*Sādhanā* Vol. 37, Part 6, December 2012, pp. 747–761. © Indian Academy of Sciences

# Error analysis to improve the speech recognition accuracy on Telugu language

N USHA RANI* and P N GIRIJA

Department of Computer and Information Sciences, University of Hyderabad, Hyderabad 500 046, India
e-mail: usha552@yahoo.com

MS received 8 October 2011; revised 4 July 2012; accepted 16 July 2012

**Abstract.** Speech is one of the most important communication channels among the people. Speech Recognition occupies a prominent place in communication between the humans and machine. Several factors affect the accuracy of the speech recognition system. Much effort was involved to increase the accuracy of the speech recognition system, still erroneous output is generating in current speech recognition systems. Telugu language is one of the most widely spoken south Indian languages. In the proposed Telugu speech recognition system, errors obtained from decoder are analysed to improve the performance of the speech recognition system. Static pronunciation dictionary plays a key role in the speech recognition accuracy. Modification should be performed in the dictionary, which is used in the decoder of the speech recognition system. This modification reduces the number of the confusion pairs which improves the performance of the speech recognition system. Language model scores are also varied with this modification. Hit rate is considerably increased during this modification and false alarms have been changing during the modification of the pronunciation dictionary. Variations are observed in different error measures such as F-measures, error-rate and Word Error Rate (WER) by application of the proposed method.

**Keywords.** Speech recognition; pronunciation dictionary modification method; error analysis; F-measure.

## 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 affect the speech recognition system. Environmental condition, prosodic variations, recording devices, speaker variations, etc., are some of the key factors which affect much in getting good percentage of speech recognition accuracy. Much efforts have 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

*For correspondence

contains many errors. In speech recognition system, it is extremely difficult to deal such errors. The techniques are being investigated and applied on the speech recognition system to reduce the error rate by increasing the speech recognition accuracy.

It is very 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 obtained from speech recognition system and correct those errors by imposing suitable methods. This results in increasing the accuracy of the speech recognition system by reducing error rate. Care should be taken to develop good transcription and pronunciation dictionary. It is important to take more care in training the speech recognition system. The pronunciation dictionary is a mapping table which consists 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 causes incorrect results at the decoding phase.

## 2. Speech recognition system

Speech recognition is the process of receiving the speech waveform (acoustic signal) and converting the acoustic signal in to the text form which is similar to the uttered speech. Speech signal is one-dimensional time-varying signal. The aim of the speech recognition is to recover the most likely word sequence $W = w_1,\ldots..,w_n$ from the available possible sequence hypothesis for the given sequence acoustic features $A = a_1\ldots..,a_n$ for the acoustic signal.

P(W|A) is computed as

$$P(W|A) = \prod_{a_i} \max_{w_i} P(w_i|a_i).$$

Applying the Bayes rule:

$$\arg\max_{w} P(W|A) = \arg\max_{w} \quad P(A|W)P(W).$$

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'. Sub-words units such as phones form the acoustic model. To map the sub-word units to the word sequences, 'lexicon' or 'pronunciation dictionary' is necessary. In the present work, Sphinx-3 speech recognition system is used for training and testing. Sphinx-3 is a large vocabulary, speaker independent, continuous and HMM (hidden Markov models) based speech recognition system.

### 2.1 *Hidden Markov Models (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 determined in first process.

### 2.2 *Learning*

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.

### 2.3 *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.

### 2.4 *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.

## 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 (Pellegrini & Transcoso 2010). Different types of errors will affect 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 (Minnen *et al* 2006). Evaluation of speech recognition has become the practical issue. Information retrieval through speech input is one of the practical issues. In this, it is necessary to determine the word error rate and accuracy of retrieval of desired document (McCowan *et al* 2005). Hits rates are calculated to know the correct recognitions. The harmonic means 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 (Makhoul *et al* 1999). Many errors occur at different stages. Lexicon or pronunciation dictionary 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 (Bourlard *et al* 1996). 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 (Davel & Martirosian 2009). 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 (Chen *et al* 2000; Kwang *et al* 2002). 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. Misrecognized words occur due to the absence of all pronunciation variations by all speakers used in training. This absence also cause for the low performance of the speech recognition system (Martirosian & Davel 2007).

