Speech Data Acquisition for Voice based Agricultural Information Retrieval

Tejas Godambe and Samudravijaya K
Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai 400 005
Email: tejas.godambe@gmail.com, chief@tifr.res.in

Abstract

Speech data acquisition is the first step towards building a speech recognition system. The accuracy of recognition depends on the speech data used to train the system. This paper describes the measures taken for collecting apt speech data from Maharashtra’s farmers, to develop a robust speech recognition system as a part of voice interface for agricultural information retrieval.

Keywords: Automatic Speech Recognition, Data Acquisition System

1. Introduction

India is an agricultural country. Despite the advent of technology in recent years, agriculture has still remained the major source of income for majority of the population. In order to provide a stimulus to the agricultural growth, the Ministry of Agriculture, Government of India set up a website at http://agmarknet.nic.in/ that displays prices of commodities traded in various mandis and districts all over India. The website also archives prices of commodities traded in various mandis in India on a daily basis. This archival information is useful to farmers and traders for market analysis. The benefits of such information can be reaped if farmers are able to access this information easily. Presently, illiteracy, ignorance, lack of knowledge of English, computers and internet are acting as hurdles in accessing the information from the website. On the other hand, if this information can be made available just by speaking over a phone line, it would foster a large number of farmers. This would need implementation of a voice interface to the agmarknet website using Automatic Speech Recognition (ASR) and Text to Speech Synthesis (TTS) technologies.

With the spread of telecom technology, mobile communication has no longer remained a utility of only the rich. Availability of inexpensive mobile phones, reliable and low-tariff networks have steered the use of mobile phones even in rural areas. The notion of a voice interface stands out because of the cheap and hassle free mode of usage it aspires to provide for agricultural information retrieval to the people of India.

The Technology Development for Indian Languages (TDIL) [1] programme of the Department of Information Technology (DIT) [2] has initiated a nationwide project that effects such a voice interface in six Indian languages, which are Marathi, Hindi, Tamil, Telugu, Bangla and Assamese. This project is being implemented by a consortium of seven institutions, which are Indian Institute of Technology (IIT) Madras, IIT Bombay, IIT Kanpur, IIT Guwahati, International Institute of Information Technology (IIIT) Hyderabad, Tata Institute of Fundamental Research (TIFR) Mumbai and Centre for Development and
Advanced Computing (C-DAC) Kolkata [3]. IIT Bombay and TIFR Mumbai are jointly involved in developing the Marathi ASR system.

In this paper, we describe the process of acquiring speech data for training an ASR system for Marathi language that will form the core of a voice interface to the website providing information about agricultural commodities.

2. Voice Interface and ASR

A voice interface will enable farmers to obtain the price of commodities over phone. The three sub-systems of this system will be: i) ASR engine to recognize the uttered district, mandi and commodity name ii) a web interface to get price of the desired commodity from agmarknet website and iii) TTS system to speak out the commodity price to the farmer.

The performance of an ASR significantly depends on the quality of speech data used to train the system. By quality, we mean, how well the train data represents the actual field data. Higher the quality, the better will be the performance of the ASR system. In addition to good quality, large amount of speech data is also essential for robust estimation of the model parameters, for building statistical models of speech sounds.

Maharashtra spans a vast geographical area. It comprises of 34 districts and many dialectal regions such as Vidarbha, Marathwada, Kokan etc. The climatic conditions, and therefore the varieties of commodities grown in various districts also differ. Around 235 commodities are traded in 343 mandis in Maharashtra. Other than this, every individual speaks with different emotions and voice; recorded speech varies with different age, gender, phone handset and service provider. The background noise also differs depending upon the location of the caller. The acquired field data should contain all these speech and environmental variations, so that the system learns to handle such variations during actual use.

The recognition accuracy of ASR system can suffer due to large variations in speech signal in field. The performance will improve if the number of words to be recognized at any given point is reduced. This can be done if the district of the farmer is known because the number of commodities grown in a given district is smaller than in the entire state. So, in the current implementation, the system will seek the district name from the farmer. This will reduce the search space for ASR system. In addition, the farmer may want to know the price in neighbouring mandi. So the system will ask the name of the mandi as well.

To summarize, the system will seek 3 pieces of information: name of district, mandi and commodity name.

The following section describes the approach adopted for acquiring such varied data, and philosophy behind the approach.

