Speaker-Independent Automatic Speech Recognition System for Mobile Phone Applications in Punjabi

Puneet Mittal[1(✉)] and Navdeep Singh[2]

[1] BBSB Engineering College, Fatehgarh Sahib, India
puneet.mittal@bbsbec.ac.in
[2] Mata Gujri College, Fatehgarh Sahib, India
navdeep_jaggi@yahoo.com

Abstract. Speaker-independent Automatic Speech Recognition (ASR) system based mobile phone applications are gaining popularity due to technological advancements and accessibility. Speech based applications may provide mobile phone accessibility and comfort to people performing activities where hand-free phone access is desirable e.g. drivers, athletes, machine operators etc. Similarly, users with disabilities like low vision, blindness and physically challenged may use it as an assistive technology. Development of ASR system for a specific language needs accurate, reliable and efficient acoustic model having language-specific pronunciation dictionary. Punjabi language is one of the popular languages worldwide having more than 150 million speakers. Three acoustic models- continuous, semi-continuous and phonetically-tied are developed based on three pronunciation dictionaries- word, sub-word and character based. Analysis of performance results validate Punjabi language principle "One word one sound" by having better accuracy and reliability for character based pronunciation dictionary than others. Further, phonetically-tied model outperforms others in terms of accuracy, word error rate and size due to reasonable number of Gaussians.

Keywords: Acoustic model · Language model · Punjabi · CMU sphinx · Dictionary

1 Introduction

Automatic speech recognition (ASR) is the transcription of speech signal into readable text to identify and process human voice. Speech provides vocalized communication through large vocabularies having different words formed out of phonetic combination of sound units called phoneme. Based upon the vocabulary size, a word may have phonetic representation as a word itself for small vocabulary, syllable-based or sub-word based representation for large vocabulary, or character-based representation for languages having characters with distinct sound. ASR systems are being widely used in various applications for desktop, laptop and hand-held devices like mobile phones, where each application has its own set of requirements. High speed is desirable for real-time applications while accuracy is the key concern for command and control

S.M. Thampi et al. (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*, Advances in Intelligent Systems and Computing 678,
DOI 10.1007/978-3-319-67934-1_33

applications and dictation applications. Efficient space utilization and high speed is desirable for mobile phone applications while high speed is expected from desktop applications having ample storage space available.

ASR systems have been an active area of research for almost last six decades. Research started in this field in the year 1952 with the development of Aurdey [4], a speaker-dependent speech recognizer having 97–99% digit recognition accuracy. It was followed by DoD's DARPA Speech Understanding Research (SUR) program [10] and Carnegie Mellon's "Harpy" speech-understanding system [16] having ability to recognize 1011 words. Hidden Markov Model (HMM) based methods gained popularity in 1980s and are still being widely used.

A major revolution in this field came in the year 1990 with the development of Sphinx [15]. Sphinx is an accurate, large vocabulary, speaker independent, continuous speech recognition system. It introduced three acoustic models- continuous [19], semi-continuous [7] and phonetically-tied [6]. They differ in the way their mixture of Gaussians is built, that is used to compute the score of each frame. In continuous model every senone has its own set of gaussians thus the total number of gaussians in the model is about 150 thousand. It requires much processing to compute the mixture efficiently. In semi-continuous model, there are total 700 gaussians for use with different mixtures to score the frame. Due to the smaller number of gaussians semi-continuous models are fast, but because of more hardcoded structure their accuracy is low as compared to continuous models. Phonetically-tied models (PTM) use about 5000 gaussians thus providing better accuracy than semi-continuous. It achieves almost same accuracy as of continuous model with less processing and storage requirements. So, it is significantly faster than continuous models and can be used for mobile applications.

ASR system requires development of an efficient acoustic model based on language specific pronunciation dictionary. This paper proposes the development of efficient acoustic model for Punjabi language that can be used to build ASR system for mobile phone applications. Section 2 covers the related work in the field of ASR followed by problem formulation in Sect. 3. Section 4 gives introduction to Punjabi language while proposed methodology is explained in Sect. 5. Section 6 gives detailed development of ASR system for Punjabi language. Results are analyzed in Sect. 7 followed by discussion and Conclusions are given in Sect. 8.

