The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Marathi Speech Recognition

Supriya Paulose[1,2], *Shikhamoni Nath*[2], *Samudravijaya K*[2]

[1]Computer Engineering, Mukesh Patel School of Technology Management and Engineering, Mumbai 400056, India
[2]Centre for Linguistic Science and Technology,
Indian Institute of Technology Guwahati, Guwahati 781039, India
supriyapaulose@gmail.com, shikhanath2034@gmail.com, samudravijaya@gmail.com

Abstract

The details of the implementation of a Marathi Automatic Speech Recognition (ASR) system as well as the associated speech database is given in this paper. The speech database consists of more than 15000 speech files. Over 1500 speakers read 10 sentences each into their mobile phone. Speech was recorded over telephone channel. The text corpus consists of 3400 sentences consisting of over 10000 unique words.

The ASR system was implemented using Kaldi toolkit. Acoustic models of various characteristics were implemented and 3-fold validation tests were conducted. The word error rate of recognising test data is 24%. Experiments were conducted to study the effect, on the performance of the system, of (a) manual annotation of non-speech events such as cough, babble etc., and (b) discarding of those training speech files which could not be recognised well. The results of these experiments and the lessons learnt are presented.

Index Terms: speech recognition, Marathi, DNN-HMM, annotation of non-speech events.

1. Introduction

Recent advances in spoken language technologies have resulted in bringing the dream of spoken conversations with machines closer to reality. The most successful technology involves employing a Deep Neural Network (DNN). Estimation of a large number of parameters of a DNN demands very large amount of speech data. This becomes a bottleneck in the process of implementing speech systems for languages lacking in such linguistic resources. Most Indian languages, with the possible exception of Hindi, are deemed to be under resourced languages. Here, we describe our effort to develop an Automatic Speech Recognition (ASR) system for Marathi language spoken in western part of India.

Initial efforts of automatic recognition of spoken sentences in Indian languages focussed on the Hindi language. In 1998, a hierarchical speech recognition system that could recognise Hindi sentences spoken with pause between words was reported [1]. In 2004, Kumar et al. reported the development of a large-vocabulary continuous Hindi speech recognition system using a hybrid approach that combines rule-based and statistical approaches [2]. The latest in the series is a Hindi ASR system [3] that was implemented using kaldi toolkit.

During the past decade, development of a couple of Marathi continuous speech recognition systems were reported. In 2005, ASR systems were implemented for 3 Indian languages, including Marathi [4]. Another Marathi ASR system was implemented that used HMMs to model the monophones [5]. These systems used Hidden Markov Model (HMM) in conjunction with Gaussian Mixture Model (GMM) to capture acoustic characteristics of phones. Both the systems used clean speech from a couple of hundred speakers recorded using a few mobile handsets. Here, we present the implementation of a Marathi ASR system that used narrowband speech data collected from 1500 farmers residing in 34 districts of Maharashtra.

2. Speech data

In this section, we give a brief account of the linguistic resources used to train and test Marathi ASR system. A detailed account of creation of text and speech corpora is given in [6].

2.1. Text corpus

The text corpus consists of 340 sentence sets, each containing 10 sentences. Each sentence set was automatically generated by pooling sentences from many sources, viz. six sentences from books, two proverbs, one sentence from online stories, and one digit sequence (of length seven digits), to incorporate variety. The text contains 11, 200 unique words.

2.2. Speech data recording

Speech data was collected from about 50 speakers in each of the 34 districts spread over 6 geographical areas of Maharashtra state as shown in Fig. 1.

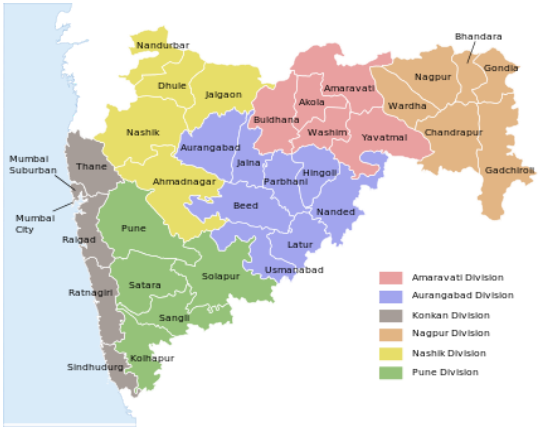

Fig. 1: The 34 districts of the state of Maharashtra (source: wikipedia).

10.21437/SLTU.2018-49

The speakers called a number to read 10 sentences using their own mobile phone. Metadata was recorded using DTMF input. The speech database consists of 15 hours of read speech. About 20% of the speakers were female speakers.

