International Journal of Speech Technology (2019) 22:219–230
https://doi.org/10.1007/s10772-019-09593-x

Development and analysis of Punjabi ASR system for mobile phones under different acoustic models

Puneet Mittal[1] · **Navdeep Singh**[2]

Received: 7 June 2018 / Accepted: 14 January 2019 / Published online: 28 January 2019

Abstract
Speech technology is widely gaining importance in our daily life. Speech based mobile phone applications are becoming popular in masses due to their usability and ease of access. Speech technology is helping people, with disabilities like blindness and physical abnormalities, to access and control mobile phone applications through voice, without using keypad or touchpad. Punjabi is one of the widely spoken language in various parts of the world. In this paper, an automatic speech recognition (ASR) system for mobile phone applications in Punjabi has been proposed and implemented for four different acoustic models- context independent, context dependent untied, context dependent tied, and context dependent deleted interpolation models. The proposed ASR is evaluated at 4, 16, 32 and 64 GMMs for performance analysis in terms of parameters like accuracy, word error rate and storage space required. It is observed that context dependent untied models outperform others by having better accuracy and lower word error rate, while context independent models require less storage space than others. The choice of fruitful acoustic model depends upon the available storage space as well as desired recognition accuracy. Mobile phones having limited resources may use context independent models, while context dependent untied models can be used to develop ASR system for high end mobile phones.

Keywords Acoustic model · ASR · Context dependent · Context independent · HMM · Speech recognition

1 Introduction

Speech is the natural and easiest communication medium for human-to-human interaction in our day-to-day life. Handling electronic devices like computers, laptops, tablets, and mobile phones through speech is quite challenging and intricate. Automatic speech recognition (ASR) on a machine requires conversion of spoken dialogues into text. With the developmental increase in ASR technology, people with disabilities like blindness and physical abnormalities can easily access and control mobile phone functions and applications. The execution of correct command depends upon the recognition capabilities of ASR system in use. If the system remains unable to correctly recognize words spoken by the user, no action or incorrect action may be taken. So, a reliable ASR system is desirable.

In recent years, demand for developing efficient user interface providing speech recognition is growing rapidly due to increased usability of mobile phones (Taylor 2010). It has diverse applications like voice dialing, short messaging service (SMS), call routing (hands-free communication), web browsing, home automation, voice search, speech-to-text processing, providing hands- and eyes-free communication, and controlling various devices by voice. In mobile phones, several technological challenges like ambient environmental noise, limited availability of hardware platforms, side language coverage requirements and cost limitations are being faced. The mobile ASR system needs to be ideal to be accepted by users. As mobile phones are eight times slower than normal desktop systems ("Why your smartphone won't be your next PC | Digital Trends" n.d.), the processes followed and applications developed for desktop systems cannot be directly used on mobile phones. It requires development of efficient ASR systems for mobile phones, while

✉ Puneet Mittal
pmittal.cse@gmail.com

Navdeep Singh
navdeep_jaggi@yahoo.com

[1] Department of CSE, BBSBEC, Fatehgarh Sahib, Punjab, India

[2] Department of Computer Science, Mata Gujri College, Fatehgarh Sahib, Punjab, India

keeping in mind various challenges related to their limited capabilities and resources. The acoustic phonetic context of speech unit needs to be modeled carefully to build an acoustic model. The phonetic context can be independent, as in case of monophone modeling; or it can be dependent, as in case of Triphone modeling. Effect of context has been shown by various researchers (Thalengala and Shama 2016; Thangarajan et al. 2009).

ASR systems for various European languages like English, French, Spanish and German (Adda-Decker et al. 1999; Ferreiros and Pardo 1999; Radeck-Arneth et al. 2015; Yang et al. 2011) have been well explored, while Asian and African languages have not received much attention of researchers (Besacier et al. 2014; Satori and ElHaoussi 2014). This is primarily due to their limited scope and unavailability of speech and text corpora for these languages. With the increase in population and expansion of mobile markets, focus of researchers has shifted towards low resource Indian languages like Hindi, Punjabi, and Gujarati (Aggarwal and Dave 2011; Hasnat et al. n.d.; Patel and Virparia 2011; Sarma et al. 2017).