2 Related Work

Various ASR systems have been proposed by researchers from time to time. Most of the applications like Google Voice Search [23] are in English, Spanish or other European languages. Wang et al. [27] developed ASR for Chinese having complete recognition of continuous Mandarin speech with large vocabulary. Walha et al. [26] developed ASR for Standard Arabic language using HTK toolkit. Satori et al. [22] trained a model for Amazigh using CMU Sphinx tools having 92.89% accuracy for 16 GMM. Naing et al. [18] developed large vocabulary continuous speech recognition system for Myanmar language using deep neural network approach. Researchers are also working on other Asian languages like Japanese, Korean [24] etc. Indian languages like Hindi [11], Assamese [1], Tamil [25], Bengali [3] etc. are also being explored. Till now, little work

has been done on speech recognition in Punjabi [12–14]. Dua et al. [5] proposed isolated word ASR system for Punjabi using HTK toolkit with overall system performance of 95.63% for a limited vocabulary having 115 Punjabi words.

3 Problem Formulation

Mobile phones have become future communication instruments by replacing computers and laptops with the advent of better hardware, computation and storage capabilities, and battery technology improvements. Speaker-dependent applications embedded in mobile phones are being ignored by the majority of users due to usability, accuracy, speed and storage constraints. Speaker-independent applications as a low cost, high capacity alternative to speaker-dependent applications are being developed to provide user-friendly, accurate, fast and low memory interface [17] for simple features like phone dialing and dictation to complex command and control features. It covers speech based applications like continuous digit dialing, name dialing, command and control for menus and navigation systems, games, and interactive man-machine interfaces.

Speech based mobile phone applications provide accessibility and comfort in situations where a person is driving a vehicle or doing some activity and needs to dial a phone number, send SMS or use GPS etc. It acts as an assistive technology for users with disabilities like low vision, blindness and physically challenged.

Punjabi is the native language of people of Punjab state in India. It is spoken by more than 150 million native speakers worldwide. According to a report by the Commissioner for Linguistic Minorities [2], 91.69% people speak Punjabi in Punjab state. 62.52% people of Punjab live in rural area [20]. People from rural areas of Punjab cannot use speech based applications built in foreign languages. So there is a need to develop speaker-independent Punjabi based applications for mobile phones. Currently, there is no acoustic model specifically built for mobile phone applications in Punjabi. This paper aims to build efficient acoustic model which can be used to develop speaker-independent mobile phone applications for Punjabi.

4 Punjabi Language

Punjabi is an Indo-Aryan language [21] widely spoken in countries like India, Pakistan, Canada and UK. It is spoken by more than 150 million native speakers worldwide. Gurmukhi and Shahmukhi scripts are used for Punjabi in India and Pakistan respectively. Gurmukhi script being alphasyllabary in nature consists of two types of symbols- consonants and vowels. It is written from left-to-right and is spelled phonetically. Gurmukhi script is based on “one sound one symbol” principle.

Punjabi is a meaningful collection of sentences made up of words where each word is a collection of phones [9]. Punjabi is formed based on phones or sounds having 41 alphabets and 9 dependent vowels of Gurmukhi script. Out of 41 alphabets, 38 are consonants (from ਸ to ਲ਼) while 3 alphabets (ਉ, ਅ, ਏ) are used in independent vowel form. In addition to these, 3 auxiliary signs are also available as shown in Table 1. Words in Punjabi are formed from different combinations of consonants and dependent vowels

Table 1. Punjabi character set.

Consonants	ੲ	ਸ	ਹ							
	ਕ	ਖ	ਗ	ਘ	ਙ					
	ਚ	ਛ	ਜ	ਝ	ਞ					
	ਟ	ਠ	ਡ	ਢ	ਣ					
	ਤ	ਥ	ਦ	ਧ	ਨ					
	ਪ	ਫ	ਬ	ਭ	ਮ					
	ਯ	ਰ	ਲ	ਵ	ੜ					
	ਸ਼	ਖ਼	ਗ਼	ਜ਼	ਫ਼	ਲ਼				
Independent vowels	ਅ	ਆ	ਇ	ਈ	ਉ	ਊ	ਏ	ਐ	ਓ	ਔ
Dependent Vowels		ਾ	ਿ	ੀ	ੁ	ੂ	ੇ	ੈ	ੋ	ੌ
Auxiliary Sign	ਂ	ਁ	ੱ							

like ਕਾ(ka), ਕਿ(Ki), ਕੀ(Ke), ਕੁ(Ku), ਕੂ(Koo), ਕੇ(Kae), ਕੈ (Kai), ਕੋ(Ko), ਕੌ(Kau). For example, the word ਚਾਰ is a combination of consonants ਚ and ਰ with vowel ਾ forming ਚ ਾ ਰ (CVC). A word may have a vowel at the beginning followed by one or more consonants and vowels e.g. ਇੱਕ. The words are joined together to form sentences as per language rules to make the sentences meaningful. In this paper Punjabi character set having 38 consonants, 10 independent vowels, 9 dependent vowels and 3 auxiliary signs is considered.