## 4. 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 obtained from the speech recognizer (decoder) to analyse the type of error. REF (Reference) indicates the transcription used for the sphinx speech recognition system and HYP (Hypothesis) indicates the hypothesis obtained from the decoder of the sphinx speech recognition system.

### 4.1 *Type 1*

Misrecognition occurs due to substitution of a word 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
SENTENCE
Correct = 75.0% 3
Errors = 25.0% 1

The word CHITTOORKU is recognized in the place of CHITTOORKI. This type of substitution is due to confusion between the two words as the distance between their phone sets is very small. Because of this confusion, 75% accuracy is obtained.

### 4.2 *Type 2*

Misrecognition occurs because of the substitution of multi words in the place single word and also inserting extra new word which increases the word error rate.

REF: YASHWANTHPUR EKSPRES KAACHIGOODAKU EPPUDU VASTHUNDHI
HYP: AVUTHUNDHI VAITING VELLE VELLE VAITING E VUNTUNDHI VELLE VUNDHI VAITING
SENTENCE
Correct = 0.0% 0
Errors = 200.0% 10

YASHWANTHPUR ==> VELLE VAITING
KAACHIGOODAKU ==> VUNTUNDHI VELLE
EPPUDU ==> VELLE VUNDHI

From the above case, it is observed that the more words are recognized in the place of single word. In this case, error rate drastically increases. More insertions will occur in this situation.

### 4.3 *Type 3*

Misrecognitions occur due to the substitution of single word in the place of multiple words. In this case, words are deleted which in turn degrades the performance of the system.

REF: RIJERVESHAN ELA CHEYAALI
HYP: SABARI MAYL
SENTENCE
Correct = 0.0% 0
Errors = 100.0% 3
ELA CHEYAALI ==> MAYL

### 4.4 *Type 4*

This type of error occurs due to the occurrence of new word given at the decoding phase or the out of vocabulary. This type of errors mostly occurs in very large vocabulary continuous speech recognition systems. The decoder some times fails to map the approximate word.

## 5. Pronunciation dictionary modification method (PDMM)

After analysing the different types 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. Confusion pairs is useful to analyse the errors occurred in the recognition results. Two words are confused means that the phone set corresponding to the two words are mostly similar in nature. In the confusion pairs, it is necessary to observe the number of times (frequency of the word) the word is confused in recognizing the correct word. If the frequency is n in confusion pairs, then n number of times that particular word is recognized as incorrect due to confusion. This confusion is measured using Levenshtein distance method. It is necessary to reduce the value of n. Let $W_i$ and $W_j$ be a confusion pair obtained from the speech recognition system decoder. Update the phone set of $W_j$ with $W_i$ in the pronunciation dictionary of the decoder. This method reduces the frequency of the confusion pairs in turn reduces error rate by improving word accuracy of the speech recognition system.

### 5.1 *Experimental results*

5.1a *Speech corpus*: Telugu language is one the south Indian languages which is used to develop the speech corpus. 10 Male and 10 female speakers are asked to utter 50 queries related to Telugu railway inquiry. Totally 1000 queries are used in the speech recognition system. Sophisticated microphone is used for the recording speech corpus in a noise free environment.

### 5.2 *Variations of language model scores before and after modification of pronunciation dictionary modification method*

5.2a *Case 1: Here $k = 20$*: In this case, total speech is divided in to 20 data sets (each data set consists of *one speaker's utterances*). One data set *(one speaker's utterances)* is taken randomly for testing and remaining 19 speakers are taken for the training. The accuracy obtained in this case is 97.89% with five confusion pairs. Out of the 50 sentences, only 3 sentences are decoded as wrong. After applying the pronunciation dictionary modification method on the decoder dictionary, the accuracy increased to 100% by reducing confusion pairs to zero. It has been observed from tables 1 and 2 that the LMSCORE (language model score) is being changed after the application of pronunciation dictionary modification.