Punjabi, one of the popular Asian languages having more than 100 million speakers worldwide (Shackle n.d.), is based on the principle of 'one sound - one symbol'. Being phonetic in nature, it is having one to one correspondence between spoken utterances and written symbols like consonants and vowels. This is something unlike English, where British pronunciation and American pronunciation are quite different ("Pronunciation guide for English and Academic English Dictionaries at OxfordLearnersDictionaries.com" n.d.). Each language has distinct phonetic structure and sounds that constrains the arrangements in which phonemes may be combined to form words. In Punjabi, words are pronounced in same way as they are written. Based upon motivations from growing mobile market and need of user-friendly voice interface for Punjabi language users, this work focuses on development of Punjabi ASR system for mobile phone applications.

This paper describes procedures for acoustic modeling of Punjabi Speech. Number of experiments have been attempted to account for some contextual effects of our models. Context Independent and various Context Dependent (Untied, Tied and deleted interpolation) models have been studied in this paper. An attempt has been done to examine the effect of acoustic phonetic context of a speech unit on its acoustic realization for Punjabi language, and find the most suitable Punjabi based acoustic model for mobile phones.

Rest of the paper is organized as follows: Brief review of literature is given in Sect. 2. Section 3 presents development of Punjabi ASR system followed by training and testing of proposed ASR in Sect. 4. The results are given in Sect. 5, while Sect. 6 gives concluding remarks for the presented work.

2 Literature review

Researchers have been working on developing speech applications for mobile phones and other embedded devices. Schmitt et al. (2008) laid emphasis on the constraints and limitations ASR applications are confronted with, under different architectures. Ruan et al. (2016) compared text entry and speech based dictation on mobile phones. They found that speech based dictation is 3 times faster than text entry. Nkosi et al. (n.d.) described the morphological driven approach to the creation of comprehensive and broadly representative Northern Sotho pronunciation dictionary for ASR. They developed the dictionary using dictMaker tool, that gives pronunciation of words. If pronunciation is not correct, the developed dictionary specifies the correct pronunciation. They developed an acoustic model using HTK having word accuracy claim of 63.9%. Thalengala and Shama (2016) developed Isolated ASR for Kannada. They have built two types of dictionaries: (a) phone level and (b) syllable level. Kannada news database is used for building pronunciation dictionaries. They obtained overall word recognition accuracy of 60.2% and 74.35% respectively for monophone and triphone acoustic models. They concluded that the performance of an ASR system may be improved by choosing a suitable acoustic model depending upon the vocabulary size. Beulen et al. (1997) presented state tying for context dependent phoneme models. They proposed decision tree based state tying on the VERBMOBIL corpus. They concluded that the gain due to state tying is lower than on the WSJ task as the context dependency of phones in the German language is not as high as in the English language. This highlights that context dependent models may not be suitable for every language. Each language has distinctive features that need to be modeled carefully to generate optimal acoustic phonetic model.

Walha et al. (2012) trained ASR for 10 digits for Standard Arabic (SA). Accuracy of 98.62% has been determined using the SA connected-digits corpus and 94.02% using the continuous SA speech corpus.

Lučić et al. (2015) analyzed the performance of audio games designed for visually impaired pupils to realize the impact of speech based applications. It was found that these applications are quite helpful for the pupils in understanding the world and learning to perform everyday tasks. They observed that the applications should be designed keeping in mind different disability levels and age groups. The initial success of introducing the ASR and TTS technologies to the visually impaired pupils resulted in motivation for further research in this field. Dua et al. (2012), and Kumar and Singh (2017) have worked on Punjabi Speech Recognition having primarily focused application development for desktop computers. The literature gives directions

towards the development of ASR system for resource limited mobile devices for less explored languages like Punjabi.

3 Development of Punjabi ASR system

The development of proposed ASR system is discussed in this section. Initially, Punjabi language consonants and vowels, their phonetic transcription, and classification based on manner of articulation is given in subsection 3.1. Further, speech and text corpus preparation for Punjabi language are outlined in subsection 3.2, having technical details of recordings for training and testing of proposed ASR system. Finally, four different acoustic models and their working is given in subsection 3.3 to finalize possible set of alternatives for the proposed system.

