Development of Kannada Speech Corpus for Continuous Speech Recognition

Anand H. Unnibhavi
Dept. of Electronics and Communication
Basaveshwara Engineering College
Bagalkot, India

D. S. Jangamshetti
Dept. of Electrical and Electronics Engineering
Basaveshwara Engineering College
Bagalkot, India

ABSTRACT
The paper presents, development of Kannada speech corpus for speaker independent continuous speech recognition. Speech corpus plays a key role in construction of Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) synthesis. The speech corpora is developed for the age group between 21 years to 45 years. Speech corpus for ASR system is developed by collecting text corpus in which data is recorded corresponding to the text corpus followed by Transliteration (phonetic representation of the text corpus) and finally a pronunciation dictionary is developed.

Keywords
ASR, TTS, G2P, Speech corpus

1. INTRODUCTION
Speech is the most effective and common way of communication between human. Human beings have long been motivated to create computer that can understand and talk like human. Demands for practical ASR systems in smart-phones have rapidly increased due to their convenience and user friendliness for information access [1]. Meanwhile, in recent years, ASR techniques have took a great leap forward with the help of Deep Neural Network (DNN) based approaches. The computers which can understand the spoken language have some of the applications in the domains like agriculture, health care and government services. Most of the Information in the world is digital and is accessible only to a few people who can understand the language [2]. Language technologies will provide the ways in the form of natural interfaces so that digital information can reach to the masses and facilitate the exchange of information across different people speaking different languages. Automatic Speech Recognition (ASR) deals with conversion of acoustic signals into text transcription in speech utterances. Even after years of extensive research and development, accuracy in ASR remains a challenge to researchers. There are number of well known factors which determine accuracy. The prominent factors include variations in context, speakers and noise in the environment. Therefore research in automatic speech recognition has many open issues with respect to small or large vocabulary, isolated or continuous speech, speaker dependent or independent and environmental robustness. In recent years, many free ASR software packages have been made available to the community. Examples are CMU Sphinx, Julius, and Kaldi [3]. The Speech technologies play a crucial role in multi-lingual societies such as India which has about 1652 dialects/native languages. Hindi written in Devanagari script is the official language and the other 17 languages recognized by the constitution of India are: 1) Assamese 2) Tamil 3) Malayalam 4) Gujarati 5) Telugu 6) Oriya 7) Urdu 8) Bengali 9) Sanskrit 10) Kashmiri 11) Sindhi 12) Punjabi 13) Konkani 14) Marathi 15) Manipuri 16) Kannada and 17) Nepali [4]. For having better performance of the speech recognizers it is inevitable to have speech corpus of that particular language. The first speech database is created by MIT with TIMIT database [5]. To test and build any recognition system, building speech corpus is an important task. IBM has built a large vocabulary continuous speech recognition system in Hindi [6] by bootstrapping existing English acoustic models. The amount of work done in Indian languages has not yet reached a critical level which can be used as real communication tool, where as the work done in speech technology for the English and other European language has reached to achieve higher accuracy rate. Speech recognition system has been built by HP Labs India and IBM research lab [7], which involved Hindi speech corpus collection and subsequent system building. There is lot of scope to develop language technology systems using Indian languages which are of different variations. To achieve such ambitious goals, the collection of standard speech databases is prerequisite. In this paper the development of speech database in kannada language for building large continuous speech recognition system is presented. The paper is organized as follows: Section II describes the History of kannada language along with classification of kannada characters. Section III deals with the basic requirement of text corpora such as sources of collecting text, speaker selection and software used for recording the speech. The section also describes grapheme to phoneme conversion, transliteration which converts kannada letters to roman symbols representing kannada phonemes in to its corresponding Devanagari script and word level dictionary. The concluding remarks are stated in the section IV.