**Table 1.** Before PDMM.

| Confused word | LMSCORE |
|---|---|
| CHEYYAALI | −164073 |
| PLAATFAAM | −200905 |
| CHITTOORKU | −698020 |
| E | −602023 |
| TAIM | −678545 |

**Table 2.** After PDMM.

| Word corrected | LMSCORE |
|---|---|
| CHEYAALI(1) | −133825 |
| PLAATFAARM(1) | −179710 |
| CHITTOORKI(1) | −580316 |
| EHTIYAM(1) | −571262 |
| TAIM | −634494 |

**Table 3.** Before PDMM.

| Confused word | LMSCORE |
|---|---|
| SHRI | −585199 |
| SABARI | −519364 |

**Table 4.** After PDMM.

| Corrected word | LMSCORE |
|---|---|
| SREE(1) | −585199 |
| CHEYAALI(1) | −583517 |

**Table 5.** Before PDMM.

| Confused word | LMSCORE |
|---|---|
| THIRUMALA | −453529 |
| BUKING | −519364 |
| DELLI | −620681 |
| PURIKI | −309689 |

5.2b *Case 2: $k = 10$*: In this case, total corpus segmented into equal 10 data sets (each data consists of *2 speakers*). From the 10 data sets, one dataset (*2 speakers*) randomly selected for testing. The confusion pairs are reduced from 2 to 0 after applying the PDMM. In this case, it is interesting point to be noted that the language model score does not change for the words SHRI and SREE even after applying PDMM. Still the error rate reduced such that the accuracy increased to the percentage of 100. The variation of LMSCORES is shown in tables 3 and 4.

5.2c *Case 3: $k = 5$*: In this case, total corpus is divided into 5 data sets (each data set consists of *4 speakers*) and one data set (*4 speakers*) is selected randomly for testing. The confusion pairs are reduced to increase the accuracy. The variations of language model scores after modifying the dictionary are observed from tables 5 and 6.

5.2d *Case 4: $K = 4$*: In this case, total corpus is divided into 4 datasets (each data set consists of *5 speakers' utterances*) and one data set (*5 speakers' utterances*) is selected randomly for testing. The confusion pairs are reduced from 14 to 1 after the dictionary modification performed in the decoder. The variations of language model scores (LM SCORES) are observed in the following tables. The language model scores of the correct word is substituted, thus the confusion reduced with this modification of the decoder dictionary which are shown in tables 7 and 8.

**Table 6.** After PDMM.

| Corrected word | LMSCORE |
|---|---|
| ELA(1) | −583517 |
| VAITING(1) | −519364 |
| VELLE(1) | −81898 |
| ENNI(1) | −81898 |

**Table 7.** Before PDMM.

| Confused word | LMSCORE |
|---|---|
| EVARINI | −603562 |
| SABARI | −519364 |
| TRAIN | −517806 |
| PEYREMITI | −584752 |
| PURIKI | −603562 |
| KRAANTHI | −620681 |
| VUNTUNDHI | −630922 |
| TRAIN | −503109 |
| EVARINI | −603562 |

**Table 8.** After PDMM.

| Corrected word | LMSCORE |
|---|---|
| SABARI(1) | −519364 |
| RIJERVESHAN(1) | −453529 |
| ELA(1) | −147781 |
| EMITI(1) | −81898 |
| THIRUPATHI(1) | −519364 |
| NUNDI(1) | −81898 |
| VELUTHUNDHI(1) | −243845 |
| ELA(1) | −147781 |
| TRAIN(1) | −503109 |

5.2e *Case 5: $k = 2$*: In this case, Corpus is divided into two sets consists of 10 speakers in each set. Randomly selecting one among the two sets is used for testing. The number of confusion pairs reduces by the application of decoder dictionary modification. With this modification, the accuracy of the speech recognition increased in considerable manner.

The tables 9 and 10 denote the language mode score variation when 10 speakers are used for training and 10 speakers are used for testing. More words are recognized incorrectly due to confusion between the phone sets.

Tables 9, 10 and 11 denote the total words recognized and the number of confusion pairs occurred without any modification performed in the dictionary and after performing modification in dictionary.

