Spoken digits in Hindi and Indian English

Basabdatta Sen Bhattacharya (1), Aiswarya Subramanian (2), Purbayan Chatterjee (1) and Sounak Dey (3)

(1) Department of Computer Science and Information Systems, Birla Institute of Technology and Science (BITS) Pilani, Goa Campus, Goa, India

(2) MathWorks, Bangalore, India

(3) TCS Research & Innovation, Tata Consultancy Services, Kolkata, India

Abstract

The data consists of spoken digits from one to ten in Hindi and English with regional accents from across India. There is a distinct lack of availability of speech data set in Hindi, which is spoken in large parts of India. On the other hand, English is the most commonly used language in work places across urban India in current times. The data sets of spoken English available currently as open access or otherwise are mainly by native English speakers. Our work is an initial attempt to collate spoken data in Hindi and Indian English. The nomenclature of each data file contains information about the specific spoken digit, language, and the gender of the speaker. To distinguish the data files while maintaining speaker anonymity, each data file is tagged with a numerical index. In addition, there are a set of data files that do not contain gender information and may be treated as unlabelled data. Both the labelled and unlabelled data sets can be used for supervised and unsupervised learning applications respectively, particularly in the area of Speech Recognition and Spoken Digit Classification.

Introduction

There are several data set that are available freely over the internet for training and testing neural networks for speech recognition and spoken digit classification [1]. In Neural Networks and Machine Learning pedagogy, data sets are often available over Kaggle for project based training [2]. Thus, students are unaware of data collection procedures and the associated standard practices in the area. This work was carried out by the students enrolled in the Neural Networks course, BITS F312, Semester 1, Academic Year 2020-21, Birla Institute of Technology and Science (BITS) Pilani, Goa Campus. The primary objective has been to provide a holistic experience to the students where, based on their review of existing literature [3, 4], they have collected data of spoken digits from one to ten in Hindi, and English as spoken in India. All participants are native of India and were residing in the country at the time of data collection.

India is a land of many languages that are distinct in both spoken and written forms. Thus, there is a need for a common language(s) for inter-regional (also referred to as ‘inter-state’) communication. The two most widely used languages for inter-state communication in India are Hindi and English. Besides, large parts of India speak Hindi for communication in day-to-day activities, and their respective native languages are either pure or some variants of Hindi, all of which are believed to have originated from Sanskrit. Also, it is the most commonly spoken language for communication in the Indian hinterlands. At the same time, there are large parts of India where the local languages are not related to Hindi. In these parts, English is the preferred mode for inter-regional communication (i.e. where local language cannot be used) for informal as well as formal purposes. Moreover, the literate urban Indian in current times often prefer English for inter-regional communication. Also, English is a consistent feature and/or

component in all official communications across India. However, the English spoken in different parts of the country are accentuated with regional tones, while the sentence structures are influenced by the nuances of the local language. Thus we term the spoken English in India as ‘Indian English’ (the interested may refer to the Wikipedia page for an informal read). Our data set contains accents from most states of India, each of which have different regional language and accent. Below we provide a brief description of our data collection and processing methods.

Methodology

The speech data set in both Hindi and Indian English are collected by eight cohorts, each cohort consisting of four to five members working in a team. As each cohort worked independently, the data collection and processing techniques vary. Most of the data collected are either in person on a recorder app of a mobile handset, or via one-to-one online communication over social apps. In addition, some cohorts obtained audio clips from freely accessible content on social media sites. All cohorts attempted to capture the variability in regional accents across India, in both Hindi and English. The nature of the source data can be categorised as “field recordings,microphone conversation and web collection” [5].

The raw audio files have spoken digits recorded consecutively as a single audio file by each speaker. Each of these raw audio files is then cut into chunks, where each chunk is an audio file containing one spoken digit by one speaker. These single-spoken digit data files are then pre-processed using the *Audacity* software tool, as well as with bespoke scripts using Python libraries for example *librosa, noisereduce*. Unwanted pauses are detected using the python library *pydub.* In addition, shifting and stretching for both higher and lower frequencies are applied. Finally, the files are padded by silence bits of required duration to ensure that every released audio file is of length 1.2 seconds. This uniformity in length will help in training and testing on the dataset when used for machine learning applications. Background noise are mostly retained, although some of the data are either recorded especially in a noise-free environment, or cleaned after recording to avoid abrupt noises such as car horns. We believe that these variability in the recording environment will in turn provide variability to the data set, which is desirable for building robust and reliable Speech Recognition applications.