2. KANNADA LANGUAGE CLASSIFICATION
Kannada also known as Canarese or Kanarese is a Dravidian language spoken by the kannada people in India, especially in the state of Karnataka (ಕರ್ನಾಟಕ) and by linguistic minorities in the states of Andhara pradesh, Tamil Nadu, Goa, Maharashtra, Kerala, and Telangana. The language has nearly 40 million Kannada native speakers who are called Kannadigas and a total of 50.8 million speakers according to a 2001 census. Kannada is one of the scheduled language (Scheduled languages are the ones included in 8th schedule of the constitution) of India and the official and administrative language of the state of Karnataka [10]. The Kannada language is written using the Kannada script and this language has evolved from the 5th-century Kadamba script. Kannada is attested epigraphically for about one and a half millennia. Literary old Kannada flourished in the 6th-century Ganga dynasty and during the 9th-century Rashtrakuta Dynasty. Kannada has an literary history of over a thousands of years. Kannada is a Southern Dravidian language and according to

Dravidian scholar Sanford B. Steever, its history can be conventionally divided into three periods: Old Kannada from 450–1200 CE, Middle Kannada from 1200–1700, and Modern Kannada from 1700 to the present. Kannada alphabet is popularly known as Aksharamale or Varnamale and the current Varnamale list consists of fifty characters. The fifty basic characters are classified into three categories such as Swaras (vowels), Vyanjanas (consonants) and Yogavahakas (part vowel, part consonants). There are fourteen vowels and are called swaras, two Yogavahakas and are called as Anusvara and Visarga. The Vyanjanas are classified into structured and unstructured consonants. The structured consonants are further classified into five groups according to which the tongue touches the palate of the mouth while pronouncing these characters [11].

3. SELECTING TEXT CORPORA

To develop a speech database, the basic requirement is to record the grammatically correct text corpus from different speakers. The text corpus should be correct in terms of typography and grammar. Phonetically rich sentences can be selected from a large set of text. The old source of collecting text data is from Books, Magazines and periodicals.Text should be in electronic form, so that it can be processed using computer. Therefore the text can be taken as input by typing the data in printed form from the sources like articles, periodicals, magazines and online available sources such as articles and online kannada news papers. Representative corpus of the language is collected by crawling content from websites of kannada newspapers, entertainment news, children story books, sports news, political news and science news. Totally thousand kannada text sentences are collected for developing speech database.

3.1 Speaker selection

Speech data is collected from the 10 native speakers of the kannada language who were comfortable in speaking and reading the language. The speakers were chosen such that all the diversities attributing to the gender, age are sufficiently captured. The recording is done in minimal background disturbance. Any mistakes made while recording have been undone by re-recording or by making the corresponding changes in the transcription set. Total of 1000 kannada sentences are collected from kannada news papers, story books etc. for utterances. Each speaker is asked to utter 10 sentences each. 10Speakers * 100 sentences = 1000 sentences. To record the sentences Wave surfer software is used with a sampling frequency of 16 kHz with bit depth of 16 bits. Wave Surfer is an audio editor widely used for studies of acoustic phonetics. It is a simple but fairly powerful program for interactive display of sound pressure waveforms [12], spectral sections, spectrograms, pitch tracks and transcriptions. The data is recorded in the silent room where background noise is absent.