2.3. Annotation and lexicon

Speech data contained a significant amount of different kinds of nonspeech events as well as incomplete and invalid words. So, a scheme of flagging the occurrence of such non-speech sounds was followed. Such non-speech sounds were grouped into 9 categories: (i) background noise, people speaking in the background, (iii) machine generated narrow band sound, (iv) pause, (v) vocal noise, (vi) impulsive sound, (vii) hesitation sound, (viii) laugher, (ix) 'hmm' sound. Transcribers listened to each speech file and marked non-speech sounds with one of the 9 filler-labels. The 14,662 utterances, corresponding to 3400 reference sentences, collected from around 1500 literate speakers contain 94,306 spoken words (complete and incomplete) and 15,155 filler labels.

2.3.1 ILSL

A pronunciation dictionary was created for all words in the lexicon. The labels for phones followed the Indian Language Speech Labels scheme [7]. The dictionary has 13662 entries.

2.4. Division of speech database into 3 sets

The sentence dataset was divided into three equal and balanced sets of speakers namely set A, set B and set C. Each set contains 520 speakers on average, equally distributed across the 34 districts and male-female classes within these districts. This allows us to perform 3-fold cross validation experiments.

3. Experiments

This section describes the experimental setup and results of the experiments conducted.

3.1 Signal processing and models

Mel Frequency Cepstral Coefficients [8] were used to used represent acoustic characteristics of a frame of speech. HMM was used to model the quasi-stationary and statistical nature of speech signal.

3.1.1. Kaldi toolkit

Kaldi [9] is a popular, open source and evolving toolkit for implementation speech recognition systems using various types of models for representing linguistic units. It has provision for using HMMs in conjunction with GMMs or sub space GMM (sGMM) or DNNs. A user can use various kinds of language models using external language model toolkits. In our experiments we used a simple model: bigram language model as estimated by IRSTLM toolkit [10].

The bigram language model was trained using the transcripts of train data alone. Since a sentence was read by multiple speakers, the language model thus trained was adequate to recognise test speech.

Kaldi toolkit allows us to train HMMs to represent different types of linguistic units. The ASR system that utilises HMMs to model context independent phones is denoted as 'Mono' here. Depending on the details of feature and model normalization and adaptation, Kaldi scripts permit us to train 3 types of context dependent phones. These are denoted as Tri1, Tri2 and Tri3 respectively. When a subspace GMM (instead of regular GMM) is used to model a state of HMM, the resulting ASR system is denoted as sGMM in this paper.

3.2 Experimental results and discussion

The performance of ASR system is expressed in terms of Word Error Rate defined as 100 (I+D+S)/N where I, D and S denote the number of word insertion, deletion and substitution errors in the test set, and N denotes the total number of words in the reference transcription.

The second row of Table 1 shows the Word Error Rate (WER) of the baseline systems using different types of acoustic models. While the WER decreases as the level of sophistication of models increase, the WER of test data is very high. The lowest error rate is achieved by subspace GMM model.

Inspection of errors revealed that quite a few speech files do not contain speech, but just background noises. When such speech files that do not contain speech sounds were removed from training data as well as test data, the WER decreased slightly as shown in the 3rd row of Table 1.

3.2.1 Relevance of marking non-speech sounds

Further analysis of errors showed that most word level 'errors' in the decoder output are due to insertion or deletion of filler sound labels. Since these non-speech events do not convey useful information in case of speech recognition, we deleted all such filler labels from both decoder output and reference transcription, just before evaluation. This reduced WER significantly as shown in the 4th row of Table 1. Finally, we removed all such non-speech (filler) labels from all transcriptions and retrained systems with just phone level transcriptions. In other words, non-speech events were not modelled by this system. This reduced WER by a small margin. This observation seems to indicate that ignoring all non-speech events in training data seems to help model speech phones better. We had thought that marking the presence of non-speech events would help the system to detect and ignore non-speech events, and thus focus better on the speech sounds. One possible explanation for the unexpected result is that the 5 emitting state silence model seems to capture all types of non-speech sounds adequately.

Table 1: *Word Error Rates (%) of the baseline ASR system and its derivatives. As the WERs in the last row show, ignoring the labels for non-speech sounds during both training and scoring results in the lowest WER.*

Salient features of systems	Mono	Tri1	Tri2	Tri3	sGMM
Initial system	49.3	43.4	41.9	40.1	38.8
Files without speech removed	47.3	40.4	39.8	38.6	37.1
Ignore noise labels during scoring	35.3	26.5	25.7	24.2	21.8
Ignore noise labels all the time.	35.2	26.3	25.5	24.4	21.6

3.2.2 *Error analysis*

The *score_kaldi.sh* script of kaldi toolkit permits to visually see the types of errors the speech decoder has committed. One can sort the test data according to the number of word level errors the decoder has made. An analysis of such decoding errors w.r.t. *training* data may reveal severe transcription errors in training data. Our hope was to identify such speech files with severe transcription errors and eliminate such files from training set, thus hopefully increasing the quality of training data. The degree of decoder error can be computed as the ratio of no. of word errors to the no. of words in the utterance. If the number of errors is more than the number of words in the reference transcription, then the value of the relative WER will be greater than 1. Figure 2 shows a histogram of such relative word error rate corresponding to training set of about 10,000 speech files.