3.1 Punjabi language

As per USA's Central Intelligent Agency (CIA) ("The World Factbook—Central Intelligence Agency" n.d.), Punjabi language is widely spoken in countries like Pakistan, India and Canada having 48%, 2.8% and 1.4% speakers out of total population respectively. With more than 100 million speakers worldwide, it is the highest spoken language in Pakistan, one of the official languages in India, and third language in Canadian parliament. Punjabi is a member of Indo-Aryan branch of the Indo-European language family, evolved from Sanskrit through Prakrit. It is influenced by Persian and Arabic languages having many Persian and Arabic words ("Persian Influence on Punjabi (Shahmukhi and Gurumukhi) Language | Universal Urdu Post" n.d.). There is a little influence of English on Punjabi too. Being one of the twenty-two official languages recognized by the constitution of India, it is the official working language of Punjab, an Indian state.

3.1.1 Punjabi phonetics

In India, the Gurmukhi script is used to write Punjabi ("History of Punjabi Language & Gurmukhi Alphabet | Trumbull, CT Patch" n.d.). It is a Brahmic script derived from the Laṇḍā script. Gurmukhi script is alpha syllabary in nature. In Gurmukhi there are thirty-eight consonants called akhar, 10 vowel symbols called laga matra, two nasal sound symbols (bindi and tippi) and one symbol adhak to duplicate the sound of consonant. Any word in Punjabi is formulated by using the appropriate combination of Punjabi symbols. These symbols can also be referred as phonemes of Punjabi. The consonants are classified into plosives, velars, nasal, lateral, trill, flap and fricatives ("Punjabi/Phonetics–Wikibooks, open books for an open world" n.d.) based on their manner of articulation. Table 1 presents the list of Punjabi consonants. The phonetic transcription of the consonants ਮ, ਪ, ਬ, ਫ਼, ਨ, ਸ, ਜ਼, ਲ, ਕ and ਗ is same as that of English consonants *m, p, b, f, n, s, z, l, k, g*. Sounds like, ɖ (ੜ) as '*rd*' in guard, ʃ (ਸ਼) as '*sh*' in shoe, tʃ (ਚ) as '*tch*' in catch, are similar to English sounds but phonetic representation is given by a special symbol. [t̪, t̪^{h}, ʈ, ʈ^{h}] sound same as /t/ and [d̪, ɖ] sound like /d/ in English. But in Punjabi these are distinct sounds. [t̪, d̪] are sounds with tongue touching the teeth. [t̪^{h}] sound is like (th) sound in English, ʈ is pronounced like /t /in total and ɖ as /d/ in guard.

Vowels in Punjabi are used in two forms: Independent and dependent. Independent vowels occur independently in a word, but dependent vowels are used along with the consonant to give it appropriate sound. Punjabi has ten independent vowels: (ਆ).ɑː,(ਏ)eː, (ਐ)ɛː,(ਅ) ə, (ਈ) iː, (ਇ) ɪ, (ਓ) oː, (ਔ) ɔː, (ਊ) uː and (ਉ) ʊ. When a vowel is used along with the consonant, the sound of vowel is incorporated in consonant, such vowel is called dependent vowel. It is represented as (ਪਾ).pɑː,(ਪੇ)peː, (ਪੈ)pɛː,(ਪ) pə, (ਪੀ) piː, (ਪਿ) pɪ, (ਪੋ) poː, (ਪੌ) pɔː, (ਪੂ) puː, (ਪੁ) pʊ. Table 2 presents position of these vowels in articulation.

Punjabi and English vowels are quite different. The vowel ɑː is pronounced as 'a' in car, eː as 'a' in pale, ɛː as 'e' in bell

Table 1 Classification of consonants based on manner of articulation

Consonants							
		Labial	Dental/ Alveolar	Retroflex	Palatal	Velar	Glottal
Nasal		m ਮ	n ਨ	ɳ ਣ	ɲ ਞ	ŋ ਙ	–
Stop/Affricate	Tenuis	p ਪ	t̪ ਤ	ʈ ਟ	ʧ ਚ	k ਕ	–
	Aspirated	p^{h} ਫ	t̪^{h} ਥ	ʈ^{h} ਠ	ʧ^{h} ਛ	k^{h} ਖ	–
	Voiced	b ਬ	d̪ ਦ	ɖ ਡ	ʤ ਜ	g ਗ	–
Fricative	Voiceless	f ਫ਼	s ਸ	–	ʃ ਸ਼	(x ਖ਼)	–
	Voiced	–	z ਜ਼	–	–	(ɣ ਗ਼)	–
Flap		–	ɾ ਰ	ɽ ੜ	–	–	–
Approximant		ʋ ਵ	l ਲ	ɻ ਲ਼	j ਯ	–	ɦ ਹ

