DEVELOPMENT OF PHONEME DOMINATED DATABASE FOR LIMITED DOMAIN T-T-S IN HINDI

Archana Balyan

Department of Electronics and Communication Engineering, Maharaja Surajmal Institute of Technology, Affiliated to GGSIPU, New Delhi, India

ABSTRACT

Maximum digital information is available to fewer people who can read or understand a particular language. The corpus is the basis for developing speech synthesis and recognition systems. In India, almost all speech research and development affiliations are developing their own speech corpora for Hindi language, which is the first language for more than 200 million people. The primary goal of this paper is to review the speech corpus created by various institutes and organizations so that the scientists and language technologists can recognize the crucial role of corpus development in the field of building ASR and TTS systems. This aim is to bring together all the information related to the recording, volume and quality of speech data in speech corpus to facilitate the work of researchers in the field of speech recognition and synthesis. This paper describes development of medium size database for Metro rail passenger information systems using HMM based technique in our organization for above application. Phoneme is chosen as basic speech unit of the database. The result shows that a medium size database consisting of 630 utterances with 12,614 words, 11572 tokens of phonemes covering 38 phonemes are generated in our database and it cover maximum possible phonetic context.

KEYWORDS

Speech database, Phonemes, Metro Rail Passenger Information System, Hidden Markov Model, Text- to-Speech

1. INTRODUCTION

The objective of speech data collection is to primarily build speech recognition system (ASR), Text – to- speech synthesis systems (TTS) and speech translation system for various Indian languages. As there is ever growing demand for customized and domain specific voices for use in corpus based synthesis systems, it is essential that standard methods should be established for creating these databases as it directly affects the quality and accuracy of the system.

In a country like India, with over 22 officially recognized languages, the language barriers can impose hurdles in the development of related fundamental Information communication technologies (ICT). In such multilingual environment, if speech technology is integrated with cross- language speech to speech translation systems, it could be of great help to remove this blockade so that services and information can be provided across languages more easily. This shall be of great cultural and economic value for a diverse country like India. This basic infrastructure will speed up the creation of large size spoken language corpora and subsequently lead to the development multilingual technologies such as multi-lingual speech translation, multi-lingual speech transcription, and multi-lingual information retrieval. In the field of development of ASR, TTS and speech translation systems, advances have been largely promoted by the Ministry of the Information and Communication Technologies (MCIT), Govt. of India (GoI) via selected set of Indian academic and research institutions in consortium mode. The resources

DOI : 10.5121/ijaia.2017.8301

including speech and text corpora collected in these efforts abide by the copyright restrictions of the sponsor [1].

Hindi is the National language of India and is spoken by majority of population (about 41%) residing mostly in northern, central, western and eastern regions of the country [2]. Hindi is written in the Devanagari script which contains 13 vowels, 33 consonants and 3 consonant clusters with special symbols [3]. In section 2, the paper focuses on the Hindi resources being developed, which can be used for research in computational linguistics. In section 3, the methodology adopted for creating our own speech database (phoneme) is presented. In section 4, the challenges faced during database creations are discussed. In section 5, conclusions are presented.

2. DEVELOPMENT OF SPEECH CORPORA: A SURVEY

In modern linguistics, a corpus is referred to as a large collection of structured text in written or spoken form to be used mainly for linguistic research [4]. If imperative linguistic information such as tags, labels etc is added to the text corpus, it is called annotated corpus. Development of TTS requires formalized knowledge for speech generation on all linguistic levels (morphological structure, the spelling variations and word sense analysis), it can be the best resource to acquire various language phenomenon. A speech database for automatic speech recognition system for travel domain has been developed at C-DAC, Noida. The training data consists of recording by 30 female speakers in the age group of 17-60 years. The recording environment was noise-free and echo cancelled. The duration of recorded sentences is approximately 26 hrs and consists of 8,567 sentences comprising of 74,807 words [6].

At IIT, Kharagpur, Hindi corpora have been developed from Broadcasted news bulletin . The total duration of speech corpus is 3.5 Hrs. The recording was done in the studio in a Noise free environment for 19 speakers (6 Male and 13 Female) [5].

The IIIT Hyderabad has collected speech databases for usage in speech synthesis systems for multiple Indian languages. The Wikipedia text corpus for Hindi language consists of 44100 sentences, 1361878 words, 942079 syllables and 1466610 phones. The optimal text selection comprises of 1000 sentences, 8273 words, 19771 syllables and 30723 phones. The duration of the recorded speech corpus is 1.12 hours with average duration of sentence being 4.35 sec. The speech data was recorded in noise free studio environment using a laptop and a headset microphone connected to a zoom handy recorder [6].

