

Error Analysis and Improving the Speech Recognition Accuracy on Telugu Language

N. Usha Rani¹ and P.N. Girija²

¹ Department of CIS, University of Hyderabad, Hyderabad, India
& Sri Venkateswara University College of Engineering,
Sri Venkateswara University, Tirupati, AP, India.

² Department of CIS, University of Hyderabad, Hyderabad, India

Abstract. Speech is the one of the most important communication channel among the humans. Speech recognition occupies prominent place in communication between the humans and machine. Several factors are affecting the accuracy of the speech recognition system. Much effort has been done to increase the accuracy of the speech recognition system. Still erroneous output is generated in current speech recognition system. Static pronunciation dictionary plays key role in the speech recognition accuracy. The required phoneme of the word changes to the phoneme of the some other word. Modification in the dictionary in the decoder of the speech recognition system reduces the number of the confusion pairs which automatically increase the accuracy. Hit rate is considerably increased and false alarms have been changed during the modification of the pronunciation dictionary. Also this proposed method observed the variations on different error measures such as F-measures, Error-Rate and WER by applying this proposed method.

Keywords: Speech Recognition, Pronunciation Dictionary, Error Analysis.

1 Introduction

Speech is one of the easiest modes of interface between the humans and machines. In order to interact with machine through speech, several factors affecting the speech recognition system. Environmental condition, prosodic variations, recording devices , speaker variations etc., are some of the key factors which affect much in getting the good percentage of speech recognition accuracy. Much effort has been incorporated in increasing the performance of the speech recognition systems. In spite of the increased performance, still the output of the speech recognition system contains many errors. In speech recognition system, it is extremely difficulty in dealing such errors. The techniques being investigated and applied on the speech recognition system to reduce the error rate by increasing the speech recognition accuracy. It is very much important to record the speech in good environment with sophisticated recording device. Back ground noise can influence much on the recognition accuracy. Speakers should record the speech clearly, so that good acoustic signals will be generated which will be used for both training phase and decoding phase. It is

important to detect the errors in speech recognition results and then correct those errors by imposing suitable methods. This results in increasing the accuracy of the speech recognition system by reducing error rate. It is important to take care more in dictionary to train the system. The pronunciation dictionary is a mapping table of the vocabulary terms and the acoustic models. It contains the words to be recognized. Incorrect pronunciation in the lexicon causes the incorrectness in training phase of the speech recognition system, which in turn results incorrect results at the decoding phase.

2 Speech Recognition System

Sphinx 3 speech recognition system is used for training and testing. Sphinx is a large vocabulary, speaker independent, continuous and HMM based speech recognition system.

Hidden Markov Model (HMM)

HMM is a method of estimating the conditional probability of an observation sequences given a hypothesized identity for the sequence. A transition probability provides the probability of transition from one state to another state. After particular transition occurs, output probability defines the conditional probability of observing a set of speech features. In decoding phase, HMM is used to determine the sequence of (hidden) states (transitions) occurred in observed signal. And also it determines the probability of observing the particular given state of event that has been determine in first process.

Learning algorithm

Baum welch algorithm find the model's parameter so that the maximum probability of generating the observations for a given model and a sequence of observations.

Evaluation Problem

Forward-backward algorithm is used to find the probability that the model generated the observations for the given model and a sequence of observations.

Decoding Problem

Viterbi algorithm is used to find out the most likely state sequence in the model that produced the observation for the given model and the sequence of observations.

Probabilistic Formulation

Let $A = \{A_1, A_2, \dots, A_t\}$ and $W = \{W_1, W_2, \dots, W_m\}$ be a sequence of acoustic observations and words used. Given acoustic observations A , the probability of word sequence W is $P(W|A)$

$$\begin{aligned} \text{Bayes Rule : } \operatorname{argmax}_W (P(W | A)) &= \operatorname{argmax}_w (P(W, A) / P(A)) \\ &= \operatorname{argmax}_W (P(W, A)) = \operatorname{argmax}_W (P(A | W) * P(W)) \end{aligned}$$

A model for the probability of acoustic observations given the word sequence, $P(A | W)$, is called an "acoustic model. A model for the probability of word sequences, $P(W)$, is called a "language model".