**Table 9.** Before PDMM.

| Confused word | LMSCORE |
|---|---|
| VUNDHI | −488498 |
| AVUTHUNDHI | −603562 |
| VAITING | −610032 |
| VELLE | −529082 |
| VELLE | −528094 |
| VAITING | −620463 |
| E | −621451 |
| VUNTUNDHI | −620577 |
| VELLE | −517663 |
| VUNDHI | −487510 |
| VAITING | −610032 |
| VAITING | −519364 |
| THIRUMALA | −453529 |
| RAAYALASEEMA | −453529 |
| RAAYALASEEMA | −519364 |
| TRAIN | −517806 |
| VUNDHI | −488498 |
| AVUTHUNDHI | −603562 |
| VAITING | −610032 |
| VELLE | −529082 |
| VELLE | −528094 |
| VAITING | −620463 |
| E | −621451 |
| VUNTUNDHI | −620577 |
| VELLE | −517663 |
| VUNDHI | −487510 |
| VAITING | −610032 |
| VELLE | −517663 |
| VUNDHI | −487510 |
| VAITING | −610032 |
| RAAYALASEEMA | 519364 |
| TRAIN | −517806 |
| SABARI | −519364 |
| KERALA | −519364 |
| EPPUDU | −456160 |
| PALANAADU | −677520 |
| VASTHUNDHI | −570350 |
| SABARI | −519364 |
| KERALA | −519364 |
| EPPUDU | −456160 |
| PALANAADU | −677520 |
| VASTHUNDHI | −570350 |

**Table 10.** After PDMM.

| Word(after) | LMSCORE |
|---|---|
| VUNNAAYI(1) | −81898 |
| YASHWANTHPUR(1) | −610032 |
| EKSPRES(1) | −81898 |
| VELLE(1) | −588409 |
| YASHWANTHPUR(1) | −620463 |
| EKSPRES(1) | −81898 |
| KAACHIGOODAKU(1) | −81898 |
| VELLE(1) | −526897 |
| EPPUDU(1) | −455172 |
| VASTHUNDHI(1) | −239418 |
| ANOWNSMENT(1) | −519364 |
| ELA(1) | −583517 |
| BANGALOOR(1) | −519364 |
| EKSPRES(2) | −81898 |
| VUNNAAYI(1) | −81898 |
| AVUTHUNDHI(1) | −603562 |
| YASHWANTHPUR(1) | −610032 |
| EKSPRES(1) | −81898 |
| VELLE(1) | −588409 |
| YASHWANTHPUR(1) | −620463 |
| EKSPRES(1) | −81898 |
| KAACHIGOODAKU(1) | −81898 |
| VELLE(1) | −526897 |
| EPPUDU(1) | −455172 |
| VASTHUNDHI(1) | −239418 |
| BANGALOOR(1) | −519364 |
| EKSPRES(2) | −81898 |
| RIJERVESHAN(1) | −453529 |
| THELANGAANA(1) | −519364 |
| ETU(1) | −81898 |
| VAARAMLO(1) | −81898 |
| ANNI(1) | −120440 |
| RIJERVESHAN(1) | −453529 |
| THELANGAANA(1) | −519364 |
| ETU(1) | −81898 |
| VAARAMLO(1) | −81898 |
| ANNI(1) | −120440 |

**Table 11.** Number of confusion pairs and % of accuracy before and after PDMM.

| | | No. of words recognized | | % of accuracy | | No. of confusion pairs | |
|---|---|---|---|---|---|---|---|
| No. of speakers(test) | Total no. words | Before PDMM | After PDMM | Before PDMM | After PDMM | Before PDMM | After PDMM |
| One speakers | 237 | 232 | 237 | 97.89 | 100 | 5 | 0 |
| Two speakers | 474 | 470 | 472 | 99.16 | 99.58 | 2 | 0 |
| Four speakers | 948 | 941 | 945 | 99.26 | 99.68 | 5 | 1 |
| Five speakers | 1185 | 1161 | 1178 | 97.97 | 99.41 | 14 | 1 |
| 10 speakers | 2370 | 2332 | 2358 | 98.39 | 99.49 | 17 | 3 |