At TIFR Mumbai and CDAC Noida , a general purpose database has been created for a set of 10 phonetically rich sentences. The signals were recorded directly in the digital format using two microphones in a noise free environment. Each sentence consists of two parts; the first part consists of two sentences which cover the maximum phonemes occurring in Hindi and the second part consists of eight sentences to cover the maximum possible phonetic context [8]. ILCI (Indian Languages Corpora Initiative) started by Technology Development for Indian Languages (TDIL) for building parallel corpora for major Indian languages including English. The Central Institute of Indian Languages (CIIL) also has been its resources towards resource building; the focus being on development of raw corpora for the language community rather than on the annotated corpora.

2.1. DEVELOPMENT OF CORPORA SPECIFIC TO MOBILE COMMUNICATION ENVIRONMENT

KIIT college of Engineering has collected a text corpus of 2 million words of raw messages in 12 different domains. A speech corpus of duration of approx. 4 hours has been created by digitally recording 630 phonetically rich sentences by 100 speakers, where each speaker spoke 630

sentences. The recording was done through 3 channels simultaneously: a mobile phone, a headset and a desktop mounted microphone. The annotated data sets are available for various applications such as development of language models, language features, and language translation etc in field of mobile communication. This project was undertaken in collaboration by Nokia Research Centre China [6].At KIIT Bhubaneswar, a total of 600 phonetically rich sentences, consisting of 42,801 unique words was created for Hindi language after collecting text messages in Hindi. The sentences were recorded by 100 speakers. The digital recording has been performed through 3 channels simultaneously. In the developed database, the ratio of female voice recording to male voice recording is 6:4 [9].

At the Linguistic Data Consortium for Indian Languages LDCIL [1], a database was collected in two states ; Uttar Pradesh and Bihar . In collecting the database, 650 different native speakers in were recorded distributed across three age groups (16-20, 21-50 , 51 and over). The recording of news text corpus was conducted in noisy environment at home, office and public places through recorder with an inbuilt microphone by each speaker. The speech signal files from all speakers were transcribed and labeled at the sentence level [10].

3. CREATING SPEECH DATABASE FOR METRO RAIL PASSENGER INFORMATION SYSTEM (MRPIS)

Delhi Metro Rail Corporation (DMRC) was established by the Government of India (GoI) and the Government of Delhi in March 1995 to build a new metro system in Delhi, capital city of India. For communication within metro, Intercoms are provided for emergency communication between the passengers and the driver in each coach, and on-train announcements are in Hindi and English. Under circumstances such as such as technical breakdown, accidents, change of route or any unforeseen incidents etc there is need for special announcements other than regular ones. It requires certain specific approach to handle the communication between the Metro Rail operators and passengers to address this kind of emergency situation. The use of TTS shall improve the quality of service to the passengers on board. Development of natural sounding corpus-based TTS system for Delhi Metro requires database with large number of units having adequate variations in prosody and spectral characteristics created exclusively for metro rail passenger information system.

The major steps taken to build medium size database for use in development of TTS for metro rail passenger information system (MRPIS) are as given below.

1. Selection of text sentences as appropriate to selected domain
2. Recording of selected text corpus
3. Automatic segmentation of speech corpus at phonetic level

3.1 TEXT CORPUS: SELECTION OF SENTENCES

The domain specific text corpora was so collected that these sentences not only covers the most frequent words that are announced while travelling in metro rail but also to have an optimal text corpus balanced in terms of phonetic coverage and have sufficient coverage of prosodic contexts of each type of sound to capture the acoustics of Hindi speech. Phonetic rich sentences are needed for robust estimation of the parameters in building statistical models of context Independent HMM (CIHMM) for phonemes. A sentence set is considered to be phonetically rich if it contains most, if not all, phonemes of the Hindi language.

[1] http://www.ldcil.org/resourcesSpeechCorpHindi.aspx

In order to create a medium size phonetically rich speech database, 630 sentences were collected. There are 38 distinct phonemes in the corpus. It is seen that each sentence contain minimum of 11 phonemes and 2% of the sentences contain all the 38 phonemes. The percentages of sentence sets containing at least 30 and 34 distinct phonemes are 24% and 10% respectively.

3.2 RECORDING OF SENTENCES AND DIGITIZATION

The quality of digital sound is determined by discrete parameters such as the sample rate, bit capacity and number of channels [11]. One can be conclusive that our digitization standards should be able to faithfully represent acoustic signals. The LDC-IL under CIIL has designed the standards for acquisition of the speech data according to application it is intended for and the recording devices that were used for recording the speech samples [12].

