EasyChair Preprint

№ 7181

# Building Speech Corpus in Rapid Manner to Adapt a General Purpose ASR System to Specific Domain

Mukund K Roy, Sunita Arora, Karunesh Arora and Shyam S Agarwal



December 7, 2021

# Building speech corpus in rapid manner to adapt a general purpose ASR system to specific domain

Mukund K Roy[1],Sunita Arora[2], Karunesh K Arora[3,] Shyam S Agarwal[4]

[1,2,3] SNLP Lab, CDAC Noida, UP, India
mukundkumarroy@cdac.in
sunitaarora@cdac.in
karuneshaarora@cdac.in
[4] KIIT, Gurugram, Haryana, India
karuneshaarora@cdac.in

**Abstract.** The situation prevalent due to Covid-19 has affected the traditional speech database collection process by reaching out persons in one-to-one manner. In this paper, we describe an alternate approach adopted for faster speech dataset construction in Hindi language for building domain adapted Automatic Speech Recognition (ASR) for agriculture domain. We resorted to two methods – one utilizing App for speech samples collection and second through domain specific YouTube videos. In this paper we outline building of App and several filtering (signature music, advertisements and cross talks) and post-processing steps for speech database collected through on-line videos. The paper also describes novel idea of making speech segments suitable for training an end-to-end ASR system. The process of annotating included combination of utilizing existing ASR systems and manual post correction to save time. Our experiment resulted in collection of speech data from 236 speakers through App and 106 hours of speech data through on-line videos. The experiment of re-training ASR with enhanced data reveals that exercise results in adapting it for a particular domain in a rapid manner.

**Keywords:** Speech corpus, Data collection, Speech recognition.

## 1 Introduction

The technological advancement in the area of speech technology has been substantial in the last decade. The recognition accuracy along with the performance of ASR systems have been continuously improving. This can be well attributed to the research community and thriving ecosystem of open source technologies. In terms, of technology a lot has been achieved and in the coming years we will witness even more advancements in this area.

ASR is indeed an enabler for technology and information proliferation in India, particularly in the agricultural sector. Owing to the large population of farmers in India, engaged in agricultural activities and relatively less ratio of formal education amongst

the population. Thus, ASR for agricultural domain is of absolute importance when it comes to the proliferation of information to empower farmers.
The research and developments in ASR have been collaborative efforts of many researchers across the globe. In a step by step fashion it has been addressed by various direct and cross-sectional approaches with Speech Technology. It has been evident that the technological background of our present ASR system or for that many any AI/ML system rely heavily on data. This, at the beginning was thought of a solution that could cater to the majority of problems at that time. However, in the due course of time, the researchers realized this approach was beneficial to languages with data in abundance. On the other side low-resource languages struggled to cope up with the speech data requirements of ASR or TTS systems. This has been a bottle-neck and still remains one for the low-resource languages. Nevertheless, speech data collection is not only expensive in terms of cost but also in pandemic situation becomes extremely difficult in absence of a proper crowd-sourced model.
We, in this paper, leverage the open community; YouTube to acquire Speech data for Agriculture domain. The motivation for the same has been the high availability of domain relevant videos, which thrives on active user community. This addresses the standard issues with crowd source model wherein we have seen the lack of interest among users after contributing for some time and relatively less repeat users. In revenue based crowd source models, re-users are even discouraged. Thus, if the YouTube videos are somehow channelized for ASR data this in turn could aid low resource languages across the globe to develop robust ASR system.

## 2 Literature Survey

In the last few years numerous efforts have been successful in building datasets via Crowd source strategy. Nevertheless, the approach was initially seen as solution to ever increasing data demands of ML based algorithms, soon it lost the shine. This was attributed to difficulty in keeping users motivated and engaged for contributing over longer periods of time. Though, Mozilla's Common Voice and Vox Forge dataset have gained significantly from such efforts. Many other efforts have been successful such as LibriSpeech (Panayotov et al., 2015) [12] comprising of audio books, alongside other data sets like TED talks data (Rousseau et al., 2014) [11] and Google Speech commands dataset[13]. Though none particularly cater for Hindi dataset requirement.
YouTube videos have been successfully used to build Robust ASR systems [9], [10] though all required many layers of pre-processing in order to useful for training ASR systems. Various different heuristics have to be applied to usable be curate data to desired level in order to filter out the noisy segments which we describe in the subsequent sections.