5 Proposed Methodology

Three methodologies are proposed for Punjabi phonetic representation based on words, sub-words and characters. In word-based methodology, each word is uniquely identified as an acoustic unit. As no segmentation of word into sub-words or characters is done, the word ਇੱਕ is represented as ਇੱਕ and ਤਿੰਨ as ਤਿੰਨ. In sub-word-based methodology, all characters of each word are scanned for identification of characters like consonants, dependent vowels, independent vowels and auxiliary signs to form sub-words based upon certain rules (Table 2). The word ਇੱਕ is segmented into two sub-words ਇੱ and ਕ while word ਤਿੰਨis segmented into sub-words ਤਿੰ and ਨ. In character-based methodology, each word is segmented into individual characters based upon certain rules (Table 3) and each character is stored in an array as a unique acoustic unit. The word ਇੱਕ is segmented into three characters ਇ, ੱ and ਕ while word ਤਿੰਨ is segmented into four characters ਤ, ਿ, ਂ and ਨ.

Further, three acoustic models- continuous, semi-continuous and phonetically-tied are developed for words, sub-words and characters identified with the above three methodologies at different Gaussian densities (4, 8, 16, 32, 64, 128, 256). Detailed comparative analysis of these three acoustic models will be conducted for different

Table 2. Rules for Sub-word based segmentation.

Description	Pattern	Example of sub word
Independent Vowel	IV	ਇ (ਇ-IV)
Independent vowel followed by auxiliary sign	IV-AS	ਇੱ (ਇ –IV ੱ-AS)
Consonant	C	ਕ (ਕ-C)
Consonant followed by Dependent Vowel	C-DV	ਦੇ (ਦ-C ੇ-DV)
Consonant followed by Dependent Vowel and auxiliary sign	C-DV-AS	ਝਿੰ (ਝ-C ਿ-DV ੰ-AS)
Consonant followed by auxiliary sign	C-AS	ਪੰ (ਪ-C ੰ-AS)

Table 3. Rules for character based segmentation.

Description	Pattern	Example
Independent Vowel	IV	ਇ
Consonant	C	ਕ
Dependent Vowel	DV	ੇ
Auxiliary Sign	AS	ੰ

performance parameters like Word Error Rate (WER), Accuracy, Speed, Size and Time taken to build the model. Based upon the outcome of comparative analysis, an optimal acoustic model will be recommended for the development of Automatic Speech Recognition model for Punjabi Mobile Applications.

6 Punjabi Speech Recognition System

This section describes the process of design and development of an efficient acoustic model for Punjabi automatic speech recognition system for mobile phone applications. Figure 1 shows the components of the proposed system. Initially the input speech signal is pre-processed at front end followed by extraction of acoustic features. The acoustic model, language model and dictionary are developed for the Punjabi, which are used by the speech recognition engine to identify the words spoken by the user. Speech corpus and Text corpus are the prerequisite for the development of Acoustic model and Language model respectively. The ultimate goal is to allow mobile phone to correctly recognize all words spoken by user in real time independent of vocabulary size, noise, speaker characteristics or accent.

The Punjabi ASR System is built in the training phase while recognition performance is evaluated during the testing phase. The major portion of the speech corpus is

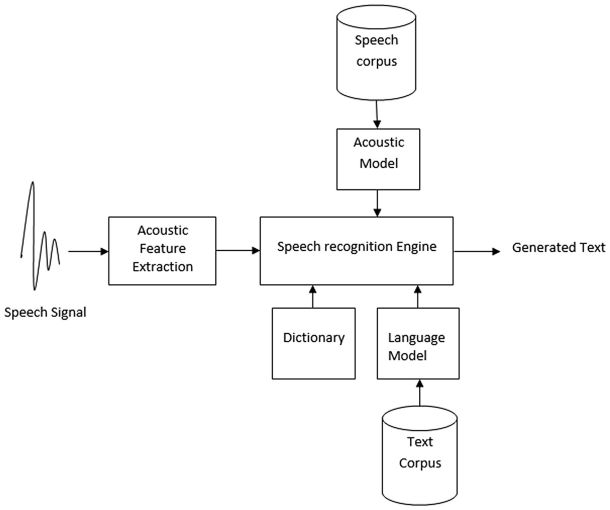

Fig. 1. Components of proposed system.

used to train the system while rest of the recordings is used for testing purpose. Training phase covers speech and text corpus preparation, acoustic feature extraction, dictionary preparation, acoustic model development and language model development. These are finally used by speech recognition engine for text generation. Testing phase covers the evaluation of performance parameters like accuracy, error rate, speed and space utilization for the developed system. This section covers the training phase in detail while the testing phase is discussed in results section.