The spoken data for each digit is saved in three folders viz. 'Female' (spoken by females), 'Male' (spoken by males) and 'Unspecified' (contains digits spoken by both females and males). Each data file is named as follows :

- {id}_{digit}_{language}_{gender},
- {id}_{digit}_{language}_{unspecified},

where ‘gender’ is specified as ‘male’ or ‘female’; where gender data is not specified, the data is tagged with ‘unspecified’; ‘language’ is specified as ‘Hindi’ or ‘IndianEnglish’; ‘digit’ is specified in words in English eg. ‘one’, ‘two’, for both languages; ‘id’ refers to a numerical index that is provided to distinguish each data file, i.e. we do not provide speaker name to maintain anonymity. The data set with gender unspecified may be treated as unlabelled data for machine learning applications that require gender specification. Details of the data files are provided in Tables 1 (Hindi) and 2 (Indian English) below. Note that some data files that were identified as corrupted during testing are removed from the dataset. Thus, there may be gap in the indices. This is outlined for each digit subdirectory in the Tables 1 and 2. Table 3 lists the characteristics of the audio data.

Table 1: Overview of data files and the data sets that for the Hindi spoken digits dataset

Label	**Name of data file/data set**	**File types** (file extension)
1_Hindi --- Female --- Male --- Unspecified	<1-34>one_hindi_female.flac <1-50>one_hindi_male.flac <1-116>one_hindi_unspecified.flac	flac
2_Hindi --- Female --- Male --- Unspecified	<1-34>two_hindi_female.flac <1-50>two_hindi_male.flac <1-102>two_hindi_unspecified.flac	flac
3_Hindi --- Female --- Male --- Unspecified	<1-34>three_hindi_female.flac <1-50>three_hindi_male.flac <1-103>three_hindi_unspecified.flac	flac
4_Hindi --- Female --- Male --- Unspecified	<1-34>four_hindi_female.flac <1-34, 36-51>four_hindi_male.flac <1-4, 6-14, 16-65>four_hindi_unspecified.flac	flac
5_Hindi --- Female --- Male --- Unspecified	<1-34>five_hindi_female.flac <1-50>five_hindi_male.flac <1-86, 88-114>five_hindi_unspecified.flac	flac
6_Hindi --- Female --- Male --- Unspecified	<1-13>six_hindi_female.flac <1-28>six_hindi_male.flac <1-88>six_hindi_unspecified.flac	flac
7_Hindi --- Female --- Male --- Unspecified	<1-34>seven_hindi_female.flac <1-50>seven_hindi_male.flac <1-102>seven_hindi_unspecified.flac	flac
8_Hindi --- Female --- Male --- Unspecified	<1-34>eight_hindi_female.flac <1-50>eight_hindi_male.flac <1-99>eight_hindi_unspecified.flac	flac
9_Hindi --- Female --- Male --- Unspecified	<1-34>nine_hindi_female.flac <1-50>nine_hindi_male.flac <1-103>nine_hindi_unspecified.flac	flac

10_Hindi --- Female --- Male --- Unspecified	<1-34>ten_hindi_female.flac <1-50>ten_hindi_male.flac <1-105>ten_hindi_unspecified.flac	flac

Table 2: Overview of data files and the data sets that for the Indian English spoken digits dataset

Label	**Name of data file/data set**	**File types** (file extension)
1_IndianEnglish --- Female --- Male --- Unspecified	<1-32>one_indianenglish_female.flac <1-53>one_indianenglish_male.flac <1-112>one_indianenglish_unspecified.flac	flac
2_IndianEnglish --- Female --- Male --- Unspecified	<1-32>two_indianenglish_female.flac <1-53>two_indianenglish_male.flac <1-108>two_indianenglish_unspecified.flac	flac
3_IndianEnglish --- Female --- Male --- Unspecified	<1-32>three_indianenglish_female.flac <1-53>three_indianenglish_male.flac <1-108>three_indianenglish_unspecified.flac	flac
4_IndianEnglish --- Female --- Male --- Unspecified	<1-32>four_indianenglish_female.flac <1-53>four_indianenglish_male.flac <1-114>four_indianenglish_unspecified.flac	flac
5_IndianEnglish --- Female --- Male --- Unspecified	<1-4, 6-33>five_indianenglish_female.flac <1-53>five_indianenglish_male.flac <1-92, 94 -118>five_indianenglish_unspecified.flac	flac
6_IndianEnglish --- Female --- Male --- Unspecified	<1-13>six_indianenglish_female.flac <1-28>six_indianenglish_male.flac <1-91>six_indianenglish_unspecified.flac	flac

7_IndianEnglish --- Female --- Male --- Unspecified	<1-32>seven_indianenglish_female.flac <1-53>seven_indianenglish_male.flac <1-105>seven_indianenglish_unspecified.flac	flac
8_IndianEnglish --- Female --- Male --- Unspecified	<1-32>eight_indianenglish_female.flac <1-53>eight_indianenglish_male.flac <1-106>eight_indianenglish_unspecified.flac	flac
9_IndianEnglish --- Female --- Male --- Unspecified	<1-32>nine_indianenglish_female.flac <1-53>nine_indianenglish_male.flac <1-105>nine_indianenglish_unspecified.flac	flac
10_IndianEnglish --- Female --- Male --- Unspecified	<1-32>ten_indianenglish_female.flac <1-53>ten_indianenglish_male.flac <1-106>ten_indianenglish_unspecified.flac	flac

Table 3: Overview of the audio characteristic values.