3. Data Acquisition System (DAS)

The data collection task has been outsourced to a firm named Uniphore [4]. Data collection will be carried out in two phases. Speech data acquired in the first phase will be used for training the system and data acquired during the second phase will be used for field testing. Uniphore’s volunteers will travel to all 34 districts in Maharashtra and collect data from 1500 farmers i.e. 45 farmers per district in each phase. The
mode of data collection is a phone line, since farmers will use a phone line only to retrieve agricultural information from the final system. Two DASs employing two dedicated phone lines have been set up at TIFR Mumbai and IIT Bombay, where the recorded data gets stored on DAS’ Hard Disk Drives. Speech data collected over a phone line is narrowband speech. Volunteers have also been provided with lightweight voice recorders to simultaneously collect wideband speech.

3.1 Overview of data collection process

The data acquisition task has been sub-divided into two divisions: Agricultural and Sentences data acquisition. Agricultural data collection involves acquiring enough utterances of district, mandi and commodity names from the farmers. The Sentences data collection involves asking speakers to read out 10 English and/or Marathi sentences from the given set. Agricultural data collection primarily aims to collect isolated words such as district, mandi and commodity names, while Sentences data collection aims to acquire continuous speech. Both data collection techniques therefore complement each other.

Prior to speech data collection, volunteer enters the metadata details of the speaker through digit keys. The need and importance of metadata entry is described in the sub-section to follow.

3.2 Metadata Entry

Figure 1 shows the Metadata entry sheet, which lists all the metadata items and the available entry options.

Other than the acquired speech data, it is better to have knowledge of these metadata items as well. After training some basic models, one can analyze and build better metadata specific models using this information. For example, item 2 in Figure 1 gives information about the farmer’s gender. We can use this information to build gender specific models later, where the parameters of a given phone model will slightly differ for men and women. Such categorization and advanced training will reduce recognition ambiguity and help boost the accuracy.

The following sub-section introduces and discusses the Agricultural session in detail.
3.3 Agricultural Data Collection

Agricultural data collection is further sub-divided into 4 blocks, which are Directed Questions, Prompted Questions, Picture Prompted Questions and General Questions. Each block has a different motto and applies a different approach for data acquisition. It is mandatory for a farmer to cover all the four blocks. The order of selecting them is left at the farmer's discretion. After each block, call goes on Hold. It allows the farmer and volunteer to relax. The volunteer can also utilize this time to familiarize the farmer with the next block of recording. In case a farmer is unable to answer some questions, a provision for re-recording such unanswered questions is also made available. Volunteer makes a note of all these unanswered questions. Re-recording takes place immediately after completion of the block.

Now, let us have a look at each block in detail.

a) Directed Questions (DQ): Here the system directs a farmer to utter a few district, mandi and commodity names. The kind of questions asked in this field are :-
a) Tell your district name 
b) Tell the name of a nearby mandi 
c) Tell any commodity name 
We can see that questions asked here are similar to the questions that will be asked by the final system. The way in which a farmer will respond to these questions during data collection and when using final system will also be the same. At both instances, i.e. during data collection and while using the final system, farmer will first recollect answer from his memory and then respond to the questions. Thus, speech data collected in this block forms a good representative of the actual field data.

b) Prompted Questions (PQ): During Directed Questions, it is possible that a farmer may utter names of only common mandis and commonly traded commodities. For example, suppose ‘Rice’ is the most commonly traded commodity in a particular district. Let ‘Ox’ be a rarely traded commodity in that district. It is likely that most farmers will utter ‘Rice’ when asked to utter a commodity name in Directed Questions block. We will have many utterances of ‘Rice’ and very few utterances of ‘Ox’. Hence, it is difficult to collect enough utterances of rarely traded commodities through Directed Questions. Similar, argument holds true for subordinate mandis as well. Also, a farmer may not take effort to say full name of a commodity, with its specific variety. For example, (s)he may say only ‘Wheat’ in place of ‘Maharashtra 2184 Wheat’. The Prompted Questions block fills this gap by asking the farmer to repeat the district, mandi and commodity names, which thus ensures that we get enough utterances of all items irrespective of their popularity and also full names of commodities with their specific varieties.

The district, mandi and commodity names pronounced to a farmer are district dependent, i.e. system won’t ask the farmer to repeat mandi or commodity names which do not belong to his district. System comes to know about the farmer’s district when the volunteer enters District code during Metadata entry.

