Marathi Speech Database

Samudravijaya K
Tata Institute of Fundamental Research,
1, Homi Bhabha Road,
Mumbai 400005 India
chief@tifr.res.in

Mandar R Gogate
LBHSST College
Bandra (E)
Mumbai 400051 India
mandargogate@gmail.com

Abstract

We present a progress report of the creation of speech database of Marathi language that will be used for development of automatic speech recognition system. The speech corpus will consist of spontaneous speech as well as phonetically rich sentences of Marathi, spoken by a variety of speakers over telephone channel. An account of the design of phonetically rich sentences, speech data acquisition and data validation are given. A statistical analysis of the phonetic richness of phonetically rich sets of sentences is presented.

1 Introduction

The need for human oriented computer interfaces cannot be over-emphasized. Spoken language interfaces form an integral part of such interfaces. Developments of automatic speech recognition and speech synthesis systems need spoken language databases. While speech databases are available in several western languages, these are far and few in oriental languages, barring a few exceptions. This paper describes the creation of a speech database for Marathi, an Indo-European language spoken in Western India.

The primary purpose of the Marathi speech database is to facilitate development of Automatic Speech Recognition (ASR) systems. The ASR system would recognize Marathi sentences spoken fluently by anybody, i.e., the system would be speaker independent. So, the ASR system needs to be trained with speech of lot of people so that it can handle variations in accent and speaking style. Since the accent varies from person to person, it is desirable to obtain at least one example of each phoneme (preferably in various phonetic contexts) of the language from every person. To achieve this goal, a person may have to speak lots of natural sentences. On the other hand, people are reluctant to spend significant time on contributing their voice. This difficulty becomes significant when speech has to be collected from a large number of people. Thus, there is a need to create a compact set of meaningful sentences that cover most phonemes of the language. A report of the ongoing work of creation of a Marathi database comprising of phonetically rich sentences spoken by many persons is given here. In addition, goal-oriented, spontaneous speech data is also being collected.

The rest of the paper is organized as follows. The creation of compact sets of phonetically rich Marathi sentences is described in Section 2. A statistical analysis of the text corpus is presented in Section 3. An account of the setup for acquisition of speech data over telephone channels is given in Section 4. Section 5 presents some conclusions.

2 Creation of phonetically rich sentence sets

In order to collect speech samples from a large number of people within the constraint that people are willing to spend only short amount of time, sets of sentences that are rich in phonetic context need to be created. Then, each person speaks one such compact set of sentences. This section describes the formation of thousands of such sets of Marathi sentences. Similar work for Hindi language is described in (Chourasia et al, 2005).

The required characteristics of set of sentences are as follows. The sentences should be meaningful and natural. They should be easy to read; so sentences should be simple and short. The number of sentences in a set should not be large, say not more than 10. The sentences in a set should contain all the phonemes of the language. In addition, it is desirable that the sentences should contain phonemes in various phonetic contexts. This assumes significance in the context of training ASR systems since context dependent phoneme models are generally used in ASR systems to take care of phonetic context dependent spectral variations of phonemes.

Although sets of phonetically rich sentences can be generated manually, it is a laborious process. On the other hand, such sets satisfying the required characteristics can be created by a program if a large corpus of electronic text is available in form that is amenable to automatic processing and statistical analysis. A source of such large electronic text corpus is an electronic newspaper in Marathi that lets public to have access to its archives.

Marathi language uses Devanagari, a character based script. A character represents one vowel and zero or more consonants. Consonant clusters are represented by combination of ligatures; so, there are hundreds of characters. Many characters share glyphs. To take advantage of such shared glyphs, some true type font designers have used codes corresponding to glyphs rather than phonemes. This makes the task of rendering Marathi text in Devanagari script easier. However, this causes problems for us, since we are interested in statistical analysis of phonemes. The grapheme-to-phoneme mapping of text in true type fonts is not straightforward, especially because of non-linear nature of Devanagari script. For example, the word "ki" is a sequence two phonemes: /k/ and /i/. However, the corresponding character in Devanagari script has the glyph corresponding to

/i/ *preceding* the consonant. Due to such non-causal nature, and sharing of glyphs among characters, the grapheme-to-phoneme mapping is complex and was to be discovered. A program was written that yields a sequence of phonemes from online Marathi text coded for true type fonts. The online sentences, that were often long, were broken into phrases, and were manually validated for syntactic and semantic correctness. From the archives of a few years, we obtained 125,000 short (4 <= number_of_words <= 10) sentences by this procedure. The transliterated text in ASCII was processed further to derive sentence sets.

CorpusCrt, a public domain sentence selection software (CorpusCrt) was used to derive 1,000 sets of 10 Marathi sentences each from the text corpus of 125,000 sentences. The program strives to increase the richness of linguistic units in the sentence sets. We specified phoneme as the basic linguistic unit. An analysis of the phonetic richness of these sentence sets is presented in the next section.