Table 2 Position of vowels in articulation

	Front	Central	Back
High	iː ɪ	–	ʊ uː
High-mid	eː	–	oː
Mid	–	ə	–
Low-mid	ɛː	–	ɔː
Low	–	ɑː	–

,ə as 'u' in hut, iː as 'ee' in meet, ɪ as 'i' in dip, oː as 'o' in bold, ɔː as 'aw' in saw ,uː as 'oo' in boot, ʊ as 'oo' in book.

Three auxiliary symbols ◌ਂ (bindī), ◌ੰ (ṭippī), ◌ੱ (adhak) do not appear independently, and are used along with vowels. The symbol ◌ਂ (bindī) adds nasal sound to a particular vowel, and is used with (ਆਂ).ɑ̃ː,(ਏ)e̯ː, (ਐ)ɛ̃ː, (ਈਂ) ĩː, (ਓਂ) oː, (ਔਂ) ɔ̃ː. Its sound is same as 'n' in band(ਬੈਂਡ), sand, bang, rang. ◌ੰ (ṭippī) also adds nasal sound to a particular vowel and is used with (ਪੰ) pə̃,, (ਪਾਂ) pɑ̃ː, (ਪੂੰ) puː, (ਪੁੰ) põ. Its sound is same as 'n' in brunch, lunch. ◌ੱ (adhak) doubles the sound of the letter. It is placed above the letter whose sound is to be doubled and indicates that the following consonant is geminate. Word without ◌ੱ(adhak) will lead to totally different meaning of word. For example ਦਸ [d̪əs]—'ten'; ਦੱਸ [d̪əsː]—'tell' (verb), ਪਤਾ [pə̪ta]—'aware of something'; ਪੱਤਾ [pə̪tːa]—'leaf'.

3.2 Data collection–speech and text corpus preparation

Development of ASR system requires collection of good quality speech and text data for training and testing. The speech and text corpus considered for this work is kept small in size to make it suitable for mobile phones with limited memory.

3.2.1 Speech corpus

Speech corpus is the collection of speech data (voice recordings) from different speakers. Data required for building the acoustic model is collected using Mobile phone app SmartVoiceRecorder ("Smart Voice Recorder for Android - Download" n.d.) in environment with normal noise conditions, and is stored in .wav file format having sampling rate of 16 KHz and 16 bit mono. Speech enhancement may be done with DOAE techniques for recordings in noisy environment (Dey and Ashour 2018). Table 3 shows technical details of recordings done.

Table 3 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	91 Unique Punjabi Words (Total 1275 words)
Number of speakers	48
Age group of speakers	18–35 years
Average recording time per speaker	8.2 min per speaker ~ 6.34 h for all speakers
Noise conditions	Normal room environment
Window type	Hamming, 25.6 ms
Frames overlap	10 ms

The recordings are made from 48 speakers having total duration of 6.34 h. For single speaker, the text corpus consists of 1275 words, out of which 91 words are distinct. It contains 38 sentences (messages), 2 commands and 10 digits. 80% of speech database is used for training and rest 20% is used for testing. Total of about 5.1 h of speech recordings are used for training and rest for testing. The division of speech database in training and testing is done arbitrarily and is unbiased.

3.2.2 Pronunciation dictionary

The Pronunciation dictionary used in this study consists of 91 words. The dictionary is not built by any linguistic professional. As Punjabi is based on the principle of 'one symbol one sound' and Punjabi writing system is similar to how it is pronounced, so, dictionary is developed considering pronunciation of individual words. The algorithm for creating pronunciation dictionary is given below-.