3 Related Work

In order to study the number of words accessed and the relevant words among the accessed words, measures are being calculated. Classification Error Rate (CER), Detection cost function (DCF) and Detection Error Trade-off (DET) are calculated to determine the errors in the speech recognition system [1]. Possible metrics have been developed to evaluate the error measures which in turn describe the performance of the system. Error division's diagrams also used to have the considerable information to user by evaluating the sensitivity and recall metrics [2]. Evaluation of speech recognition has become the practical issue. The common thing is the information retrieval is one of the practical issues. In this, it is necessary to determine the word error rate and accuracy of retrieval of desired document [3]. Hits rates are calculated to know the correct recognitions. The harmonic mean of precision and recall is one such measure used to determine the cost to the user to have different types of errors. Slot error also calculated in order to overcome the limitations of the other measures [4]. Many errors occur at different stages. Lexicon also plays a key role in speech recognition. Pronunciation dictionary consists of all the possible pronunciations of all the speakers. Different speakers may pronounce the same word differently and in some cases same speaker pronounce the same word differently in different contexts. This is due to dialectal variations, educational qualifications and emotional conditions and so on. These variations increase the word error rate. If training data covers all variations, more probability is there to improve the accuracy rate [5]. Dictionary should be developed for all the possible pronunciations. Depending on the pronunciation context and the frequency of that word also affects the accuracy of the system [6]. Compound and individual word in the training and testing influence the accuracy of the system. Proper Training is required so that all the language models built properly. It is required to remove errors corresponding to the transcript during the computation of word error rate [7] [8]. If all the acoustic models are exactly mapped to vocabulary units, then the effective word rate should be zero, practically it is difficult to achieve. Mis-recognized words occur due to the absence of all pronunciation variations by all speakers used in training which is the cause for the low performance of the speech recognition system [9]. Stochastic learning of lexicons is necessary for the spontaneous speech and lecture speech etc. [10]. Accent and speech rate also influence the pronunciation dictionary. Vowel compression and expansion are mostly observed which are very difficult to represent in the pronunciation dictionary [11]. These are also cause to occur more confusion pairs which degrades the performance of the system. Confusion pairs will increase enormously in the case of out of vocabulary [12].

4 Major Types of Errors

Signal processing and Linguistic processing influence the accuracy of the speech recognition system. If the speakers are not recorded properly, more error rate will occur at the decoder phase. So the recording should be properly maintained. The

following are the some of the errors analysis are taken from the speech decoder to analyze the type of error. REF indicates the transcription used for the sphinx speech recognition system and REF indicates the hypothesis obtained from the decoder of the sphinx speech recognition system.

Type 1

Misrecognition occurs due to substitution of one word occurs in the place of original word. This substitution reduces the performance of the speech recognition system.

REF: CHITTOORKI velle train peyremiti

HYP: CHITTOORKU velle train peyremiti

Type 2

Mis recognition occurs because of the substitution of multi words in the place of original single word and also inserting new word E which reduced accuracy.

REF: thirumala ekspres eppudu VASTHUNDHI

HYP: thirumala ekspres eppudu EEST VUNDHI

Type 3

Mis recognitions occur due to the substitution of single word in the place of multiple words. This degrades the accuracy of the system.

REF:ANOWNSEMENT ELA CHEYALI

HYP:GANTALAKU VELUTHUNDHI

Type 4

This type of error occurs due to the out of vocabulary (OOV) situation. Sometimes, decoder some times fail to map the approximate word.

REF: AJHANTHA EKSPRES EKKADIKI VELUTHUNDHI

HYP: ENTHA EKPRES EKKADIKI VELUTHUNDHI

5 Pronunciation Dictionary Modification Method

After analyzing the type of the errors, it is necessary to recover from the errors to improve the accuracy. The knowledge sources of acoustic model, lexicon and language model need improvements. Error patterns are observed from the confusion pairs obtained from the decoder of the speech recognition system. More the frequency of the error patter in confusion pairs, more the sentences in test data are recognized as incorrect. Confusion pairs is useful to analyze the errors occurred in the recognition results. This confusion pairs collects the information of frequency of every possible recognition error. The more two words are confuse each other, the closer they are. It generally refers to the hits, substitutions, deletions, insertions. If the frequency is n in confusion pairs, then n number of times that particular word is recognized as wrong. It is necessary to reduce the value of n. The number of confusion pairs also increased as the frequency of n increases. In order to reduce the number of confusion pairs, one of the recovery techniques called pronunciation dictionary modification method is to be applied. This method reduce the frequency of the error patterns in the confusion pairs to reduce the error rate In order to