6.1 Text and Speech Corpus Preparation

Text corpus is the prerequisite for the language model. Text corpus for Punjabi consists of 10 digits (0 to 9) for phone number and two commands 'saaf karo' (to clear number) and 'dial karo' (to dial number). Speech corpus is the prerequisite for acoustic model development. Speech corpora required for acoustic model are not available for Punjabi. The speech corpus for Punjabi is designed to satisfy a set of criteria, which specify the required quality of speech data and the proportional distribution of data with different speaker characteristics. Table 4 provides the technical specifications of the speech recordings.

The speech corpus is representative of native speakers of the Punjabi who are comfortable in speaking and reading the language. The speakers having all the diversities attributing to the gender, age and dialect are chosen for recordings. Every speaker has its own style of speaking, especially male and female voices are quite different. Figures 2 and 3 show waveform representation of digits 0-9 for male and female voices. Male speakers are generally having higher pitch and frequency than the female speakers. Speech recordings of 50 speakers are recorded for 10 digits (0 to 9) and two commands- 'saaf karo' (to clear) and 'dial karo' (to dial). Out of these recordings, the training set consists of 6 h 25 min of speech from 35 speakers (18

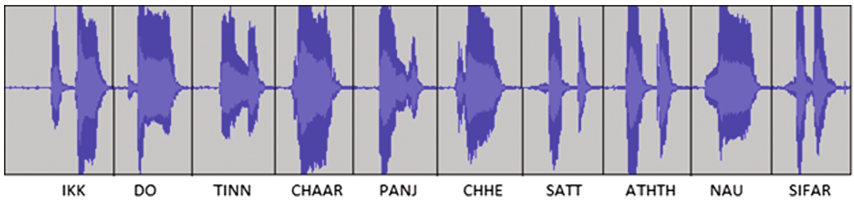

Fig. 2. Waveform of 10 Punjabi digits in male voice.

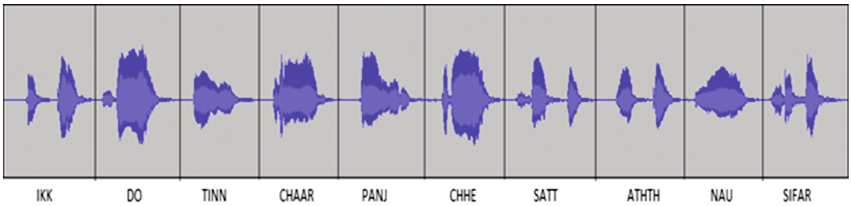

Fig. 3. Waveform of 10 Punjabi digits in female voice.

Table 4. Technical Details of Recordings.

Parameter	Value
Sampling rate	16 kHz
Number of bits	16
Number of channels	1, Mono
Audio data file format	.wav
Corpus	Punjabi 10 digits and 3 words
Number of speakers	50
No. of male speakers	25
No. of female speakers	25
Range of age group of speakers	18–35 years
Average recording time per speaker	11 min per speaker ∼9 h for all speakers
Number of recordings per speaker	118
Total number of recordings	5900
Size of raw speech	1010 MB
Condition of noise	Normal life
Window type	Hamming, 25.6 ms
Frames overlap	10 ms

Table 5. Punjabi dataset description.

Data set	Number of tokens	Number of speakers	Speakers' gender		Total number of recordings	Recording time
Punjabi_corpus (training)	13 (10 digits and 3 words)	35	18 female	17 male	4130	6 h 25 min
Punjabi_corpus (testing)	13 (10 digits and 3 words)	15	7 female	8 male	1770	2 h 45 min

female and 17 male) while testing set comprises of 2 h 45 min of speech from 15 speakers (7 female and 8 male). Mobile phone is used to collect recordings having minimal background disturbance. Speakers were asked to utter digits in sequence as well as at random for better accuracy and the recordings were stored in.wav files. Any mistakes made while recording have been undone by re-recording or by making the corresponding changes in the transcription set. Table 5 provides details of training and testing data sets.