Algorithm for character level representation of word in Pronunciation Dictionary

Character(w, length, c, loc)

(w refers to Punjabi word, c is character array to store each Punjabi character in word and loc is location)

1. For each w, do step 2 and 3
2. for loc=1 to w.length
3. Store each character of w in character array, c[loc]
4. Exit

Table 4 Excerpt of Punjabi dictionaries considered

Punjabi Word	Punjabi word in English	English meaning	Character/symbol based Phonetic Representation
ਮੈਨੂੰ	Mainu	Me	ਮ ੈ ਨ ੂ ੰ
ਜਾਵੇਗੀ	Jaawegi	Will be	ਜ ਾ ਵ ੇ ਗ ੀ
ਦਿਵਾਲੀ	Diwali	Festival of lights	ਦ ਿ ਵ ਾ ਲ ੀ
ਗੁਰਪੂਰਬ	Gurpurab	Birth anniversary of Guru	ਗ ੁ ਰ ਪ ੂ ਰ ਬ
ਆਵਾਂਗਾ	Aawanga	Will come	ਆ ਵ ਾ ਂ ਗ ਾ
ਇੰਤਜ਼ਾਰ	Intazaar	Wait	ਇ ੰ ਤ ਜ਼ ਾ ਰ
ਵਧਾਈਆਂ	Vadhaaian	Congratulations	ਵ ਧ ਾ ਈ ਆ ਂ

Example: Consider a Punjabi word, ਤੁਸੀਂ (you) having length 5, where length of a word is the number of characters present in it. Character array will be of length 5 and stored as ਤ,ੁ , ਸ, ੀ , ਂ . Here, ਤ and ਸ are consonants, ੁ and ੀ are dependent vowels and ਂ is an auxiliary sign, bindi. 47 distinct phones are used in this dictionary. Table 4 represents excerpt of two Punjabi dictionaries developed using above algorithm.

3.2.3 Feature extraction and analysis

Feature extraction is one of the most significant steps in speech recognition. Mel Frequency Cepstral Coefficients (MFCC), one of the best known feature extraction techniques, is based on the perception behavior of human ear. Parametric and acoustic–phonetic features are extracted from the speech signal. To improve the signal, pre processing is done by removing the unwanted and superfluous data from it. Speech processing engine uses this preprocessed speech feature data for further processing. This acoustic feature consists of 13 dimensional MFCC with window size of 25 ms and frame shift of 10 ms.

3.3 Acoustic model

There are varieties of Acoustic models like continuous (Bahl et al. 1983), semi-continuous (Huang and Jack 1988, 1990) and phonetically tied model (PTM) (Liu and Fung 2004). These acoustic models can be differentiated based upon how Gaussian Mixture is built. In continuous models, number of Gaussian mixture is about 150,000 which is too much for computation and such models are difficult to be handled for mobile phones. Semi-continuous models have only 700 Gaussians and are quite fast but they are less accurate. Phonetically Tied models use 5000 Gaussians and have accuracy akin to that of continuous models ("Acoustic Model Types—CMUSphinx Open Source Speech Recognition" n.d.). They are faster than continuous models thus making it apposite for the mobile phones. Being focused on development of proposed ASR system for mobile phone applications, only PTM based acoustic model development is worked upon.

In continuous speech, correct recognition of spoken words is a major challenge. It is quite difficult to distinguish between spoken words, as there are no visible boundaries present. Moreover, there is a lot of variability in speech due to dissimilar gender - male or female, age group - child, adult or old person, environment- noisy or clean, and various other factors. Various statistical approaches are available to build ASR models, irrespective of such variability. Hidden Markov Model (HMM) is one such approach, based on probabilistic sequence model, which computes a probability distribution over possible sequences of labels and choose the best label sequence, from a given sequence of units (words, letters, morphemes, sentences). In HMM, output observation is produced using output probability function of speech feature vectors, which are modelled with mixtures of Gaussian distributions, where the sequence of states is hidden. It uses Baum-Welch algorithm (Baum et al. 1970), also called Forward–Backward algorithm, to compute the maximum likelihood of speech signal feature vector. It is based on expectation–maximization (EM) algorithm (Dempster et al. 1977), which is an iterative method to find maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. Viterbi algorithm, a dynamic programming algorithm, is used for searching the optimal path through HMM model. It finds the most likely sequence of hidden states, called the Viterbi path, and provides a sequence of observed events.