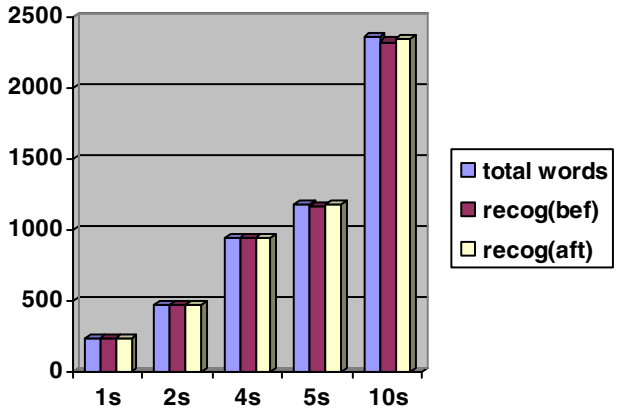


**Figure 1.** Total words recognized before and after PDMM.

Figure 1 shows the total number of words recognized before and after the modification in the decoder dictionary. Figure 1 clear shows the improvement in recognition of words after the modification applied on the decoder dictionary. Number of confusion pairs obtained before and after modification method applied in the decoder dictionary is shown in figure 2. Confusion pairs are reduced after the modification in the decoder dictionary is clearly seen in figure 2.

From the speech recognition decoder, the substitutions (Sub), insertions (Ins), deletions (Del), misrecognitions (Mis recog), total errors occurred for the given data are examined. From these values, word error rate (WER) and error-rate will be calculated as follows:

$$\text{WER} = (\text{Sub} + \text{Del} + \text{Ins})/\text{N}. \tag{1}$$

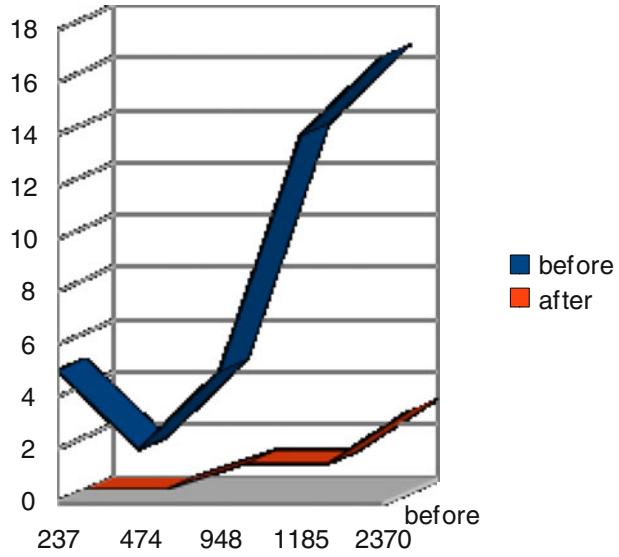


**Figure 2.** Total words vs confusion pairs.

**Table 12.** WER and error-rate before application of PDMM.

| No. of speakers | Total words | Recognized words | Sub | Ins | Del | Mis recog | Errors | WER | Error-rate |
|---|---|---|---|---|---|---|---|---|---|
| 1 speaker | 237 | 232 | 5 | 3 | 0 | 5 | 8 | 3.37 | 3.33 |
| 2 speakers | 474 | 470 | 2 | 0 | 2 | 2 | 4 | 0.84 | 0.84 |
| 4 speakers | 948 | 941 | 6 | 2 | 1 | 6 | 9 | 0.95 | 0.94 |
| 5 speakers | 1185 | 1161 | 19 | 0 | 5 | 19 | 24 | 2.03 | 2.03 |
| 10 speakers | 2370 | 2332 | 35 | 21 | 3 | 35 | 59 | 2.48 | 2.46 |

**Table 13.** WER and error-rate after the application of PDMM.

| After | Total words | Recognized words | Sub | Ins | Del | Mis recog | Errors | WER | Error-rate |
|---|---|---|---|---|---|---|---|---|---|
| 1 speaker | 237 | 237 | 0 | 3 | 0 | 0 | 3 | 1.27 | 1.25 |
| 2 speakers | 474 | 472 | 0 | 0 | 2 | 0 | 2 | 0.42 | 0.42 |
| 4 speakers | 948 | 945 | 2 | 2 | 1 | 2 | 5 | 0.53 | 0.52 |
| 5 speakers | 1185 | 1178 | 3 | 1 | 4 | 3 | 8 | 0.68 | 0.67 |
| 10 speakers | 2370 | 2358 | 9 | 23 | 3 | 9 | 35 | 1.47 | 1.46 |