Audio Characteristic	Characteristic Value
Audio type	FLAC
Bit depth	16 bit
Sample rate	16 kHz
Number of channels	Single
Sample Format	Linear PCM

Applications of the Data

This data set is intended primarily for Speech Recognition and Spoken Digit Classification applications. Considering that the data set consists of simple spoken digits, it can be used by students as well as teachers for learning and pedagogy respectively. At the same time, advanced applications that intend to include languages spoken in India may use this for preliminary training and testing purposes. Also, machine learning enthusiasts can use the data to develop fun projects.

The data set is labelled in terms of gender, digit, and language, and can be used with supervised learning algorithms using any of the three attributes. Furthermore, a part of the data set where gender is not specified can be treated as unlabelled data and can be

used in unsupervised learning applications where gender information is the key, for example Gender specification. On the other hand, for Language identification applications, all files in the data set can be treated as labelled data. Overall, the data set can be used for research as well as for projects under the broad area of Natural Language Processing.

For convenience of the users, we have provided a Google Colab Notebook (SpeechRecognitionDataAPI.ipynb) which can be used for basic functionalities such as removing noise, removing unwanted spaces in between the audio etc. Some of the libraries that we have used are specified here for interested users:

- **scipy.signal.spectrogram –** To get the spectrogram
- **noisereduce.reduce_noise –** To remove unwanted noise
- **pydub –** To remove unwanted silence in audio

Contributors to the Corpora

The authors acknowledge the contribution of the Academic cohort of Semester I, Academic Year 2020-21, at Birla Institute of Technology and Science (BITS) Pilani, Goa Campus, who registered for the Neural Networks (BITS F312) course. BSB has been the Instructor-in-Charge for the course; PC has been one of the Teaching Assistants who handled this exercise. The list of all registered students who are contributors to this work are as follows:

Pulkit Purwar, Shashank Mittal, Chetan Gupta, Savio Jomton, Aditya Jain, Rochishnu Banerjee, Avil Aneja, Ashrut Kumar, Rijul Ganguly, MS Nandan, Vanishka Kapoor, Sanjay Krishnan, Surya Vatsalya, Sri Hari Chidella, Aman Vaya, Rajat Goyal, S Sai Vineet, Yash Narang, Vedant Chavda, Moitrish Majumdar, Akhil Tarikere, Rajath Reghunath, Abichal Ghosh, Ninad Sunil Mhalgi, Saurabh Wandhekar, Satyansh Singh, Abhinav Shankar, Prateek Goyal, Karan Shetty, Mayank Sheoran, Shivam Chandak, Shubh Shah, Soumik Dhua, Susmit Wani, Nilesh Kumar Gupta, Vinayak Shukla, Avdhoot Bhandare, Amit Chauhan, Siddharth Sharma, Atishay Jain, Ram Kishan Vooka, Vidit Lohia, Piyush Ranjan Deb, Subhjeet Pati

Author contributions

BSB and SD conceived of the data collection plan. AS handled all post-processing, organising, cleaning and testing of the data files, as well as final editing and handling of the document files. All authors contributed to the writing of this document and supporting materials.

Funding

Basabdatta Sen Bhattacharya is supported by Insitutional Grant no. BPGC/RIG/2018-19/ , and Institutional Grant no. GOA/ACG/ 2019-20/ Oct/02 , of the Birla Institute of Technology and Science (BITS) Pilani, Goa Campus, Goa, India.

References

[1] Free Spoken Digit Dataset https://github.com/Jakobovski/free-spoken-digit-dataset

[2] Free Spoken Digit Dataset in Kaggle https://www.kaggle.com/joserzapata/free-spoken-digit-dataset-fsdd

[3] Andrew R. Freed. “Data Collection and Training for Speech Projects”. https://medium.com/ibm-watson/data-collection-and-training-for-speech-projects-22004c3e84fb

[4] Marco Noel. “Watson Speech-To-Text: How to Train Your Own Speech “Dragon” — Part 1: Data Collection and Preparation”.

[5] LDC UPenn catalog https://catalog.ldc.upenn.edu/search