3 Statistical analysis of sentences

One thousand sets of phonetically rich sentences were automatically created as described in the previous section. In this section, we carry out an analysis of the phonetic characteristics of these sentences in order to get a feel for the effectiveness of the sentence selection procedure.

One of the pre-requisites of sentence selection was that each set of 10 sentences (to be read by one person) should contain at least one example of each phoneme of the language. The sentence selection program, CorpusCrt, picked those sentences which contain phonemes that occur rarely in natural text, thus enhancing the phonetic richness of each sentence set. Of course, the success of this program depends on the phonetic richness of the input text. Nevertheless, we can say that the sentence sets are phonetically rich if most of them contain at least one example of each phoneme.

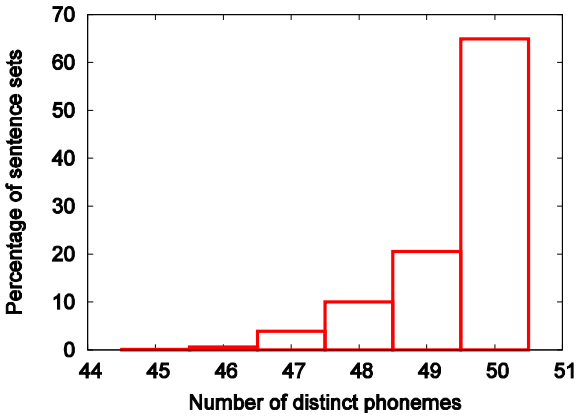

Figure 1: Distribution of sentence sets as a function of the number of distinct phonemes they contain. About two-thirds of the sentence sets contain all the phonemes of the language. About 22% sentence sets do not contain one of the 50 Marathi phonemes.

Figure 1 shows the percentage of sentence sets as a function of the number of *distinct* phonemes they contain. There are 50 Marathi phonemes. The mode of the graph is at 50 phonemes. About 2/3rd of the sentence sets contain at least one token of each and every Marathi phoneme. Even the least phonetically rich sentence set contains at least 45 phonemes out of 50. It would be interesting to discuss about the phonemes that did not occur in some sentence sets.

Information about top 10 rare phonemes is given as a bar graph in Figure 2. The graph shows the percentage of sentence sets that contain these rare phonemes. Out of 10 rare phonemes, 6 are asphirated stop consonants; 2 are retroflex sounds; the rest are velar and palatal nasals. Such a tilt of the distribution of rare phonemes towards asphirated stops is to be expected since asphirated phonemes are known to occur rarely in natural text [see figure 4 of (Arora et al 2003)]. There are 10 asphirated stop consonants in Marathi. So, it is difficult to have all of the 10 rare phonemes in addition to other phonemes in just 10 sentences. The most rare phoneme is the velar nasal (/ng/); it does not occur in 64 out of 1000 sentence sets.

It is interesting to ask whether the CorpusCrt, the sentence selection program, has done a good job. The answer will be in the affirmative if the phonetic richness of the compact set of 10,000 sentences is better than that of the general text

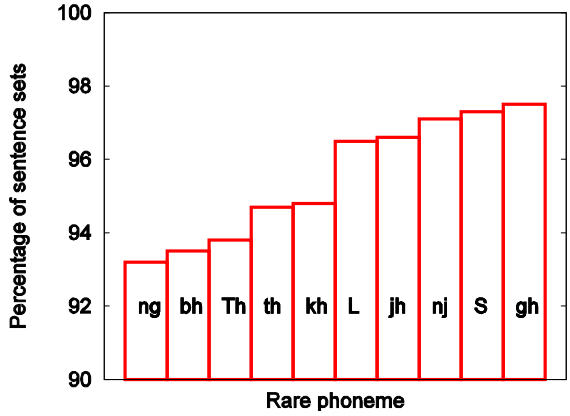

Figure 2: Percentage of sentence sets in which the top 10 rare phonemes are present. The velar nasal /ng/ is the rarest phoneme; it occurs in only 93.6% of the sentence sets.

corpus of 125,000 sentences. The `phonetic richness' of a text corpus is high if the relative frequencies of rare phonemes are high. Relative frequency of a phoneme is the ratio of the number of times the phoneme occurred in a set to that of all the phonemes in the database. The phonemes were arranged according to increasing relative frequency and cumulative relative frequencies were computed. Figure 3 shows cumulative relative frequency curve of the general text corpus and the phonetically rich sentence corpus (y-axis on a log scale). It is clear that the cumulative relative frequency curve of the phonetically rich compact set (filled blue circles) rises faster than that of the general text corpus (red stars). For example, in the general text corpus of 125,000 sentences, the rarest phoneme was unvoiced, aspirated affricate (/ch/) whose relative frequency was 0.03%. The relative frequency increased to 0.195% in the phonetically rich compact set, corresponding to an increase by a factor of 6.5. The CorpusCrt program has done such a good job that the phoneme /ch/, which was the rarest phoneme in the general text corpus, is no longer a rare phoneme in the selected sentence sets; this can be seen from Figure 2.