reduce the number of confusion pairs, it is desirable to change the phonemes of confused words in the dictionary, so that correct recognition occurs. After modification in the pronunciation dictionary, it is given to the decoder. In other words, it lists the possible phone pronunciations that represent the uttered word with associated probability. This method is used to minimize the errors which are discussed in the previous section.

6 Clustering Method for Random Selection for Testing Set

Conventional clustering is implemented for testing. In hard clustering, one member (here speaker) should belong to one cluster and it should not share the membership value with other clusters. In other words, the membership value for each member should be either 0 or 1. The clusters for these speakers into K non-empty groups or classes $C_1, C_2 \dots C_K$ are a partition in such a way that each cluster should have exactly same size, i.e (i) $C_i \cap C_j = \Phi$ for all i & j (disjoint classes) ; (ii) $C_1 \cup C_2 \cup \dots C_K = X$ (all objects belongs a class) and (iii) $size(C_i) = size(C_j)$

7 Experimental Results

7.1 Speech Corpus

Telugu language is one of the widely used South Indian Languages. 10 male speakers and 10 female speakers are uttered with sophisticated microphone. 50 queries related to Telugu Railway Inquiry, thus it becomes total 1000 queries.

7.2 Results

The following table denotes the percentage of accuracy and the number of confusion pairs occurred before and after modification of the dictionary.

Table 1. No. of Confusion Pairs and % of accuracy (before and after PDMM)

No of Speakers	Total words	Total words recognised		% of accuracy		Confusion pairs	
		Before PDMM	After PDMM	Before PDM	After PDMM	Before PDMM	After PDMM
1S	237	234	235	98.734	99.156	01	0
2S	474	467	469	98.523	98.945	02	0
4S	948	934	938	98.523	98.945	04	0
5S	1185	1161	1178	97.975	99.409	15	0
10S	2370	2281	2343	96.245	98.861	41	08

From the speech recognition decoder, the substitutions (SUB), insertions (INS), deletions (DEL), misrecognitions, total errors collected for the given data. Word Error Rate (WER) and ERROR-RATE are determined as follows:

$$WER = (S+D+I)/N \text{ --- (1) } \&\& \text{ ERROR-RATE} = (S+D+I)/N+I \text{ (2)}$$

Table 2. WER and ERROR-RATE before the application of PDMM

Before	SUB	INS	DEL	Mis recog	Total Errors	WER	ERROR-RATE
1S	01	0	02	01	03	1.265	1.265
2S	02	0	04	02	07	1.265	1.265
4S	04	0	10	04	14	1.477	1.477
5S	13	02	05	13	26	1.687	1.684
10S	26	07	09	30	105	1.772	1.767

Table 3. WER and ERROR-RATE before the application of PDMM

After	SUB	INS	DEL	MIS Recog	Total Errors	WER	ERROR-RATE
1S	0	0	02	0	02	0.844	0.844
2S	0	0	04	0	05	0.843	0.843
4S	0	0	10	0	10	1.055	1.055
5S	0	02	05	0	09	0.590	0.589
10S	05	06	08	07	49	0.802	0.799

From the above tables 2 and table 3, the hits (number of correctly recognized) and false alarms (number of words incorrectly recognized as true) are determined. Hit rate is improving and False Alarm is reducing with the modification in the dictionary, which are shown in the following figures.