## 3 Data collection: A rapid way

In this work, we adopted two methodologies; one for preprocessing the YouTube videos and the other to transform the same to a usable corpus for ASR model preparation. The first methodology is completely automatic process for which steps are as follows:

a. Crawl and Extract URLs from agriculture related channels on YouTube.
b. Download the videos and extract audio using ffmpeg tool.
c. Perform internal speech transcription using available ASR systems along with the word time offset boundary and then some manual correction was performed.
d. Create bucket of 15 words and mark the time range using start-time of first word and end-time of last word.
e. Usually time duration of 15 words should be between 7-10 seconds long.
f. On observation we found that longer time duration bucket segments have music, advertisement and other kinds of spoken/non-spoken noise. We filtered out those buckets having time duration more than 10 seconds.
g. Finally the audios were splitted based on the remaining bucket's time boundaries using ffmpeg tool.

The second methodology involved manual cleaning of various noises present in the audio like signature tunes, advertisements, music etc. The cleaned audios were then transcribed using available ASR systems and using word bucket algorithm, the audios were splitted.

With the use of second methodology, we could get good quality corpus but it was time consuming laborious process. Also, this methodology doesn't handles inherent Transcription and word boundary marking errors done by Speech API. First methodology is much faster but lossy as filtering process is drops word buckets. Though, it could be excellent choice for rapid development of speech corpus.

In our subsequent experiments we have used, corpus extracted using second methodology.

## 4 Methodology

Our work predominantly is focused on domain adaptation of ASR system. We begin with some details describing our existing ASR model, which is trained on 100 hours of audio collected from 1500 male and female speakers. Lexicon count in corpus is 37K unique words of general domain. And the language model has been built over the text corpus of training and test set. In this paper, we have combined two-step process to build an agriculture domain ASR model. We evaluated existing ASR model against systems built from both the experiments with respect to WER percentage.

### 4.1 ASR Model

Existing ASR described here has been built using state of the art Kaldi toolkit. Acoustic model training in Kaldi is a pipeline process. The first step consists of creating a basic HMM-GMM acoustic model (called mono) in which the HMM states

model context-independent phones. This model is used to force-align the train dataset in order to have rough estimations of the phones boundaries needed for a more robust model.
In the second step, the forced-alignments are used to train context-independent phone models, also called tri-phonemes. Just as before, this newly created HMM-GMM model (called tri1) is used to produce better forced-alignments of the training dataset to be used in training the next, more complex acoustic model.
In the following steps, several other HMM-GMM acoustic models are created iteratively by (i) applying Linear Discriminant Analysis (LDA) [9] and Maximum Likelihood Linear Transform (MLLT) [10] transforms on the input features ( tri2 model), (ii) performing speaker adaptive training ( tri3 model) and finally (iii) applying the maximum mutual information (MMI) [11] training criteria ( tri3-mmi model).
While the speech features used for training HMM-GMMs are traditional, 13-dimensional MFCCs along with their first and second order derivatives, from this step onwards, the acoustic models are trained using both MFCCs or filter-banks plus i-vectors. Moreover, for further training steps basic data augmentation is performed by applying volume and speed perturbations on the original speech signals.
The best HMM-GMM acoustic model (tri3-mmi) is further used to force-align the training dataset to produce high-quality alignments for the DNN-based models. These alignments are used to train the various deep networks that can be implemented using Kaldi NNET3 library, among which the most popular are the following: the pure TDNN [4], the factored TDNN (TDNN-F) [5], the CNN-TDNN and the TDNN-LSTM [6]. These DNN architectures can work in conjunction with the more traditional, cross-entropy objective function or with the newly introduced lattice-free MMI (LF-MMI) objective function [5]. The first one was the default in Kaldi NNET2 library, while second one is one of the innovations in Kaldi NNET3. The models using LF-MMI are also called chain models, as the objective function was inspired by the Connectionist Temporal Classification (CTC) [6], another objective function that allows training from scratch, without any pre-aligned data.

### 4.2 Domain Adaptation

In this paper, we explored a very simple technique of adapting a general domain ASR to domain specific one. Without affecting the acoustic model, Lexicon and Language model FSTs can largely help in changing the behavior of ASR on particular domain. In Kaldi pipeline, we filtered out unnecessary lexicons from the baseline general domain lexicons and enriched it with Agri domain lexes. Also the monolingual data used for Language Model was enriched with Agri domain sentences. The HCLG decoding graph is then updated with these language components and used with baseline Acoustic model. Experiment showed significant improvement in WER for utterances pertaining to Agri-domain test utterances. In latter section, WERs for the different trials have been given which indicates how quickly ASR can be effectively adapted.

### 4.3 Dataset

The first Dataset that we used to build baseline ASR has 1500 speakers' data of 100 hours with 37K unique lexicons and 84000 utterances [14].
The second dataset collected from mobile app had 236 speakers with 5900 utterances constituting approx 10 hours of audio. Total unique lexicon found was 3000 only (excluding lexicons already present in first dataset). We segregated 750 utterances as test set for evaluating our baseline model and other domain specific ASR model.
The third dataset that we collected and cleaned constituted 15 hours of clean data. We have also enriched our lexicon with additional 18K agriculture terminologies. Summary of all the three datasets is given in Table I.
The two models discussed here were built with Single GPU (NVIDIA RTX 2080), 16 core, Intel i7, 32 GB RAM system.