6.2 Acoustic Feature Extraction

The training starts with the process of feature extraction. It is one of the most important and crucial steps in speech recognition. In this step parametric and acoustic-phonetic speech features are extracted from the recordings and stored in.mfc file. The unwanted and redundant speech signals are removed to improve the recognition accuracy and pre-processed necessary speech signals are forwarded to the speech recognition engine. The acoustic feature consists of first and second derivatives of 13 dimensional Mel Frequency Cepstral Coefficients (MFCC). The window size of 25 ms and frame shift of 10 ms is considered for MFCC.

6.3 Pronunciation Dictionary

ASR relies on the comprehensiveness of pronunciation dictionary that maps words to their corresponding pronunciation forms in terms of their phonetic representation in a specific language. As discussed earlier, pronunciation dictionary for Punjabi may have word-based, sub-word-based or character-based phonetic representation. So, three pronunciation dictionaries are created for the proposed system.

The word-based dictionary consists of 13 words, Sub-word based dictionary consists of 22 sub-words and character-based dictionary consists of 24 unique characters,

Table 6. Rules Pronunciation of Punjabi digits.

Punjabi Digits in English	Digits in Numerical form	Digits in Punjabi – Word based Phonetic Representation	Character based Phonetic Representation	Sub word based Phonetic Representation
Ikk	1	ਇੱਕ	ਇ ੱ ਕ	ਇੱ ਕ
Do	2	ਦੋ	ਦ ੋ	ਦੋ
Tinn	3	ਤਿੰਨ	ਤ ਿ ੰ ਨ	ਤਿੰ ਨ
Chaar	4	ਚਾਰ	ਚ ਾ ਰ	ਚਾ ਰ
Panj	5	ਪੰਜ	ਪ ੰ ਜ	ਪੰ ਜ
Chhe	6	ਛੇ	ਛ ੇ	ਛੇ
Satt	7	ਸੱਤ	ਸ ੱ ਤ	ਸੱ ਤ
Athth	8	ਅੱਠ	ਅ ੱ ਠ	ਅੱ ਠ
Nau	9	ਨੌ	ਨ ੌ	ਨੌ
Sifar	0	ਸਿਫਰ	ਸ ਿ ਫ ਰ	ਸਿ ਫ ਰ

Table 7. Pronunciation of Punjabi words.

Punjabi Words in English	Words in Punjabi (Gurmukhi Script)	Character based Phonetic Representation	Sub word based Phonetic Representation
Saaf	ਸਾਫ਼	ਸ ◌ਾ ਫ਼	ਸਾ ਫ਼
Karo	ਕਰੋ	ਕ ਰ ◌ੋ	ਕ ਰੋ
Dial	ਡਾਇਲ	ਡ ◌ਾ ਇ ਲ	ਡਾ ਇ ਲ

representing 10 digits and 03 words of Punjabi. Tables 6 and 7 show the phonetic representation of Punjabi digits and words in English form, Numerical form, word-based, sub-word-based and character-based.

6.4 Acoustic Model Development

An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word called as phoneme. It represents the relationship between the recorded speech and the phonemes. From the speech corpus, 70% of recordings by 50 speakers have been used as a statistical base from which the acoustic model has been developed. HMM based acoustic model trainer Sphinxtrain has been used to create statistical representations for each phoneme in Punjabi. The words are represented as sequence of phonemes where each phoneme has its own HMM having a sequence of states. From each speech recording, sequence of feature vectors are extracted and computed. The basic 3-state HMM model is used for each Punjabi phoneme having one state for the transition into the phoneme, one for the middle part and one for the transition out of the phoneme which join models of HMM units together in the ASR engine.

6.5 Language Model

Language model is a probability distribution over sequence of words. It is used for searching the correct word sequence by estimating the likelihood of the word based on previous words. CMU – Cambridge statistical language modeling toolkit (2016) has been used to develop language model for Punjabi. Text corpus of Punjabi having digits and commands is used for language model development.

7 Experimental Results

The proposed ASR system is developed having the ability to convert real-time speech into text and recognize the digits and commands spoken by the user. To evaluate the system performance against desired performance parameters testing is performed. Pocketsphinx [8], speech recognition system for hand held devices is used as decoder. From the speech corpus, 30% of recordings by 50 speakers are used for testing purpose. The experiments included training and testing of three acoustic models with three

pronunciation dictionaries on different GMMs. Each model has been evaluated for the following performance parameters:

Word Error Rate (WER): It is defined as the sum of word errors divided by the number of reference words. It takes into account three error types: *substitution* (the reference word is replaced by another word), *insertion* (a word is hypothesized that was not in the reference) and *deletion* (a word in the reference transcription is missed). Word error rate can be calculated as:

$$WER = (S + I + D)/N \tag{1}$$

where S, I and D represent substitution, insertion and deletion errors respectively while N is total number of reference words.