4 Speech data acquisition setup

The Unique Selling Point of voice interfaces is the ability to access an information system remotely using. The ubiquitous telecommunication system enables one to access information on an 'anytime anywhere' basis if the information server is speech enabled. For this reason, we want to build speech database that will aid development of Marathi speech recognition system for telephony speech. This section gives a brief account of the data setup; see (Samudravijaya 2006) for details.

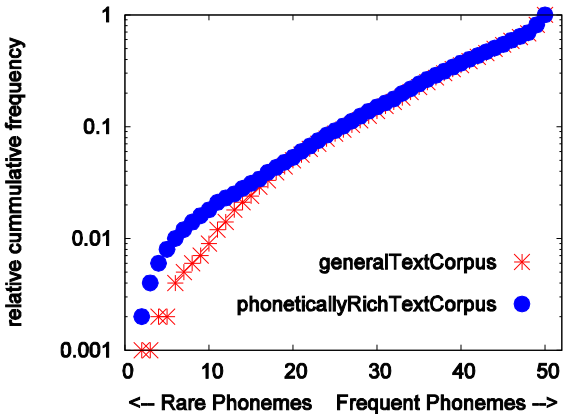

Figure 3: The graph shows the cumulative relative frequency curves of the general text corpus (red starts) and the phonetically rich sentence corpus (filled blue circles). Clearly, the relative frequencies of the rare phonemes have increased in the phonetically rich text corpus as indicated by the faster rise of the curve of the phonetically rich compact set than that of the general text corpus.

A Computer Telephony Interface (CTI) card was used to setup a data collection environment. Marathi speakers are asked to read one set of 10 phonetically rich sentences printed in Devanagari script on a sheet. The data acquisition setup not only records the speech data, but also permits re-recording specified sentence, if needed. In addition to narrowband speech recorded by the CTI card, we also collect wideband speech by recording the same speech by a lightweight digital recorder cum flash drive that is temporarily attached to the telephone instrument of the caller. We also plan to collect spontaneous, conversational speech in a Wizard of Oz mode. Here, the caller is required to interact, in voice only mode, with a machine in order to get specific information such as availability of reservation in a particular train by a particular class on a pre-specified date (Samudravijaya 2006). We have thoroughly tested the speech data collection setup and have collected speech data from a few dozen persons. The goal is to collect speech samples from hundreds of people of different age groups who speak various dialects of Marathi.

The collected data needs to be validated and transcribed. A pronunciation dictionary has to be generated. While this task is easy thanks to near one-to-one correspondence between Devanagari graphemes and Marathi phonemes, there are quite a few pronunciation rules, some of which can be automated. The rest have to be manually handled because they call for morphological analysis of words. The task of augmenting/editing the pronunciation dictionary is likely to be tougher in case of spontaneous speech.

5 Conclusions

We presented an account of the creation of Marathi language speech database. Through a statistical analysis, we showed that the sentences to be read by speakers are phonetically rich. Hence, the corresponding speech data should be valuable for training speaker independent, continuous speech recognition systems for Marathi language.

6 Acknowledgements

We thank Poonam, Swankita, Jessy Verghese, Lokhandwala Tayabi, Sampat Desai, Shraddha Dalvi, Kashyap Patel and Shailesh Khole for their help in accumulation of the text corpus of Marathi.

References

Arora Karunesh Kr., Sunita Arora, Vijay Gugnani, V N Shukla and S S Agrawal, 2003. GyanNidhi: A Parallel Corpus for Indian Languages including Nepali. Presented at Information Technology: Challenges & Prospects (ITPC – 2003), May 23-26, Kathmandu, Nepal; <http://www.cdacnoida.in/technicalpapers/ PaperNepal.pdf>

CorpusCrt: <http://gps.tcs.upc.es/veu/personal/ sesma/sesma/CorpusCrt/php3>

Chourasia V., Samudravijaya K., Chandwani M. 2005 "Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database", Proc. O-COCOSDA 2005, Indonesia, pp. 132--137.

Samudravijaya K, 2006. Development of Multi-lingual Spoken Corpora of Indian Languages. Proc. of Int. Symp. on Chinese Spoken Language Processing (ISCSLP 2006), Lecture Notes on Artificial Intelligence, LNAI 4274, pp. 792-801. Springer-Verlag, Berlin.