It is desirable that the names of popular commodities are spoken more often than that of rarely traded commodities in a district. However, a farmer is asked to repeat just 8 commodity names during the Prompted Questions block. So, a method of determining a relative frequency of prompting commodity names was used. This scheme is described below.

Consider the example of Ahmednagar district. For the past 5 years, Wheat has been traded on 5358 occasions in all mandis of Ahmednagar, while Goat has been traded just once. The difference in trading frequencies of Wheat and Goat is too large. It is understandable that we need more repetitions of Wheat to well train the models, but we also need some minimum number of repetitions of Goat so that system does not make mistake in recognizing Goat. We cannot record 5358 repetitions of Wheat for every repetition of Goat. This won’t be practical. Thus there was a need to determine the number of times a common
traded commodity like Wheat will be prompted to a farmer and number of times a rarely traded commodity like Goat will be prompted. The number of times a commodity is prompted was decided as the logarithm to the base 2 of its frequency count. Also, it will be better if every farmer is prompted a variety of commodities within 8 instances in Prompted Questions block.

![Figure 2: List of commodities traded in Ahmednagar district with descending logarithmic frequency count](image)

![Figure 3: List of 8 shuffled commodity sets prompted to farmers during PQ block.](image)

Figure 2 shows the commodities arranged in a descending order of their logarithmic frequencies. After logarithmic compression, DAS does not prompt the rare and moderately traded commodities as infrequently to a farmer as it would have prompted before compression. We still need to shuffle the list in Figure 2, so that each farmer gets to repeat a variety of commodity names in Prompted Questions block as shown in Figure 3.

Thus, the goal of collecting some minimum number of repetitions of rarely traded commodities in addition to more repetitions of commonly traded commodities is achieved after logarithmic compression and shuffling.

c) Picture Prompted Questions (PPQ): During Prompted Questions block, a pre-recorded voice speaks the name of a commodity and prompts the farmer to repeat the name. Here a farmer may try to mimic the pronunciation, speed, loudness etc. of the commodity prompts. He may not say names in his natural way. In contrast, when using the final system, a farmer will naturally utter district, mandi and commodity names. There the farmer will not have any reference prompt to imitate. Therefore, there is a need to collect spontaneous responses which probably cannot be obtained through Prompted Questions. But, Picture Prompted Questions serve to fulfill this purpose. The volunteer displays a sheet containing 10 unique pictures of commodities and asks farmer to utter the commodity names one by one. This way spontaneous response is being collected from the farmer. These commodity picture sheets are prepared district wise, i.e. a sheet for a particular district contains pictures of only those commodities which are grown/traded in that district.

PPQ gives an added advantage too. It makes easy to collect the local names of the commodities, since farmers are more likely to utter the local names of commodities rather than the formal names as found on the agmarknet website.
Figure 4 shows a commodity picture sheet for Akola district. It contains pictures of 10 different commodities. But, many districts trade more than 10 commodities. Suppose 20 commodities are traded in Akola district, then, volunteer carries 2 picture sheets for Akola. Volunteer displays sheet1 to farmer1, sheet2 to farmer2, and again sheet1 to farmer3 in a circular fashion, so that almost equal number of utterances get collected for all commodities in all sheets.

Also, wherever it was felt that a particular commodity is difficult to recognize from its picture, its name is embedded on the picture itself. For example, observe Kale Til written on the last picture in the first row of Figure 4. We can see three similar instances of this argument in Figure 4. Observe last row in Figure 4. It is a single picture which shows two different images of cotton.

![Figure 4: Commodity Picture Sheet](image)

The intention behind including such pictures containing multiple images is, to make the commodity easily recognizable to the farmer. To segregate these multi-imaged pictures from the normal ones, such pictures have been inserted at the bottom of the picture sheet.

d) General Questions (GQ): All the three blocks discussed uptil now, intended to collect isolated words i.e. district, mandi and commodity names by different means. We also wanted to collect spontaneous speech for general analysis and future work. So, in this block, the farmer is required to respond to questions related to farming. The type of questions asked here are:-
1) Do you own a cycle?
2) Tell today’s date, month and year
3) What measures do you take when pests attack crops?