**Table 14.** Improvement of hit rate and reducing the false alarms.

| | Hits | | False alarms | |
|---|---|---|---|---|
| No. of speakers | Before | After | Before | After |
| 1 speaker | 232 | 237 | 8 | 3 |
| 2 speakers | 470 | 472 | 2 | 0 |
| 4 speakers | 941 | 945 | 8 | 4 |
| 5 speakers | 1161 | 1178 | 24 | 4 |
| 10 speakers | 2332 | 2358 | 56 | 32 |

$$\text{Error-rate} = (\text{Sub} + \text{Del} + \text{Ins})\big/[\text{N} + \text{Ins}]. \tag{2}$$

After examining the number of confusion pairs, pronunciation dictionary is modified by giving priority to the highest frequency rate of confusion pairs. The WER and error-rate is calculated for different datasets before the modification in the dictionary are shown in table 12.

After applying the modification method, the following values are tabulated in table 13. In this case also WER and error-rate is determined from the same equations (1) and (2).

From the above tables, hits (number of correctly recognized) and false alarms (number of words incorrectly recognized) are obtained. The obtained values of hits and false alarms are tabulated in table 14. It is observed that the hit rate is improving with the modification in the pronunciation dictionary, where as false alarms rate is reduced. This shows the performance of the speech recognition system has been improved.

Figures 3 and 4 show the improvement of the hits and reduction of the false alarm rate after the modification in the decoder dictionary.

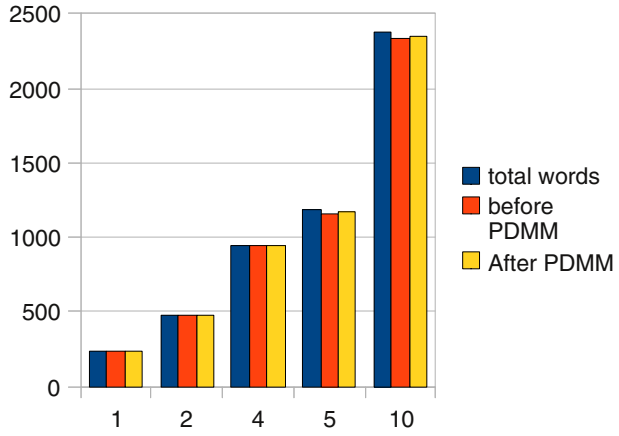


**Figure 3.** Hit rates before PDMM and after PDMM.

From the detection error trade off (DET) curves, it is possible to examine the errors occurred from the speech recognition decoder and in turn observe the performance of the speech recognition system. Figures 5 and 6 show the DET curves obtained before and after the PDMM. From these, it is observed that the reduction of false alarm improves the performance of the speech recognition system.

### 5.3 *F-measure*

F-measure is the weighted combination of the precision and recall. It is also used to call the error measurement technique to evaluate the performance of the system. Precision in speech recognition is the number of correctly recognized to the sum of number of correctly recognized, substitutions and insertions. Recall in speech recognition is the number of correctly