Accuracy (%WAcc): It is defined as the percentage of words correctly recognized by the speech recognition system. It can be calculated as

$$\%WAcc = 100 - \%WER \tag{2}$$

where %WER is the percent word error rate.

Build Time: It is the amount of time taken to build the acoustic model from the training data. It starts with the feature extraction and finishes when acoustic model is fully built.

Decoder Speed: It is the average time taken by the acoustic model to recognize a word. It specifies the CPU time taken by the decoder to recognize speech of one second duration. An average speed of 0.02 xRT (Real time) means that the decoder takes 0.02 s of CPU time to recognize speech of one second duration. The speed of decoder increases with decrease in average time.

Memory Size: It specifies the storage space required to store the fully built acoustic model.

ASR model having minimum WER, build time and memory size with maximum accuracy and decoder speed are desirable for optimal performance. Results of the three models are analyzed and compared to recommend optimal model for development of ASR for Punjabi mobile applications.

7.1 Performance Analysis of Continuous Acoustic Model

The results of Continuous acoustic model for different pronunciation dictionaries are shown in Table 8. It can be observed that the model attains maximum accuracy of 97.5% for character-based dictionary having WER of 2.5. The maximum accuracy for word and sub-word based dictionary is 85.6% and 94.5% having WER of 14.4% and 5.5% respectively.

It is worth noting that the accuracy of continuous model initially increases with increase in GMMs but decreases for very high value of GMMs. This happens due to the

Table 8. Continuous models.

GMM	WER (%age)			Accuracy (% age)			Time (mins)			Decoding speed (xRT)			Size (Mb)		
	W	S	C	W	S	C	W	S	C	W	S	C	W	S	C
4	23.1	5.6	**2.5**	76.9	94.4	**97.5**	27	38	**37**	0.02	0.02	**0.03**	0.45	0.37	**0.35**
8	16.7	**5.5**	2.5	83.3	**94.5**	97.5	20	**49**	34	0.02	**0.02**	0.03	0.75	**0.69**	0.68
16	**14.4**	5.6	3	**85.6**	94.4	97	**29**	68	35	**0.02**	0.04	0.03	**1.33**	1.34	1.51
32	14.9	6.6	3.8	85.1	93.4	96.2	47	109	60	0.03	0.07	0.04	2.49	2.63	2.82
64	17.3	8.5	5	82.7	91.5	95	213	187	695	0.04	0.2	0.06	4.83	5.23	5.44
128	19.2	9.7	6	80.8	90.3	94	387	368	1234	0.07	0.22	0.12	9.5	10.2	10.5
256	20.5	10.3	7	79.5	89.7	93	975	667	1845	0.1	0.24	0.2	14.6	15.6	15.8

W-Word based, S-Sub-word based and C-Character based

presence of own set of senone Gaussians in continuous acoustic models, that increases drastically with the increase in GMMs thereby hampering the efficiency of mixture computation. So, it is not advisable to build continuous models above 16 GMMs. The character based continuous model outperforms others in terms of accuracy, WER, build time and space requirements. Its only limitation is that the speed of decoder is low that can be neglected at the price of high accuracy.

7.2 Performance Analysis of Semi-continuous Acoustic Model

The results of Semi-Continuous acoustic model for different pronunciation dictionaries are shown in Table 9.

Table 9. Semi continuous models.

GMM	WER (%age)			Accuracy (% age)			Time (mins)			Decoding speed (xRT)			Size (Mb)		
	W	S	C	W	S	C	W	S	C	W	S	C	W	S	C
4	60.9	45.7	33.4	39.1	54.3	66.6	11	20	21	0.01	0.01	0.01	0.19	0.07	0.042
8	36.2	17.7	14.8	63.8	82.3	85.2	12	25	31	0.01	0.01	0.01	0.21	0.09	0.64
16	23.6	11.0	8.8	76.4	89.0	91.2	17	36	41	0.01	0.01	0.01	0.25	0.13	0.11
32	18.0	5.8	4.2	82.0	94.2	95.8	30	62	71	0.01	0.01	0.01	0.33	0.22	0.20
64	10.7	5.2	3.9	89.3	94.8	96.1	51	114	123	0.01	0.01	0.01	0.49	0.4	0.38
128	**10.4**	5.0	3.3	**89.6**	95.0	96.7	**237**	226	243	**0.01**	0.01	0.01	**0.83**	0.75	0.74
256	12.9	**4.5**	**3.0**	87.1	**95.5**	**97.0**	803	**506**	**490**	0.01	**0.01**	**0.01**	1.44	**1.45**	**1.44**

