph between phoneme error rate and amount of training. Detail of phoneme training data and error rate has been given in Appendix A.

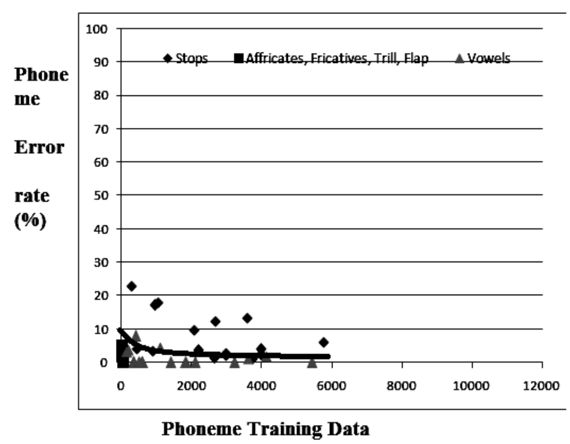

Fig. 2. Phoneme error rate of Experiment-3

The recognition results of Experiment-2 has been described in Table-7.

Table 7 Experiment-3 Recognition Results

No. of tied states	1000
Language weight	11
Word error rate	20.5%

Figure-3 shows the graph between phoneme error rate and amount of training. Detail of phoneme training data and error rate has been given in Appendix A.

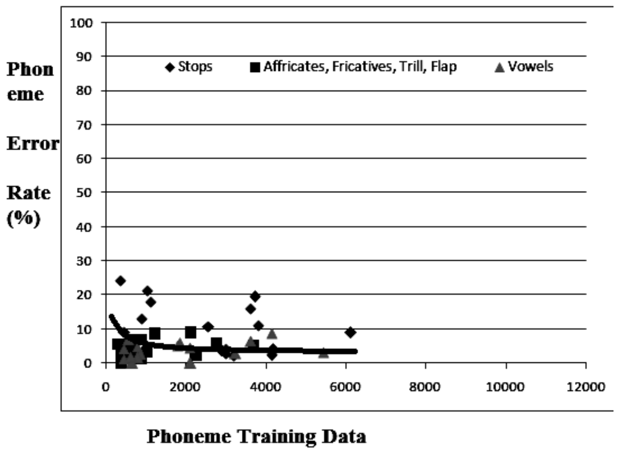

Fig. 3. Phoneme error rate of Experiment-3

Table-8 shows the original and optimal training data used in Experiment-1 & 2 of phonemes category respectively. The last column describes the reduction in training data of each category.

6. Discussion

The minimal corpus selection algorithm has selected 18,590 training utterances in Experiment-2 out of 30,983 training utterances in Experiment-1. The phonetically balanced corpus selection algorithm has selected 20,145 training utterances. Read and spontaneous speech utterances have also been reduced to 8,135 and 10,455 out of 12,393 and 18,590 respectively. Test utterances have been kept same to compare the recognition results of three experiments. Figures-1, 2 and 3 show the amount of training data and phoneme error rate for Experiments-1, 2 and 3 respectively. Phonemes have been divided in three categories on the basis of degree of opening of the vocal tract i.e. (i) stops, (ii) fricatives, affricates, trill, flap and (iii) vowels. Figure-1 shows that the general trend is that error rate decreases with the increase of training data (as indicated by the solid line). However, this is not true in a few cases. For example 'N' stop has 10,980 training samples but its error rate is still 51.60%. Further, stops generally show higher error rates than other category.

As discussed in [3], different phonemes require different amount of training data to achieve maximum accuracy. The current work confirms this observation and ensures that this aspect can be utilized to achieve equally accurate recognition system with considerably reduced training data. Figure-2, as summarized in Table-7, shows overall reduction of 50.7% and 56.49% of training data by using reduced and minimal phonetically rich speech corpora and reduction phoneme in error rate.

Table 8 Reduction In Training Data of Phoneme

Phoneme Category	Original Training Data (1)	Phonetically Rich Training data (2)	Minimal Training Data (3)	Phoneme Error rate (%-%) (1)-(2)-(3)	Reduction in Training Data (%) (1)-(2)/(1)-(3)
Stops	90148	44079	40287	60.7-10.6-7.8	51.1/55.3
Vowel	60254	29826	27559	41.5-3.3-1.2	50.5/54.2
Fricatives, Affricates, Trill and Flap	41424	20660	15611	56-4.3-1.78	50.1/62.3
Overall reduction in speech corpus	191826	94565	83457	----------	50.7/56.49

For vowels same accuracy has been achieved with less amount of training data e.g. 'AAN' has same accuracy with reduced 5432 training utterances. Interestingly, some phonemes show better accuracy when trained on less amount of training data e.g. by selection of reduced training data (3970) for 'N' phoneme, error rate of this phoneme has been decreased from 51.60% to 4.0%. 'P_H' and 'D_D_H' stops shows similar trend. Reducing training data for these phonemes decrease the phoneme error rates from 63.33% to 23% and 83.3% to 18.1% respectively.

7. Conclusion

The current work shows that speech corpora can be collected more efficiently by analyzing the phoneme error rates. This algorithm can be further modified to collect optimal speech corpora. The effectiveness of this algorithm has to be explored on other languages.

8. 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", *Oriental COCOSDA 2010 conference, Nov. 24-25, 2010,* Katmandu, Nepal.

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

[3] Irtza, S. and Hussain, S. "Minimally Balanced Corpus for Speech Recognition*", in the Proceedings of 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13),* IEEE, Sharjah, 2013.

[4] J. Daniel & J. H. Martin, Speech and Language Processing: *An introduction to natural language processing computational linguistics, and speech recognition, 2005.*

[5] 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.