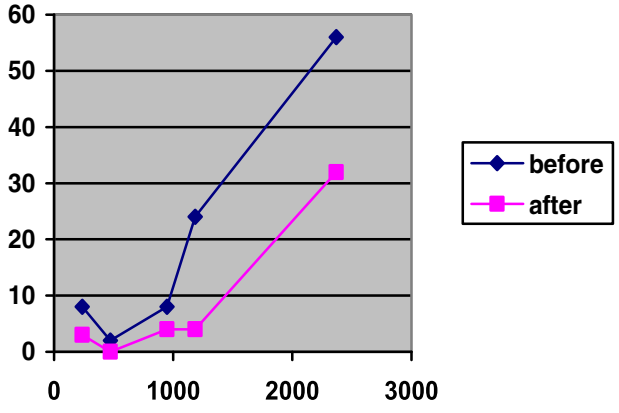


**Figure 4.** Total words vs false alarms before and after PDMM.

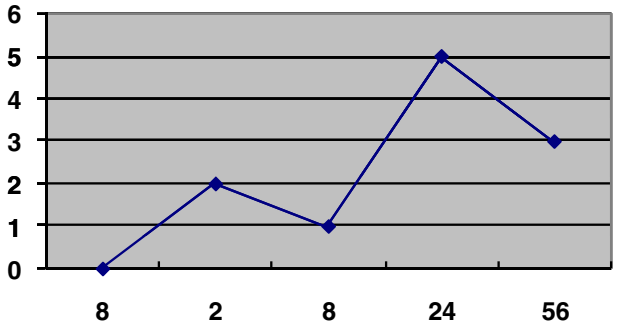


**Figure 5.** DET curve (before PDMM).

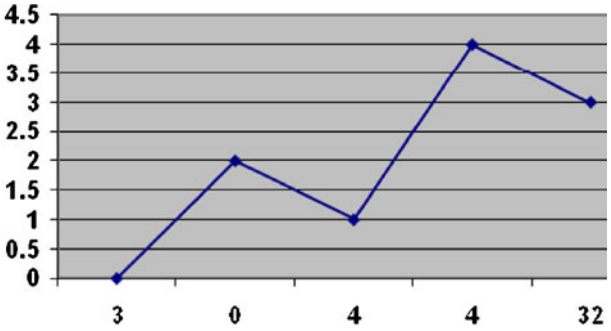


**Figure 6.** DET curve (after PDMM).

**Table 15.** Precision and recall values before and after PDMM.

| | Precision | | Recall | |
|---|---|---|---|---|
| No. of speakers | Before PDMM | After PDMM | Before PDMM | After recall |
| 1 speaker | 0.9666 | 1 | 0.9789 | 1 |
| 2 speakers | 0.9957 | 1 | 0.9915 | 0.9957 |
| 4 speakers | 0.98 | 1 | 0.98 | 0.99 |
| 5 speakers | 0.99 | 1 | 0.99 | 1 |
| 10 speakers | 0.98 | 0.99 | 0.98 | 0.99 |

recognized utterances to the number of correctly recognized, substitutions and deletions. F-measure is calculated using the precision and recall using the following formulae.

Let C = Number of correctly recognized words; S = Number of substitutions
I = Number of insertions; D = Number of Deletions

$$\text{Precision} = \text{C}/(\text{C} + \text{S} + \text{I}) \tag{3}$$

and

$$\text{Recall} = \text{C}/(\text{C} + \text{S} + \text{D}). \tag{4}$$

**Table 16.** F-measure before and after PDMM.

| No. of speakers | F-measure(before) | E = 1-F(before) | F-measure(after) | E = 1-F(after) |
|---|---|---|---|---|
| 1 speaker | 0.97 | 0.03 | 1 | 0 |
| 2 speakers | 0.99 | 0.01 | 0.99 | 0.01 |
| 4 speakers | 0.98 | 0.02 | 0.99 | 0.01 |
| 5 speakers | 0.99 | 0.01 | 0.99 | 0.01 |
| 10 speakers | 0.98 | 0.02 | 0.99 | 0.01 |

**Table 17.** Three errors before PDMM.

| No. of speakers | E | Error-rate | WER |
|---|---|---|---|
| 1 speaker | 3 | 3.33 | 3.37 |
| 2 speakers | 1 | 0.84 | 0.84 |
| 4 speakers | 2 | 0.94 | 0.95 |
| 5 speakers | 1 | 2.03 | 2.03 |
| 10 speakers | 2 | 2.46 | 2.48 |

**Table 18.** Three errors after PDMM.

| No. of speakers | E | Error-rate | WER |
|---|---|---|---|
| 1 speaker | 0 | 1.25 | 1.27 |
| 2 speakers | 1 | 0.42 | 0.42 |
| 4 speakers | 1 | 0.52 | 0.53 |
| 5 speakers | 1 | 0.67 | 0.68 |
| 10 speakers | 1 | 1.46 | 1.47 |