7.3 Performance Analysis of PTM Acoustic Model

The results of PTM model for different pronunciation dictionaries are shown in Table 10. It shows that the model attains maximum accuracy of 97.5% for character-based dictionary having WER of 2.5 at 8 GMMs. The maximum accuracy for word and sub-word based dictionary is 87.8% and 94.8% having WER of 12.2% and 5.2% respectively.

Table 10. PTM model.

GMM	WER (%age)			Accuracy (% age)			Time (mins)			Decoding speed (xRT)			Size (Mb)		
	W	S	C	W	S	C	W	S	C	W	S	C	W	S	C
4	26.1	9	4.7	73.9	91	95.3	12	32	46	0.02	0.02	0.02	0.21	0.1	0.078
8	16.1	6	**2.5**	83.9	94	**97.5**	18	48	**54**	0.02	0.02	**0.02**	0.25	0.16	**0.14**
16	14.3	**5.2**	2.7	85.7	**94.8**	97.3	30	**85**	91	0.02	**0.03**	0.02	0.33	**0.27**	0.26
32	13.9	6	3	86.1	94	97	65	177	186	0.02	0.03	0.05	0.49	0.5	0.49
64	**12.2**	6	2.9	**87.8**	94	97.1	**107**	275	511	**0.03**	0.05	0.05	**0.81**	0.95	0.79
128	14.8	8.2	3.1	85.2	91.8	96.9	1207	310	634	0.04	0.08	0.08	1.45	1.9	1.87
256	17.5	9.6	3.9	82.5	90.4	96.1	1475	584	1424	0.08	0.2	0.22	3.4	3.6	3.73

Results indicate that the accuracy of PTM model initially increases with increase in GMMs but decreases slightly for higher value of GMMs. Character based PTM model is having high accuracy, low WER, low build time, low decoding speed and low space requirement than others.

The performance analysis for the three acoustic models clearly indicates that results obtained with character-based pronunciation dictionary are consistent and far better than word-based and sub-word-based pronunciation dictionaries. So, it is recommended to use character-based pronunciation dictionary for Punjabi ASR system. Further, in-depth study of the three acoustic models with character-based pronunciation dictionary outcomes maximum accuracy and minimum WER at only 4 Gaussians resulting low storage requirement and build time with high decoding speed for continuous models. The time required to build the PTM model is more than continuous models but decreased decoding speed overshadows time. Small size of PTM model makes it suitable for memory-limited mobile phone applications. Semi continuous models work well at 256 GMM. In comparison to other two models their performance is not much good but their decoding speed is very low which make it usable for real time environment.

8 Conclusions

Mobile phones have become an integral part of our daily life. Numerous applications are being developed to increase the usability of mobile phones. To develop speaker-independent ASR system for Punjabi mobile applications, different acoustic models with different pronunciation dictionaries at different Gaussians are evaluated in this paper. It can be concluded that character-based dictionary is the best fit for the Punjabi while phonetically-tied acoustic model gives optimal performance for different accuracy and reliability parameters. So, a phonetically-tied model with character-based dictionary can be used for development of speaker-independent ASR system for Punjabi based mobile phone applications.