[6] A. Samoulian, *Knowledge based approach to speech recognition,* Department of Electrical and Computer Engineering University of Wollongong.

[7] L. Deng , H. Strik, *Structure-Based and Template-Based Automatic Speech Recognition*, Comparing parametric and non-parametric approaches".

[8] Nirav S. Uchat, *Hidden Markov Model and Speech Recognition*.

[9] HTK, http://htk.eng.cam.ac.uk, accessed July 2010.

[10] S. T. Abate, W. Menzel, and B. Tafila, "An amharic speech corpus for large vocabulary Continuous speech recognition," *ISCA, 2005. Ninth European Conference on Speech Communication and Technology*.

[11] L. Villase.or-Pineda, M. Montes-y Gomez, D. Vaufreydaz, and J. F. Serignat, "Experiments on the construction of a phonetically balanced corpus from the web," Lecture notes in computer science, pp. 416–419, 2004.

[12] A. Li, F. Zheng, W. Byrne, P. Fung, T. Kamm, Y. Liu, Z. Song, U. Ruhi, V. Venkataramani, and X. Chen, "Cass: A phonetically transcribed corpus of mandarin spontaneous speech," *ISCA, 2000. Sixth International Conference on Spoken Language Processing.*

[13] A. L. Ronzhin, R. M. Yusupov, I.V. Li, and A. B. Leontieva, "Survey of Russian speech recognition systems".

[14] A. Li, F. Zheng, W. Byrne, P. Fung, T. Kamm, Y. Liu, Z. Song, U. Ruhi, V. Venkataramani, and X. X. Xhen, " Cass: A phonetically transcribed corpus of mandrain spontaneous speech, *ISCA, 2000. Sixth International Conference on Spoken Language Processing.*

[15] S. T Abate, W. Menzel, and B. Tafila, "An Amharic speech corpus for large vocabulary continuous speech recognition," *ISCA, 2005. Ninth European Conference on Speech.*

[16] P. A Heeman, "The American English sala-II data collection," 2004. *Proceedings LREC.*

[17] G. Raskinis, "Building medium vocabulary isolated word Lithuanian HMM speech

recognition system," *Informatica, vol. 14, no. 1, pp.75-84, 2003.*

[18] G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S. P. Singh, R. N. V. Sitaram, and S. P. Kishore, "Development of Indian language speech database for large vocabulary speech recognition system".

[19] V. Chourasia, K. Samudravijaya, and M. Chandwant, "Phonetically rich Hindi sentence corpus for creation of speech database," *Proc. O-Cocosda,p. 132-137, 2005.*

[20] Y. C. Yio, M. S Liang, Y.C. Chiang, and R. Y. Lyu, "Biphone rich versus triphone rich: a comparison of speech corpora in automatic speech recognition," pp.194-197, 2005. Cellular Neural Networks and their applications, *2005 9th International workshop.*

[21] A. C. Kelly, H. Berthelsen, N. Campbell, A. Chasaide, C. Gobl , "Corpus Design Techniques for Irish Speech Synthesis Phonetics and Speech Laboratory", SLSCS, *Trinity College Dublin, Ireland, 2006.*

[22] W. Chai, P. Cotsomrong, S. Suebvisai, S. Kanokphara, Information Research and Development Unit National Electronics and Computer Technology Center, Phonetically Distributed Continuous Speech Corpus for Thai Language, *COCOSDA, 2003.*

Appendix

Phoneme	Experiment-1 error rate	Experiment-2 error rate	Experiment-3 error rate
P_H	63.3	23	**24.2**
T_D_H	44.4	6.25	**8.9**
B_H	54.86	17.19	**21.3**
P	42	9.8	**10.7**
G	54.1	2.1	**2.1**
TT	62.5	12.39	**15.98**
T_D	59.5	3.67	**3.676**
B	52.63	2.8	**2.8**
N	51.6	4	**4**
D_D	75.8	13.35	**19.5**
K	55.7	1.1	**11.1**
M	51.3	1.4	**4.1**
D_D_H	83.3	18.1	**18.1**
K_H	61.53	2.4	**2.4**
G_H	73.2	4.1	**9.2**
NG	86.3	3.6	**13.1**
V	72.5	2.4	**3.1**
SH	59.5	0	**5.4**
S	37	5.1	**5.9**
F	44.5	0	**6.7**
7	69.2	0	**2.1**
ZZ	56	4.9	**6.8**
X	64	0	**1.9**
R	54.37	1.1	**5.1**
T_SH	59.3	0	**3.7**

Phoneme	Experiment-1 error rate	Experiment-2 error rate	Experiment-3 error rate
D_ZZ	57.7	3.7	**2.1**
T_SH_H	40.3	2.9	**8.6**
D_ZZ_H	55.8	0	**0**
TT_H	81.8	2.9	**3.1**
RR_H	73.2	1.3	**1.3**
RR	56.8	0	**5.5**
DD_H	30	1.9	**2.1**
J	60.1	2.1	**6.7**
L	36.3	3.8	**8.9**
UUN	51.4	0	**2.7**
O	27.7	8	**4.1**
OON	42.8	0	**5.6**
E	50.2	0	**1.3**
DD	56	4.5	**7**
AAN	35.9	0	**3.2**
AE	20	0	**1.7**
AY	54.9	4.5	**4.5**
UU	54	0	**5.1**
I	35	0	**0**
II	50.8	1.1	**6.5**
AA	50.2	0	**2.9**
AEN	39.4	1.9	**8.7**
AYN	49.3	0	**0**
A	63.7	0	**1.8**
IIN	64.2	3.1	**4.8**