Each phone in acoustic model can be modeled in two ways: (i) Context Independent model, and (ii) Context dependent model. These models are discussed below:

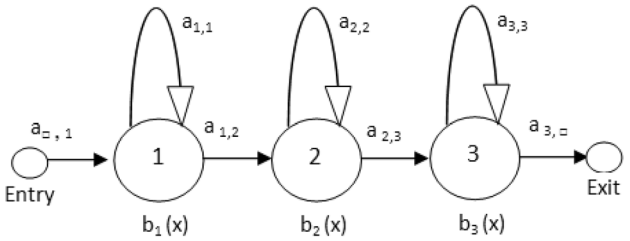

Fig. 1 3-state HMM model

3.3.1 Context independent models

It uses Context Independent (CI) modeling, also called Monophone modeling, which considers the individual phone without taking into account its left and right context. The model is developed for each phone independently, and individual HMM is built for each phone of the language. In this work, the basic 3-state HMM model (Fig. 1) is used for Punjabi phones. It has one state for transition into the phone, one for the center part and one for the transition out of the phone, and all HMM units are joined together in speech recognition system.

3.3.2 Context dependent models

It uses Context Dependent (CD) modeling, also referred as Triphone modeling, which considers that there are no well defined boundaries between phonemes in speech. Triphone modeling takes into account the effect of left and right phone on the phone under consideration. So, the phones are likely to be influenced by the presence of adjacent phones. Two identical phones having different left and/or right phones are considered as different Triphones. The basic 3-state model, considered for CI model, is also considered for CD models. Each phone in the CD Phonetically Tied Models (PTM) is modeled with GMM (Beaufays et al. 1999),

$$p(X_k|\emptyset, i) = \sum_{g=1}^{N_\emptyset} P_g^i N_g\left(X_k\right) \tag{1}$$

where $\emptyset$. represents the phone being modeled, i. represents a specific context realization or Triphone cluster of $\emptyset$, P_g^i represents the gth. CD mixture weight in cluster i, and $N_g\left(X_k\right)$. represents the gth CI Gaussian distribution evaluated for the observation X_k respectively. Three types of models built under CD modeling–Untied, Tied and Deleted Interpolation, are discussed below:

- **Untied models** : The HMMs are trained forll CD phones that appear in training data, while traing CD-ed models. In CD untied model, a separate model is built for each and every occurrence of Triphone. It results in large number of parameters. Therefore, it requires large amount of hardware resources while modeling CD phones (ex. triphones) with untied states. Untied HMM states are shown in Fig. 2.
- **Tied models**: In CD tied models, data from similar HMM states is collected together and used to train one global state called a "senone". Many groups of similar

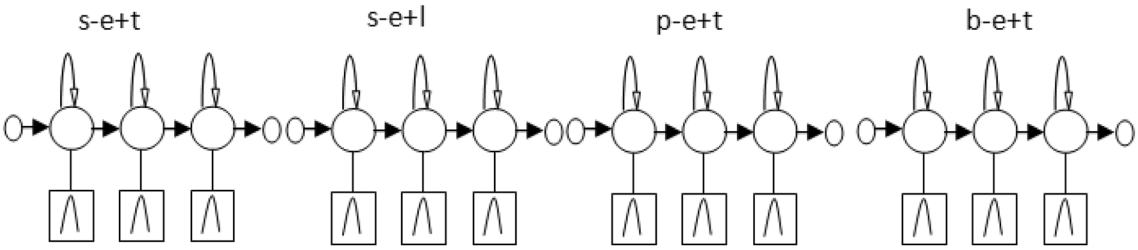

Fig. 2 CD untied states

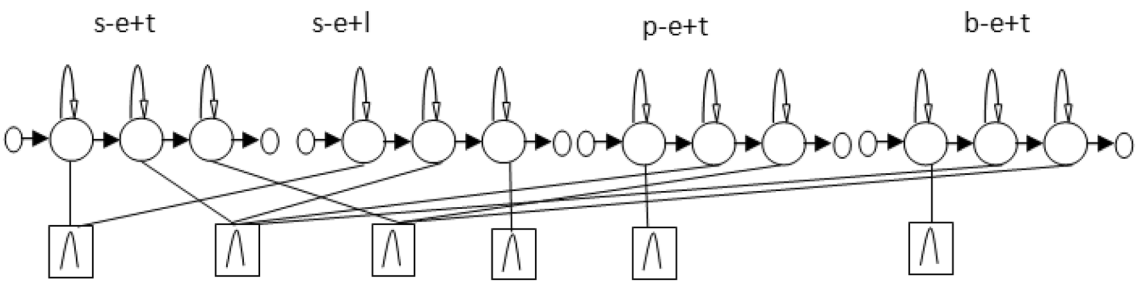

Fig. 3 CD tied states

states are formed, and the number of "senones" that are finally to be trained can be user defined. A senone is also called a tied-state and is shared across the triphones which contributed to it. As the states are being shared, it requires far less hardware resources for storage. Tied HMM states are shown in Fig. 3.