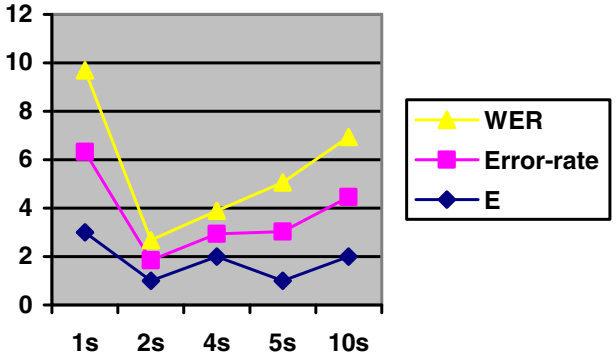


**Figure 7.** E<=Error-rate<=WER(before PDMM).

Then

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}. \tag{5}$$

From (3) and (4), the precision and recall values are obtained as follows.

From table 15, the F-measure is obtained from the (5). The values are tabulated in table 16.

It has been observed that the E<=Error-rate<=WER from tables 17 and 18. The same observation is shown in the pictorial representation in figures 7 and 8. It has been compared the results in Kwang *et al* (2002) to the result of the present work.

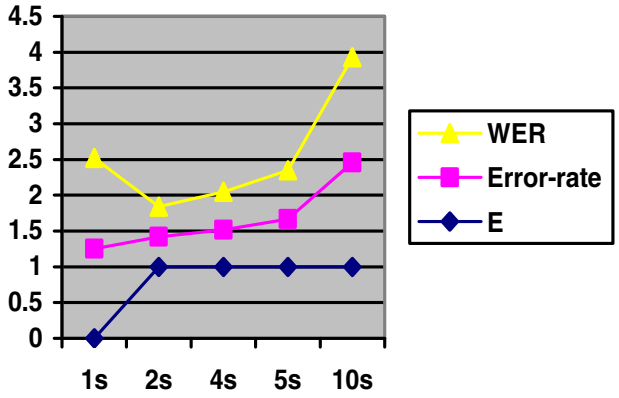


**Figure 8.** E<=Error-rate<=WER(after PDMM).

## 6. Conclusions

This paper describes the different types of errors obtained from the decoder of the speech recognition system. In the present work, the language model score variations after the error recovery have been verified. This paper also reviewed the importance of the pronunciation dictionary used in the speech recognition system. It also describes the importance of modification in the dictionary. Pronunciation dictionary modification method can be used as error recovery technique. This method minimizes the substitution errors by correcting the confused pairs. Error analysis has been performed to apply the pronunciation modification method. Three different types of error measures are determined to show the relationship among them and to verify how the values are varied in each measure.

## References

Bourlard H, Hermansky H and Morgan N 1996 Towards increasing speech recognition error rates. *Speech Communication*, 205–231

Chen Z, Lee K-F and Lee M-J 2000 Discriminative training on language model, ICSLP-2000, *International Conference on Spoken Language Processing*, 16–20

Davel M and Martirosian O 2009 Pronunciation dictionary development in resource-scarce environments, *INTERSPEECH*, 2851–2854

Kwang H, Kuo J, Fosler-Lussire E, Jiang H and Lee C-H 2002 Discriminative training of language models for speech recognition, In: *Proceedings of ICASSP*, 325–328

Makhoul J, Kubala F, Schwartz R and Weishede R 1999 Performance measures for information extraction, In: *Proceedings of DARPA Broadcast News Workshop*, 249–252

Martirosian O M and Davel M 2007 Error analysis of a pubic domain pronunciation dictionary, In: *Proceedings of PRASA*, 13–16

McCowan I, Moore D, Dines J, Gatica-Perez D, Flynn M and Wellner P 2005 *On the use of information retrieval measures for speech recognition evaluation*, IDIAP Research Report

Minnen D, Westeyn T, Starner T, Ward J A and Lukowicz P 2006 Performance metrics and evaluation issues for continuous activity recognition, In: *Performance metrics for intelligent system*, 141–148

Pellegrini T and Transcoso I 2010 Improving ASR error detection with non-decoder based features, *INTERSPEECH*, 1950–1953