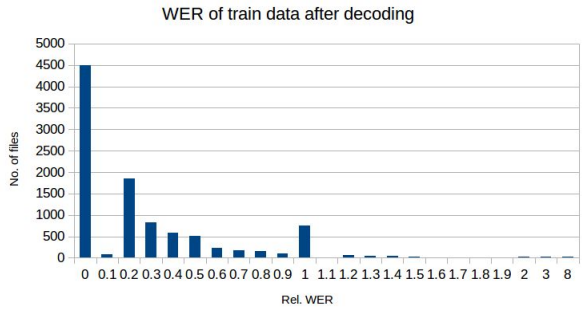

Figure 2: A histogram of relative WER (no. of word errors / no. of words in the utterance) of about 10000 speech files used for training when decoded by the trained model.

Inspection of Fig. 2 shows the presence of more than 100 training data files whose decoder output shows more word errors than the number of words in the corresponding reference transcription (i.e., relative WER > 1). Since the number of such files associated with severe decoder errors is large, we decided to eliminate certain fraction of worst training files from the training set, and retrain the acoustic models.

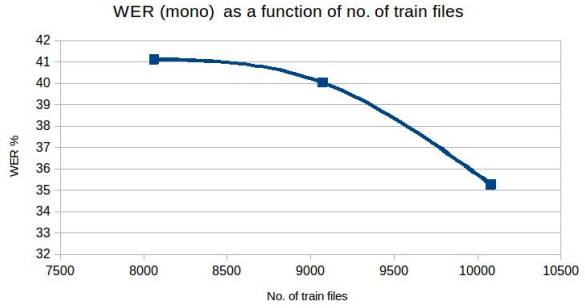

Figure 3: A chart showing a decrease in WER (%) with increasing number of training files. A monophone model was used here.

Fig. 3 shows the WER as a function of training data size. The original training set contained more than 10,000 speech files. The corresponding WER for Mono models is about 35%. When 2000 worst speech files were removed from the training data, the WER increased to 41%. When just 1000 worst speech files were removed from the training data, the WER still remained high at 40%. The lowest WER was obtained when all training files were used to train acoustic models. Thus, it appears that the amount of speech data is more critical for ASR based on sophisticated statistical models than the quality of speech data.

The above inference was reinforced by the following observation. We trained acoustic models by replacing GMMs by either deep neural network or Time delay neural networks. Contrary to our expectations, the WER increased in both cases. DNNs need a large amount of training data. The size of training data is insufficient to adequately estimate the parameters of the model, thus increasing the WER on unseen test data. In view of the of limited amount of data, we did not purse neural net based ASR systems further.

3.2.3 *Cross validation experiments.*

The WER figures presented so far correspond to the case when the dataset A was used as test data and datasets B and C were used for training models; we refer to this as fold-1. We carried out 3-fold cross validation experiments to check that the results of experiments do not significantly depend on division of data into training and test sets.

The WERs of the 3-fold experiments corresponding to 5 types of acoustic models are shown in Table 2. The general trend persists across all 3-folds. The average of the 3 word error rates corresponding to 3-folds is 21.2%, and corresponds to sGMM model.

Table 2: The word error rates (%) of 3-fold cross validation experiments are shown in this table for ASR systems using acoustic models of different detailing. The best performance is obtained by SGMM model, whose average WER (last row) is 21.2%.

Fold	Mono	Tri1	Tri2	Tri3	SGMM
1	35.2	26.3	25.5	24.4	21.6
2	37.1	26.1	25.7	24.3	21.1
3	38.0	26.3	25.6	24.0	20.9
Av	36.8	26.2	25.6	24.2	**21.2**

3.3 Comparison with similar ASR systems

While the WER of the current Marathi ASR system (21.2%) is not low, it is comparable to that of ASR systems implemented for Indian languages.