References

1. Bharali, S.S., Kalita, S.K.: A comparative study of different features for isolated spoken word recognition using HMM with reference to Assamese language. Int. J. Speech Technol. **18**(4), 673–684 (2015)
2. Commissioner for Linguistic Minorities, Ministry of Minority Affairs, Government of India. 50th Report of the Commissioner for Linguistic Minorities in India. http://www.nclm.nic.in/shared/linkimages/NCLM50thReport.pdf. Accessed 14 Jul 2016
3. Das, B., Mandal, S., Mitra, P.: Bengali speech corpus for continuous automatic speech recognition system. In: International Conference on Speech Database and Assessments Proceedings, Taiwan, pp. 51–55 (2011)
4. Davis, K.H., Biddulph, R., Balashek, S.: Automatic recognition of spoken digits. J. Acoust. Soc. America **24**, 637–642 (1952)
5. Dua, M., Aggarwal, R.K., Kadyan, V., Dua, S.: Punjabi automatic speech recognition using HTK. Int. J. Comput. Sci. **9**(4), 359–364 (2012)
6. Ho, T.H., Liu, C.J., Sun, H.: Phonetic State Tied-Mixture tone modeling for large vocabulary continuous mandarin speech recognition. In: Sixth European Conference on Speech Communication and Technology Proceedings, Hungary, pp. 883–886 (1999)
7. Huang, X., Alleva, F., Hon, H.W., Hwang, M.Y., Rosenfeld, R.: The SPHINX-II speech system: an overview. Comput. Speech Lang. **7**(2), 137–148 (1993)
8. Huggins-Daines, D., Kumar, M., Chan, A.: Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices. In: International Conference on Acoustics, Speech and Signal Processing Proceedings, pp. I-185–I-188. IEEE, Toulouse (2006)
9. Khaira, S.S.: Punjabi Bhasha Viyakarn Ate Bantar (Punjabi). Punjabi University, Patiala (2011)
10. Klatt, D.H.: Review of the ARPA speech understanding project. J. Acoust. Soc. America **62** (6), 1345–1366 (1977)
11. Kumar, K., Aggarwal, R.K.: A Hindi speech recognition system for connected words using HTK. Int. J. Comput. Sys. Eng. **1**(1), 25–32 (2012)
12. Kumar, R.: Comparison of HMM and DTW for Isolated Word Recognition System of Punjabi Language. In: 15th Iberoamerican Congress on Pattern Recognition Proceedings, SP, Brazil, pp. 244–252 (2010)
13. Kumar, Y., Singh, N.: An automatic spontaneous live speech recognition system for Punjabi Language corpus. Int. J. CTA **9**(20), 9575–9595 (2016)
14. Kumar, Y., Singh, N.: An automatic speech recognition system for spontaneous Punjabi speech corpus. Int. J. Speech Technol. **20**(2), 297–303 (2017)
15. Lee, K.F., Hon, H.W., Reddy, R.: An overview of the SPHINX speech recognition system. IEEE Trans. Acoust. Speech Signal Process. **38**(1), 35–45 (1990)
16. Lowerre, B.T.: The Harpy Speech Recognition System. Dissertation, CMU (1976)
17. Mittal, P., Singh, N.: Speech based command and control system for mobile phones: issues and challenges. In: International Conference on Computational intelligence and communication technology Proceedings, pp. 729–732. IEEE, Ghaziabad (2016)
18. Naing, H.M.S., Hlaing, A.M., Pa, W.P.: A Myanmar large vocabulary continuous speech recognition system. In: APSIPA Annual Summit and Conference Proceedings, Hong Kong, pp. 320–327 (2015)
19. Placeway, P., Chen, S., Eskenazi, M.: The 1996 HUB-4 Sphinx-3 system, In: DARPA Speech Recognition Workshop Chantilly Proceedings (1996). http://www.itl.nist.gov/iad/mig/publications/proceedings/darpa97/pdf/placewa1.pdf. Accessed 09 Sept 2016

20. Punjab Population Census data. http://www.census2011.co.in/census/state/punjab.html. Accessed 14 Jul 2016
21. Punjabi Language, Encyclopedia Britannica Online. https://www.britannica.com/topic/Punjabi-language. Accessed 05 Jul 2016
22. Satori, H., ElHaoussi, F.: Investigation Amazigh speech recognition using CMU tools. Int. J. Speech Technol. **17**, 235–243 (2014)
23. Schalkwyk, J., Beeferman, D., Beaufays, F.: Google search by voice: a case study. In: Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics Proceedings, pp. 61–90. Springer (2010)
24. Schuster, M., Nakajima, K.: Japanese and Korean voice search. In: International Conference on Acoustics, Speech, and Signal Processing Proceedings, pp. 5149–5152. IEEE, Kyoto (2012)
25. Thangarajan, R., Natarajan, A.M., Selvam, M.: Syllable modeling in continuous speech recognition for Tamil language. Int. J. Speech Technol. **12**, 47–57 (2009)
26. Walha, R., Drira, F., El-Abed, H., Alimi, A.M.: On developing an automatic speech recognition system for standard Arabic language. Int. J. Electr. Comput. Energ. Electron. Commun. Eng. **6**(10), 1138–1143 (2012)
27. Wang, H.M., Ho, T.H., Yang, R.C.: Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data. IEEE Trans. Speech Audio Process. **5**(2), 195–200 (1997)