The corpus consists of 630 phonetically- balanced Hindi sentences spoken by single male speaker. The digitized recording was done at sampling rate of 16 kHz , stored in 16 bit PCM-encoded waveform format in mono mode. The duration of recording is about 1.5 hours. The speech utterances are manually transcribed into text using INSROT[2] , the Indian Script Roman Transliteration scheme for Hindi. The Table 1 below shows the statistical analysis of the recorded speech corpus.

Table 1. Statistics of speech Corpus

Sl. no	Process	Count
1	Total number of utterances recorded	630
2	Total number of words	12,614
3	Total number of unique phonemes	38
4	Total number of phonemes recorded	11572
5	Total number of nouns recorded	78

3.3 AUTOMATIC SEGMENTATION OF SENTENCES INTO PHONEMES

The basic and essential task in building speech database is the speech data segmentation and labelling. In this work, phoneme is chosen as the basic speech unit for segmentation. The accuracy of the segmentation and labelling of phoneme directly affects the quality of TTS. Manual segmentation is labour intensive, tedious, time consuming, error prone, inconsistent and requires much effort. In order to reduce manual efforts and speed up segmentation process, automatic phonetic segmentation is preferred [13].

3.3.1 PRE-PROCESSING AND FEATURE EXTRACTION PROCESS

The recorded speech waveform is digitized at a sampling rate of 16 KHz and is pre-emphasized (α=0.99). The speech signal is uniformly segmented at interval of 25 msec (overlapping factor = 0.5). Thus, each frame consists of 400 samples. Hamming window is then applied to each frame and then FFT is performed on each frame. For each frame, 12 mel-frequency cepstral coefficients (MFCC)[14] are extracted including 12 delta Δ and 12 Δ - Δ coefficients with cepstral mean normalization (CMN) . Hence, a 36 dimension feature vector is extracted for each frame that is used as one of the inputs for building Phone models.

[2] http://www.tdil.mit.gov.in

3.3.2 HMM BASED PHONETIC SEGMENTATION

This is the process of building a model for each phoneme in the spoken speech. The work presented in this paper is based on Hidden Markov Models [15] wherein forced alignment using Viterbi algorithm [16] is applied to find the most probable boundaries for the known sequence of phonemes.

3.3.3 BUILDING HMM MODELS FOR PHONEMES

The phone models of each phoneme involves creation of HMM file, finding and storing global mean and variance for each phoneme occurring in database. The block diagram for training with HMM is given in Figure 1 below.

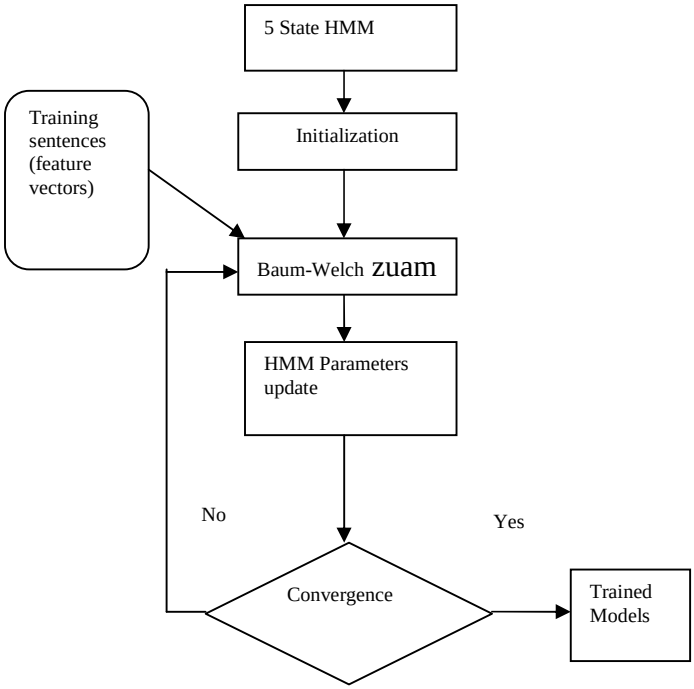

Figure 1: Block diagram for HMM Training

HMM dealing with continuous speech signals are characterized by a Gaussian pdf. In our experiment, Context Independent Phoneme HMM (CI Model) consists of M state (M=5) , for each phoneme unit and M state (M=3)for the non-speech region, i.e., the silence since silence is stationary and has no temporal structure to exploit. The phone models created so far are used to realign the training data (MFCC frames) and create new transcriptions. For alignment of frames with phones, the Viterbi re-estimation algorithm has been implemented , that starts with uniform segmentation (flat mode) and rapidly converges to the best model estimates in just few iterations and obtain optimized speaker-dependent HMM for each phoneme. Table 2 gives the parameters used for building HMM models for each phoneme.