3.3.1 *Comparison with other Marathi ASR systems.*

The Marathi ASR system reported in 2005 [4], used a speech database that contained 52 sentences each spoken by 176 speakers over narrow band channel. The speakers used one of the four mobile phones or landline phone sets. The size of vocabulary was 21,640 lexical entries. The CMU 11 toolkit [11] was used to implement semi-continuous HMMs that modeled context dependent phones; backoff trigram language model was used. The WER of the test data with the ASR system trained using mobile channel data was 23.6%. This is higher than the WER of the current system (21.2%). While the

size of the lexicons in the cases are comparable, the number of speakers in the current system is about an order of magnitude higher. In the current database, each person used his/her own mobile phone in contrast with 4 mobile phones used by all 176 speakers in [4]. Moreover, the speech data used in the current experiment has a lot of noise from field data. In contrast, in [4], the "recording is clean and has minimal background disturbance. Any mistakes made while recording have been undone by re-recording or by making the corresponding changes in the transcription set". Despite such challenges, the Marathi ASR system reported here shows lower WER. The WER of a monophone HMM based Marathi sentence recognition system was 50% [5]; this WER is higher than that of the monophone system reported here.

3.3.2 Comparison with other Indian language ASR systems.

In this section we report the WERs of ASR systems implemented using kaldi toolkit for Indian languages other than Marathi. An Assamese ASR system that could recognise isolated names was reported in [12]. The WERs for GMM-HMM models were 10.3% and 5.2% for recognition of 109 agricultural Commodity names and 27 District names in Assamese language. Due to the small size of vocabulary, the WER of this system is comparable to that of the current system. Similarly, the phone error rate of Mizo phone recognition system using sGMM-HMM was 15.7% [13].

Recently, Upadhyaya et al. implemented a Hindi ASR system using kaldi toolkit [3]. The authors used context dependent phone models and bigram as language model. The speech database contained 1000 sentences comprising of 2007 unique words, spoken by 100 speakers. The WER of the system was 14.4% . Both the lexicon size and the number of speakers in the database is smaller than those of the current Marathi ASR system.

4. Conclusions

A large vocabulary ASR system for Marathi language was implemented using speech contributed by thousands of speakers. While the recognition accuracy of the system can be enhanced further by fine tuning the parameters, a few of the lessons learnt are as follows. (1) Larger the amount of training data, lower the test WER. This is true even when about 10000 spoken sentences are used to train acoustic and language models. (2) Detailed information about the presence of irrelevant acoustic events such as babble, mono chromatic noise, cough etc. do not seem to lead to better training of acoustic models of relevant events such as phones generated by the speaker. (3) Given the limited training data and large vocabulary size, subspace Gaussian mixture model seem to model linguistic units better. (4) Deep neural networks did not yield better performance than sGMMs due to lack of sufficient data to adequately estimate a large number of parameters of the model. So, if we increase the size of speech corpus for training the system, it is likely to result in a Marathi ASR system with better performance.

5. Acknowledgements

The Marathi sentences speech database was created jointly by the speech groups of IITB Mumbai and TIFR Mumbai, as part of a project sponsored by Department of Electronics and Information Technology (DeitY), Government of India.

6. References

[1] Samudravijaya K et al., "A feature-based hierarchical speech recognition system for Hindi", Sadhana, vol. 23, pp. 313-340, 1998.

[2] M.Kumar, N.Rajput and A.Verma, "A large-vocabulary continuous speech recognition system for Hindi", IBM J. of Res. and Development, vol. 48, pp. 703-715, 2004.

[3] P. Upadhyaya et al., "Continuous Hindi Speech Recognition Model Based on Kaldi ASR Toolkit", Proc. of Int. Conf. on Wireless Communications, Signal Processing and Networking (WiSPNET 2017), Chennai, India.

[4] G.Anumanchipalli et al., "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems", Proc. of Int. Conf. on Speech and Computer (SPECOM), Patras, Greece, 2005.

[5] S.S.Chavan, S.M.Handore, "Speech Recognition using HTK toolkit for Marathi Language", IEEE Int. Conf. PCSI, pp. 1591-1597, 2017.

[6] T.Godambe, N.Bondale, K.Samudravijaya and P.Rao, "Multi-speaker, narrowband, continuous Marathi speech database", Proc. of Oriental COCOSDA conference, Nov. 2013, DOI: 10.1109/ICSDA.2013.6709844

[7] Indian Language Speech Label (ILSL12), https://www.iitm.ac.in/donlab/tts/downloads/cls/cls v2.1.6.pdf

[8] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[9] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, Karel Vesely, "The Kaldi Speech Recognition Toolkit", In Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011, Hawaii, US

[10] IRST language modeling toolkit, https://hlt-mt.fbk.eu/technologies/irstlm

[11] "About CMUSphinx", https://cmusphinx.github.io/wiki/about/

[12] A. Dey et al., "AGROASSAM: A Web Based Assamese Speech Recognition Application for Retrieving Agricultural Commodity Price and Weather Information", accepted for presentation in the Show-and-tell track, Interspeech 2018, Sep 2-6, 2018, Hyderabad.

[13] A. Dey et al., "Mizo Phone Recognition System, Proc. IEEE Indicon 2017, IIT Rourkee, December 15-17, 2017.