3.2 Grapheme to Phoneme Convertion

The Text corpus has to be phonetized so that the distribution of the basic recognition units, the phones, diphones, syllables, etc in the text corpus can be analyzed. The process of converting a sequence of letters into a sequence of phones is called grapheme-to-phoneme conversion, sometimes shortened as G2P Phonetizers. Grapheme-to-phoneme (G2P) models are key components in speech recognition and text-to-speech systems as they describe how words are pronounced. The job of a grapheme-to-phoneme [13] algorithm is thus to convert a letter string like cake into a phone string like [K EY K]. The online textual data available on the archive uses a Devnagri font named “sudipto”. The sudipto font employs UTF-8 encoding scheme [14]. Kannada is a Dravidian language spoken predominately by kannada people and is closely related to Tamil, Malayalam, Irula, Kodagu, Toda and Kota. Badaga, spoken in the Nilgiri Mountains to the south of Karnataka, is thought to be a recent off shoot of Kannada [15]. Constitution of India had recognized 17 languages and Kannada is one among these. Hindi uses character based Devnagri script; a Devnagri character represents either a standalone vowel or vowel in combination with one or more consonants. Also Kannada language is more similar to Hindi Devanagri script. The G2P involves generating Roman symbols representing Kannada phonemes from the UTF-8 code. Kannada is represented by a code that is either 2 or 3 bytes long. All vowel modifiers and most pure consonants are encoded using 2 bytes. Some of the examples of grapheme to phoneme conversion of kannada are explained in detail in Table 1. Which shows an example for kannada G2P conversion for word ‘ರಾಜಕೀಯ'. In case of letter 'ರಾ' consonant ರ is followed by vowel ಆ i.e., 'ರ + ಆ = ರಾ' and the corresponding symbol is ‘ra + aa = raa’, forಜ(consonant) the symbol is ‘ja’ similarly letter ‘ಕೀ' can be written as 'ಕ + ಈ = ಕೀ', vowel ಈ following consonant ಕ symbol is ‘ka + ii =kii’, for ಯ(consonant) symbol is ‘ya’ . In Table 2 another example is discussed for kannada G2P convertion for the word 'ಭಾರತದ'. Letter ಭಾ can be written as a combination of consonant ಭ and followed by vowel ಆ i.e., 'ಭ + ಆ = ಭಾ', The decimal sequence of complete word is 3245 3206 3248 3236 3238 with the symbol as ‘bharatada’. Similarly for the word ಈಶಾನ್ಯ the G2P conversion is given in Table 3. with decimal sequence 3208 3254 3205 3230 and is equivalent to ‘iishaanya’. In Table 4, letter ರೋ is equivalent to ‘ ರ + ೋ = ರೋ’ with G2P conversion of word ‘ಆರೋಗ್ಯ' is ‘aarogya’.

Table 1. G2P of word ರಾಜಕೀಯ

DECIMAL SEQUENCE	SYMBOL	GRAPHEME
3248	ra	ರ
3206	aa	ಆ
3228	ja	ಜ
3221	ka	ಕ
3208	ii	ಈ
3247	ya	ಯ

Table 2. G2P of word ಭಾರತದ

DECIMAL SEQUENCE	SYMBOL	GRAPHEME
3245	bha	ಭ
3206	aa	ಆ
3248	ra	ರ
3236	ta	ತ
3238	da	ದ

Table 3. G2P of word ಈಶಾನ್ಯ

DECIMAL SEQUENCE	SYMBOL	GRAPHEME
3208	ii	ಈ
3254	sha	ಶ
3205	a	ಅ
3230	nya	ಞ

Table 4 G2P of word ಆರೋಗ್ಯ

DECIMAL SEQUENCE	SYMBOL	GRAPHEME
3206	aa	ಆ
3248	ra	ರ
3275	oo	ೋ
3223	ga	ಗ
3247	ya	ಯ
3205	a	ಅ

3.3 Database collection

In this work, four different female speakers are asked to utter 300 hundred Kannada sentences each, total of 1200 (4speakers x 300 sentences = 1200 sentences) sentences are recorded using wave surfer software. The uttered sentences are sampled at frequency of 16 kHz with 16 bit depth. The details of procedure of sentences collected, creation of dictionary, pronunciation dictionary etc. are discussed below.

3.4 Selection of sentences

For creation of Pronunciation Dictionary and word level Dictionary, short sentences with a minimum of 5 and maximum of 10 words are selected for the generation of text corpus. The sentences are carefully selected so that they are meaningful and not contain any offensive or sensitive words.

3.5 Phonetic richness of sentences

Phonetic rich sentences are needed for robust estimation of the statistical model parameters of context sensitive phonemes. A set of sentences is considered to be phonetically rich if it contains all permissible triphones of the language in sufficient quantity. If there are M phonemes in a language, there can be M^n triphones. In order to develop speaker independent recognition system, it is required to collect speech data for a large number of speakers. Also it is necessary to collect sentences of minimum 5 words to maximum 10 words.