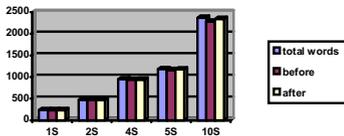


Fig. 1. Hit rates before and after PDMM

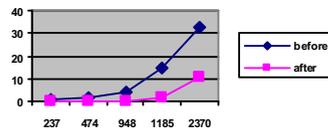


Fig. 2. False alarms before and after PDMM

7.3 F-Measure

Precision and Recall is used to measure the performance of the system. Precision is ratio of the correctly recognized words (C) to the substitutions (S) and insertions (I)

errors. Recall is ratio of correctly recognized words (C) to the substitution and deletion (D) errors. F-measure is weighted combination of precision and recall. It is also called as error measure to evaluate the performance of the system.

$$\text{Precision} = C / (C+S+I) \text{ --- (3) } \&\& \text{ Recall} = C / (C+S+D) \text{ (4)}$$

$$\text{F-measure} = (2 * \text{precision} * \text{Recall}) / (\text{precision} + \text{Recall}) \text{ (5)}$$

Table 4. Precision, Recall and F-measure values before and after PDMM

speakers	Precision		Recall		F-measure		E=(1-F)	
	Before PDMM	After PDMM	Before	(After)	before	after	before	after
1S	0.996	1	0.987	0.992	0.99	0.99	0.01	0.01
2S	0.996	1	0.987	0.992	0.99	0.99	0.01	0.01
4S	0.995	1	0.985	0.989	0.98	0.99	0.02	0.01
5S	0.987	0.998	0.984	0.996	0.98	0.99	0.02	0.01
10S	0.985	0.995	0.985	0.994	0.98	0.99	0.02	0.01

From the table 2, table 3 and table 4, it is observed that the $E \leq \text{ERROR-RATE} \leq \text{WER}$, [4] which are shown in the following figures.

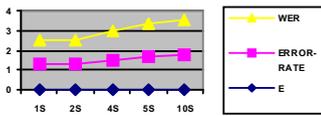


Fig. 3. $E \leq \text{ERROR-RATE} \leq \text{WER}$ (before)

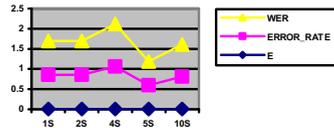


Fig. 4. $E \leq \text{ERROR-RATE} \leq \text{WER}$ (after)

8 Conclusion

In this paper, Errors are analyzed to apply the improving techniques to increase the accuracy of the speech recognition system. Error measures are discussed.

References

1. Pellegrini, T., Transcoso, I.: Improving ASR error detection with non-decoder based features. In: INTERSPEECH, pp. 1950–1953 (2010)
2. Minnen, D., Westeyn, T., Starner, T., Ward, J.A., Lukowicz, P.: Performance Metrics and Evaluation Issues for Continuous Activity Recognition. Performance Metrics for Intelligent System, 141–148 (2006)
3. McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P.: On the Use of Information Retrieval Measures for Speech Recognition Evaluation. IDIAP Research Report (2005)

4. Makhoul, J., Kubala, F., Schwartz, R., Weishede, R.: Performance Measures for information Extraction. In: DARPA Broadcast News Workshop, pp. 249–252 (1999)
5. Bourlard, H., Hermansky, H., Morgan, N.: Towards increasing speech recognition error rates. *Speech Communication*, 205–231 (1996)
6. Davel, M., Martirosian, O.: Pronunciation Dictionary Development in Resource-scarce Environments. In: INTERSPEECH, pp. 2851–2854 (2009)
7. Chen, Z., Lee, K.-F., Lee, M.-J.: Discriminative Training on Language Model. In: ICSLP 2000, pp. 16–20 (2000)
8. Hong-Kwang, Kuo, J., Fosler-Lussire, E., Jiang, H., Lee, C.-H.: Discriminative Training of Language models for Speech Recognition. In: ICASSP 2002, pp. 325–328 (2002)
9. Martirosian, O.M., Davel, M.: Error analysis of a public domain pronunciation dictionary. In: PRASA, pp. 13–16 (2007)
10. Badr, I., McGraw, I., Glass, J.: Pronunciation Learning from Continuous Speech. In: INTERSPEECH, pp. 549–552 (2011)
11. Benus, S., Cernak, M., Rusko, M., Trnka, M., Darjaa, S.: Adapting Slovak ASR for native Germans speaking Slovak. In: EMNLP, pp. 60–64 (2011)
12. Karanosou, P., Yvon, F., Lamel, L.: Measuring the confusability of pronunciations in speech recognition. In: 9th International Workshop on Finite State Methods and Natural Language Processing, pp. 107–115. Association for Computational Linguists (2011)