- **Deleted interpolation**: Deleted interpolation models (Huang et al. n.d.) improve the performance of acoustic models by reducing the effect of overfitting. It interpolates between CD and CI mixture weights iteratively. For estimating the interpolation factor, two balanced data sets are required. Data from one set estimates the interpolation factor between Context Independent and Context Dependent tied models. After that switching of two data sets is done and process is repeated iteratively till the new interpolation factors is close to the previous value. Interpolated Probability Density function ${}^{DI}_{P_i}(.)$ can be expressed as:

$$ {}^{DI}_{P_i}(.) = \lambda {}^{CD}_{P_i}(.) + (1-\lambda) {}^{CI}_{P_i}(.) \tag{2} $$

where ${}^{DI}_{P_i}(.)$ is the mixture function after deleted interpolation for CD Markov state i.; ${}^{CD}_{P_i}(.)$ is the corresponding CD mixture, and ${}^{CI}_{P_i}(.)$ is the corresponding CI mixture. The interpolation weight $\boldsymbol{\lambda}_i$ can be state-dependent, senone-dependent, or phone-dependent if we share interpolation weights at different levels.

4 Training and testing of proposed ASR

The proposed ASR system is trained and tested for 6.34 h of recordings from 48 speakers, out of which.1 h of recordings (80% of total recordings) are used for training while rest of the recordings (20% of total recordings) are used for testing. ASR system is trained using CMU- Sphinxtrain ("Training an acoustic model for CMUSphinx – CMUSphinx Open Source Speech Recognition" n.d.). It is highly reliable and well recognized platform used for development of the proposed system. It includes acoustic model training tools for building HMM based acoustic models.

Following steps are performed for Punjabi ASR system development: (i) Training of Context Independent models, (ii) Training of Context Dependent Untied models, (iii) Training of Context Dependent Tied models, and (iv) Training of Context Dependent Deleted Interpolation models. The training procedure is shown in Fig. 4.