3.6 Transliteration

Indian Script Code for Information Interchange (ISCII) is a coding scheme for representing various writing systems of India. Kannada is one of the supported indian script. It encodes the main indic script and a Roman transliteration. The coding procedure is, first phonetic Kannada rich sentences collected, is converted to Devnagari convention by comparing with the Hindi Devanagiri script. Sentences are converted to roman symbols representing kannada phonemes into its corresponding Devanagari script. This transcription and word level Dictionary are implemented by running a Perl code resulting in 2253 words, and pronunciation dictionary is created for the word level dictionary. A total of 64 distinct phones are generated from the collected sentences. A set of sentences is considered to be phonetically rich if it contains all permissible triphones of the language in sufficient quantity.

4. CONCLUSION

In this work a total of 1200 Phonetic rich sentences collected and recorded using wave surfer software. Also the work gives the procedure of collecting the text corpus, implementation of continuous speech database for these 1200 sentences of Kannada text speech corpus for speaker independent continuous speech recognition. The database includes transliteration of the collected text corpus, generation of word level dictionary which consists of 2253 words, and corresponding pronunciation dictionary is generated. A total of 64 unique distinct phones are obtained. The database will cover all the phonetic variations. This databases will be helpful to develop a Robust speaker independent Automatic Speech Recognition System.

5. REFERENCES

[1] Hay Mar Soe Naing, Aye Mya Hlaing, "A Myanmar Large Vocabulary Continuous Speech Recognition System", Proceedings of APSIPA Annual Summit and Conference 16-19 Dec. 2015 IEEE.

[2] Ahmad A. M. Abushariah, Teddy S. Gunawan,"English Digits Speech Recognition System Based on Hidden Markov Models", International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia 978-1-4244-6235- 3/10/$26.00 ©2010 IEEE.

[3] Rohit Kumar, S.P. Kishore, Anumanchipalli Gopalakrishna, Rahul Chitturi, Sachin Joshi, Satinder Singh, "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems", Proceedings of International Conference on Speech and Computer (SPECOM), Patras, Greece, Oct 2005 IEEE.

[4] Biswajit Das, Sandipan Mandal and Pabitra mitra, "Bengali speech corpus for continuous automatic speech recognition system", International Conference on Speech

Database and Assessments(Oriental COCOSDA) 2011 IEEE.

[5] Niklas Vanhainen and Giampiero Salvi, "Free Acoustic and Language Models for Large Vocabulary Continuous Speech Recognition in Swedish", Ninth International Conference on Language Resources and Evaluation (LREC'14), May, 26-31, 2014,Reykjavik, Iceland, ISBN: 978-2-9517408-8-4.

[6] Chalapathy Neti, Nitendra Rajput and Ashish Verma, "A Large Vocabulary Continuous Speech Recognition System for Hindi", IBM India Research Lab, Volume: 48, Issue 5.6, Sep. 2004 IEEE.

[7] Tejas Godambe and Samudravijaya K, "Speech Data Acquisition for voice based Agricultural Information Retrieval", Proceeding of 39th All India DLA Conference, Punjab University, Patiala, India 2011.

[8] Tejas Godambe and Samudravijaya K, "Speech Data Acquisition for voice based Agricultural Information Retrieval", Proceeding of 39th All India DLA Conference, Punjab University, Patiala, India 2011.

[9] G. V. Mantena, S. Rajendran, B. Rambabu, S. V. Gangashetty, B. Yegnanarayana and K. Prahallad, "A speech-based conversation system for accessing agriculture commodity prices in indian languages", Proceedings of IEEE Hands-free Speech Communication and Microphone Arrays Edinburgh UK, Edinburgh, UK, 2011.

[10] https://en.wikipedia.org/wiki/Kannada.

[11] http://shodhganga.inflibnet.ac.in/bitstream/10603/104462/12/12_chapter %202.pdf

[12] http://www.sif.us.es/fil/publicaciones/apuntes/mpinedaperez/Wavesurfer. Pdf.

[13] http://www.voxforge.org/home/docs/faq/faq/what-is-g2p

[14] http://www.personal.psu.edu/ejp10/symbolcodes/bylanguage/kannadach art.html.

[15] http://languagemanuals.weebly.com/uploads/4/8/5/3/4853169/kannada.pdf