Table 2. Parameters for Building Phone Model

Parameters	**Number**
Frame size	25msec
Feature vector dimension	36
No. of states to model each phoneme	5 for phonemes and 3 for silence
Topology	Left – to-right with no skip states
Number of Gaussian mixtures	05
No. of phone models	183
Convergence criterion	error less than 0.04
Training data	150-200 sentence
Test data	80-100 sentence
Platform	visual studio on Windows

3.4 DESIGN OF SPEECH DATABASE

The phonetic level segmentation for each recorded speech utterance is performed using HMM. In the database generated, each phoneme is tagged with its left and right context, information about start time and end time, duration of the phoneme and the location of the wave file to which it belongs. The speech database is prepared and stored in the format given below as shown in Figure 2.

{Prefix phoneme; phoneme; suffix phoneme}{Start time}{End time}{Duration of phoneme}{Wave file}

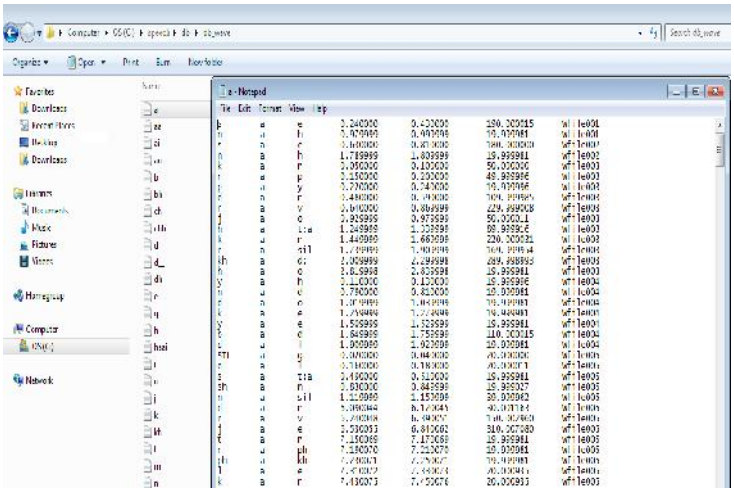

Figure 2: Screen shot of database generated for phoneme "a"

3.5 VOLUME OF DATABASE PREPARED

An analysis has been performed to calculate the volume of database prepared by taking a count of tokens generated for each phoneme occurring in the sentences selected to prepare the speech corpus. The distribution of tokens in the database is given in Table 3 given below.

Table 3. Number of tokens generated in our database

Phones	Tokens	% age	Phones	Tokens	%age
/ə/	752	6.498	t	362	3.1282
/ɑ/	1100	9.505	t^h	82	0.7086
/ɪ/	505	4.363	d	291	2.5146
/ii/	1285	11.104	d^h	97	0.83823
/U/	67	0.5789	n’	568	4.908
/u/	228	1.9702	p	301	2.6011
/ɛ/	112	0.9678	p^h	49	0.4234
/e/	123	1.0672	b	177	1.5295
/O/	331	2.806	b^h	78	0.6740
/ɔ/	56	0.4839	m	381	3.2924
/k/	756	6.5330	j	129	1.114
/k^h/	44	0.3801	r	880	7.604
/g/	217	1.8752	l	295	2.594
/g^h/	67	0.57898	v	344	2.972
/ng/	95	0.8209	s’a	208	1.797
/c/	95	0.8209	sa	87	0.7518
/c^h/	58	0.50121	s	446	3.8541
/j/	221	1.9097	h	348	3.0072
/j^h/	52	0.4493			
/ɲ’a/	85	0.7345			

From design of MRPIS database, there are altogether 11572 phoneme units from 630 sentences. It can be seen that /ɑ/ and /ii/ are the two phonemes with highest occurrence in database. They contribute to almost 11% and 9.5 % of total phoneme unit extracted. The least occurring phonemes are /p^h/ and /k^h/ contributing to 0.4234% and 0.38% of the phonemes occurring in database.

4. CHALLENGES IN DATABASE PREPARATION

Important issues involved in database preparation for development of various speech technologies are (i) creating a generic acoustic database that covers language variations (ii) designing of the recording prompts and recording of speech databases to be used by corpus – based speech synthesizers and (iii) Lack of availability of phonetically balanced material. Also, the recording of the inventory requires large number of recording sessions with long durations under strict supervision to ensure good voice quality during recording sessions. . The recording process is expensive and puts a restriction on the capability to create databases in large number of voices for each restricted domain application.