**Table 1.** I: Different datasets with their specifications

| Dataset | # of hrs. | # of utts. | # of Spks. | # of Lex |
|---|---|---|---|---|
| I | 100 | 84K | 1500 | 37K |
| II | 10 | 5.9K | 236 | 3K |
| III | 15 | 8.2K | NA | 5.5K |

## 5 Results &Discussion

As mentioned in previous sections, we conducted three sets of experiments with different sets of datasets and employing domain specific lexicons and Language Model. Here in, we name these experiments as Expt1, Expt2 and Expt3. Setup for these experiments are given in Table II.

**Table 2.** II: Experiments with different datasets and domain specific Lex and Language Model

| Experiment # | Setup |
|---|---|
| Expt 1 | Baseline model, 750 Testset cases |
| Expt2 | Baseline model+ enriched lexicon, 750 Test Set cases |
| Expt3 | Baseline model + Domain specific App Audio + YouTube audio + Enriched lexicon, Test Set cases. |

All these experiments and results presented here has been performed on Chained model and 750 test cases. Table III shows the performance of each experiment described in Table II.

**Table 3.** III: Word Error rates

| Experiment | WER(%) |
|---|---|
| Expt1 | 29.55 |
| Expt2 | 24.42 |
| Expt3 | 16.75 |

Expt3 outperformed Expt1 and Expt2 by huge margin recording lowest word error rate of 26.55%. As mentioned earlier, our test cases were purely from the Mobile App collected audios which almost replicate real world scenarios. These audios were unguided and without supervision. We also observed that test cases were recorded in different environments and sometimes due to distance from microphone while speaking captured very low voice also. We anticipated our baseline model (Expt1) to perform poorly. By enriching lexicons and Language model, we got improvement of ~5%. And merely adding 25 hours of domain specific audio we were able to get a performance gain of ~13%.

## 6 Conclusion

Results discussed above clearly indicates that any baseline ASR model can be adapted to domain specific ASR model by adding few more hours of audio data and enriching lexicons with domain specific terminologies. Though WER of baseline system on general domain is ~13%, we were able to achieve significant WER in Expt3 considering test case which almost is like real world case. Methods described in previous sections for data collection has proved to be very effective and rapid way of adapting existing ASR model to specific domain.

## References

1. P. Somervuo., “Experiments with linear and nonlinear feature transformations in HMM based phone recognition”, IEEE Int. Conf. on Acoustics, Speech and Signal processing,, vol. 1, pages 52–55 (2003).
2. M. J. Gales, “Maximum likelihood linear transformations for HMM based speech recognition”, Computer speech & language, 1998, pp. 75- 98 (1998).
3. D .Povey et. al., “Boosted MMI for model and feature-space discriminative training”, In ICASSP, pp. 4057-4060 (2008).
4. V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," INTERSPEECH, (2015).
5. D. Povey et. al., "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks", In INTERSPEECH, pp. 3743-3747(2018).
6. V. Peddinti et. al., "Low latency acoustic modeling using temporal convolution and LSTMs", IEEE Signal Processing Letters, 25(3), 373-377(2017).

7. D V. Peddinti et. al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI", In INTERSPEECH, pp. 2751-2755(2016).
8. A. Graves et al., “Connectionist temporal classification: labelling unsegmented se-quence data with recurrent neural networks,” In Proceedings of the 23rd international conference on Machine learning", , pp. 369–376(2006)
9. Hank Liao, Erik McDermott, and Andrew Senior. Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pages 368–373(2013).
10. Benjamin Lecouteux, Georges Linarès, and Stanislas Oger. Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. Computer Speech & Language, 26(2):67–89. (2012).
11. Anthony Rousseau, Anthony Rousseau, Paul Deléglise, and Yannick Estève. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In Proceedings 9th International Conference on Language Resources and Evaluation, pages 26–31(2014).
12. Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur.. Librispeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210. IEEE (2015).
13. Warden P. , Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, ArXiv e-prints, https://arxiv.org/abs/1804.03209. (2018)
14. J. Basu, S. Khan, R. Roy, B. Saxena, D. Ganguly, S. Arora, K. K. Arora, S. Bansal, S. S. Agrawal: Indian Languages Corpus for Speech Recognition. 22nd Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), 1–6 (2019)
15. S. Arora, K. K. Arora, M. K. Roy, S. S. Agrawal and B. Murthy, "Collaborative speech data acquisition for under resourced languages through crowd-sourcing", Procedia Computer Science, vol. 81, pp. 37-44, (2016).