Thus, having studied each block in detail, it is not difficult to understand and accept that all four blocks complement each other. Next sub-section describes the motivation behind acquiring continuous speech data through read sentences.

### 3.4 Sentences Data Acquisition

In contrast to Agricultural data collection task, where the aim is to collect isolated words, viz. district, mandi and commodity names, the primary goal behind sentences data collection is to acquire continuous speech. The speaker may or may not be a farmer.

Continuous speech is the simplest form of communication. When two people talk, the exchange of information does not take place with pause between words, rather it takes place through continuous speech. Through continuous speech one is able to efficiently convey the message with proper emotions, tone, speed, stress, etc. Even if the final system will be interested in isolated words (district, mandi and commodity names) only, farmers are likely to respond in continuous speech. For example, a farmer may say “The commodity name is Potato”, instead of saying only ‘Potato’, in response to the system’s demand
to utter commodity name. The final system should have knowledge to recognize ‘Potato’ from a sequence of words. Therefore it will be better to train the system with continuous speech too.

It is desirable to ask the farmer to read a set of phonetically rich sentences [5] if the number of such sentences is small due to small size of database. On the other hand, enough repetitions of many possible context dependent units i.e. triphones naturally occur when thousands of sentences are accumulated from different sources. So sets of sentences were constructed from a larger set collected from diverse sources without resorting to design of phonetically rich sentence sets.

Each set is prepared by pooling sentences from different sources in different proportions, to incorporate variety. For example 340 unique sets (3400 unique sentences) of Marathi sentences were prepared by pooling sentences from 5 internet sources [7][8][9][10][11], 21 books and digit sequences in different proportions. The proportion is a factor of the size of the source. A Marathi 10 sentences set was formed by collecting 6 sentences from books, 2 proverbs and 1 sentence from online stories, and 1 digit sequence. Figure 5 shows such a Marathi sentence set. A total of 80 unique digit sequences can be observed on the last lines of all 340 sets. This feature was incorporated to capture all possible ways of reading sequences of digits.

The sentences copied from web contained a lot of spelling mistakes, which had to be manually rectified. In addition, all sentences were manually scrutinized to eliminate sentences touching social and political sensitive topics. Some long sentences which could not be accommodated in a single line and repetitive sentences also had to be removed. For English sentences, foreign names, words which can be difficult to pronounce have also been removed. The font size was also adjusted large enough to aid the speaker for reading sentences.

Such large number of sentence sets compiled from different sources will contain adequate repetitions of all possible context dependent units, i.e. triphones, permitted in Marathi language which will aid to build a better general purpose as well as task specific ASR system. Preparation of 400 unique English sets from different sources [12][13] was also done in a similar way.

3.5 Current Status
The first phase of data collection has begun. Currently the data collection is progressing simultaneously in 8 districts, viz. Pune, Ahmednagar, Sangli, Satara, Raigad, Ratnagiri, Nashik and Yavatmaal. The transcription of the collected data is also advancing simultaneously. The benefits of the transliteration scheme used for transcription is presented in the accompanying paper [6].

4. Summary
In this paper, we saw the need for implementing a voice interface to aid the farmers of Maharashtra to easily retrieve the agricultural information from http://agmarknet.nic.in. The voice interface requires a robust ASR to recognize the uttered district, mandi and commodity name correctly. But robustness is highly dependent on speech data used to train the ASR. This speech data should be real data i.e. it should
comprise of all speech and environmental variations and therefore represent actual field data. Hence, different complementary approaches like Directed Questions, Prompted Questions, Picture Prompted Questions and General Questions were adopted to collect Agricultural data. Directed Questions intends to collect natural speech, Prompted Questions helps to collect minor mandi and rarely traded commodity names as well as commodity names with their specific varieties, Picture Prompted Questions aim to collect spontaneous responses and local names of commodities while General Questions help to collect spontaneous speech for general analysis and future work.

Apart from Agricultural data collection, continuous speech data collection through read sentences is also being done, where a literate speaker is asked to read 10 English and/or Marathi sentences from a given set constructed from diverse sources. This speech data will not only be useful for the current task, but also for developing other ASR systems.

Acknowledgement
We thank Prof. Preeti Rao, Pranav Jawale of IIT Bombay, Dr. Nandini Bondale, Tauseef Hussain and Nikita Acharekar from TIFR Mumbai for their useful contribution towards this work.

References