5. CONCLUSIONS

Development of TTS systems for DMRC will be extremely useful for metro rail passenger information systems (MRPIS). Here, a technique based on HMM has been proposed for automatic segmentation and identification of phoneme like speech units. The phoneme database is created by using the state of art technique based on HMM. The design of the database is such that each speech unit (phoneme) is available in various phonetic contexts. A medium size database has been prepared that consists of 630 utterances with 12,614 words, 11572 tokens of phonemes covering 38 phonemes occurring in database. This database is believed to augment database for application in development of TTS for metro rail. In addition to above, an in- depth

survey of efforts made in database developments for Hindi language has been performed. It discusses some core linguistic resources of Hindi language, available through various resources developed for usage in text-to-speech synthesis and speech recognition technology.

It is suggested that the recordings from various resources can be grouped into application domains that can be combined to generate inventories which can be integrated with speech synthesizers to develop TTS. Unfortunately many of the existing corpora or resources lack features that are highly desirable for their use in the scientific context. These shortcomings include problems with availability (in some cases the use of very specific interfaces is required), high costs or strict licenses that allow reuse and data collection. This paper highlights the issue of distribution constraints which needs to be solved. As some of the problems such as copyright cannot be eliminated, it would be advantageous to have more resources available electronically that can be used with fewer limitations. This shall make involvement of extensive number of institutions and industries possible in research and development of TTS in Hindi language.

REFERENCES

[1] Prahallad, Kishore, et al. "The IIIT-H Indic Speech Databases." Interspeech, 2012
[2] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_ speakers_in_India
[3] Kachru,Y. Hindi. John Benjamins Publishing company,Philadelphia, 2016.
[4] Dash, N S and B B Choudhary, “Why do we need to develop corpora for Indian languages”, International conferences on SCALLA , Bangalore, 2001,Vol. 11 in Elsevier.
[5] Sunita Arora, Babita Saxena, Karunesh Arora, S S Agarwal, “ Hindi ASR for Travel Domain,” In Proceedings of OCOCOSDA 2010, Kathmandu, Nepal.
[6] Agrawal, S. S. "Recent developments in speech corpora in Indian languages: Country Report of India." O-COCOSDA, Kathmandu (2010).
[7] Anandaswarup V, Karthika M et al., “Rapid Development of Speech to Speech Systems for Tourism and Emergency Services in Indian Languages”, In Proceedings of International Conference on Services in Emerging Markets, Hyderabad, India.
[8] Samudravijay K., P.V.S Rao and S.S. Agrawal, “Hindi Speech Data”, In proceedings of Sixth International Conference on Spoken Language Processing (ICSLP 2000),Beijing, China
[9] Shyam Agrawal, Shweta Sinha, Pooja Singh, Jesper Olsen, “ Development of Text and Speech Database for Hindi and Indian English specific to Mobile Communication Environment. In Proceeding of International Conference on The Language Resources and Evaluation Conference, LREC, Istanbul, Turkey, 2012.
[10] K.Srenivasa Rao,“ Application prosody model for Developing speech system”, International Journal of Speech Technology, 2011,Vol 11 in Elsevier
[11] Primer on Audio PC various issues associated with PC based audio technology, Online available on ttp://www.totalrecorder.com/primerpc.htm, 2016.
[12] Standards for Speech Data Capturing and Annotation, available online http://www.ldcil.org/download/samplingstandards.pdf, 2016.
[13] Balyan, Archana, S. S. Agrawal, and Amita Dev. "Automatic phonetic segmentation of Hindi speech using hidden Markov model" AI & society , Vol 27, no. 4 : 543-549,2000.
[14] Molau,S, Pitz, M, et al.,2001: Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum, proceedings, ICASSP, IEEE.
[15] Rabiner, L.R., 1989: A Tutorial on hidden markov models and selected applications in speech recognition, Proceedings. IEEE, vol. 77, no. 2, 257-286.
[16] Forney, J.D, 1978: The Viterbi Algorithm, proceedings, IEEE, vol.no.3, 268-278.

AUTHOR

Archana Balyan has obtained her BE (ECE) degree from Bangalore University and received the ME (ECE) degree from Delhi College of Engineering, University of Delhi. She has obtained her PhD degree from GGSIP University in the area of speech synthesis. She has more than 20 years of teaching experience and is an Associate Professor in Maharaja Surajmal Institute of Technology (a premier institute affiliated to GGSIP University, Delhi). She has published several papers in reputed International Journals and in the proceedings of leading conferences. Her research interests include speech synthesis, computational linguistics and Analog electronics.