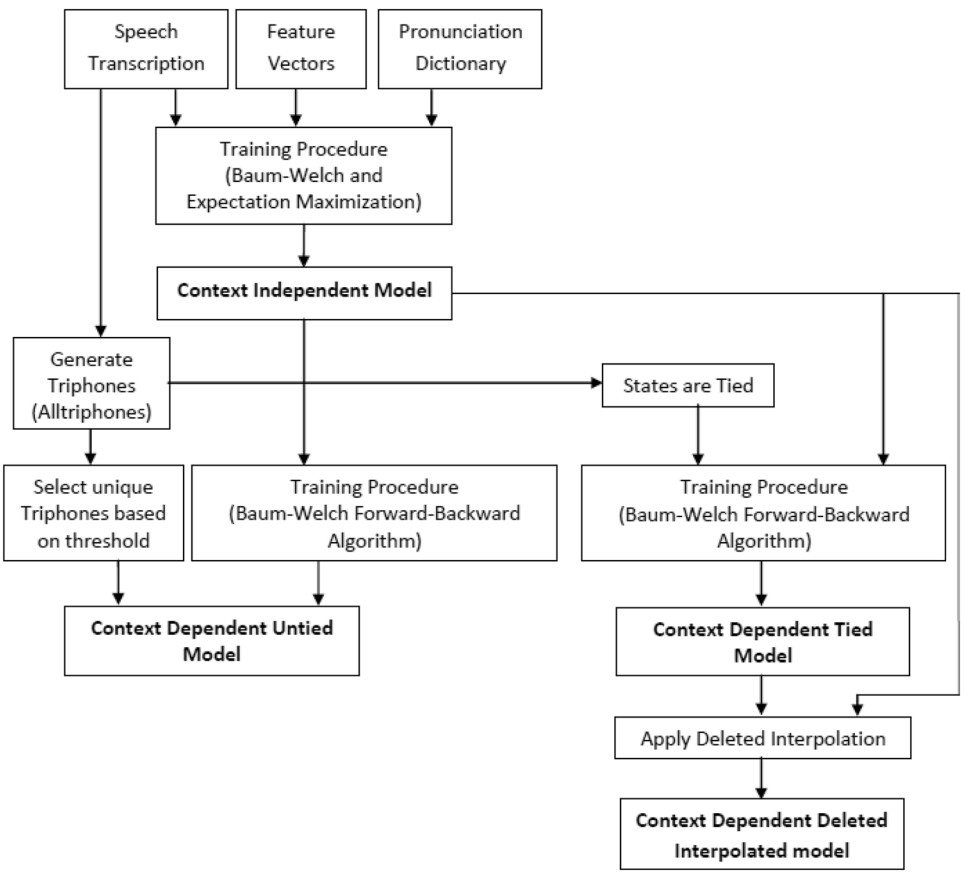

Fig. 4 Training of different acoustic models

4.1 Training of context independent model

The CI Model is built using speech transcription, feature vectors and pronunciation dictionary. The speech signal contains a lot of redundant data that is undesirable for signal processing. So, it is transformed into reduced set of features called feature vector. During training, first of all, the feature vectors are extracted from speech signal. Various features related to power, pitch, vocal tract information are extracted from the speech signal. Then, Baum Welch (Baum et al. 1970) and Expectation–Maximization (Dempster et al. 1977) algorithm are used to train the monophone based HMM model. This training procedure is repeated for 4,8,16, 32 and 64 Gaussian mixtures. In this work, 47 CI phones are generated.

4.2 Training of context dependent untied model

The CD Untied model is trained using CI model trained above. Here, the HMMs are trained for all context-dependent phones, triphones, in the training corpus. The Baum-Welch (BW) forward–backward algorithm is iteratively used for training the model. BW algorithm is executed on training data and for each iteration BW buffers are generated. A normal model definition file (mdef) contains all possible triphones from the current training dictionary. But CD untied mdef file contains only those triphones which are above the threshold. Thresholding is done to lower the number of triphones and it further condenses the size of the model. The number of unique triphones considered in this work are 645.

4.3 Training of context dependent tied models

To train the CD tied models data from all similar HMM states is collected. This is done by building the decision tree for each Triphone and triphones with same HMM states are combined. Each combined HMM state is called Senone and it is shared by all triphones. After this, decision trees are pruned to have number of leaves similar to number of senones. Mdef file contains all triphones seen and HMM states corresponding to these triphones identified with senones. After this, various states identified in triphone model definition file are tied. The training procedure is repeated for 4,8,16, 32 and 64 Gaussian mixtures.

4.4 Training of context dependent deleted interpolation model

The CI and CD Tied models developed in Sect. 4.1 and 4.3 respectively, are used to train the CD deleted interpolation model. Two balanced data sets are needed to develop the model. Even number of Baum-welch buffers are used to separate data into two different sets. The training procedure is repeated for 4,8,16, 32 and 64 Gaussian mixtures for context dependent models.

The monophones, triphones, and states considered in this work are given in Table 5. It can be observed that there are total 2768 HMM states during CD untied modelling, while 15960 states are considered for CD tied modelling as all triphones are taken up.

In tied model, the number of states is reduced to 341 after state tying. It indicates that the tied models need all phones to be modelled for only 341 states as compared to 2076 states in untied models. It results in generation of more mixture weights for untied modelling.

4.5 Testing of acoustic models

The acoustic models developed above are tested using speech corpus collected from different speakers. Out of total recordings (speech corpus), 20% data is used for testing the models. Total 20 acoustic models are developed, having 5 models for each model type. PocketSphinx Recognizer is used as decoder for testing purpose.

Table 5 Monophones, triphones and states

Category	Value
Number of monophones	47
Number of unique triphones (for CD_Untied modelling)	645
Total number of HMM states (emitting and non-emitting) (for CD_Untied modelling)	2768
Number of states of all phones (Untied states)	2076 ((645 + 47)*3)
Number of triphones (for CD_Tied modelling)	3943
Total number of HMM states (emitting and non-emitting) (for CD_Tied modelling)	15,960
Number of states of all phones (after state-sharing) (Tied States)	341

Table 6 WER for different acoustic models

GMM	CI	CD_Untied	CD_Tied	CD_DelInterp
4	27.8	27.2	25.9	28.5
8	22.0	23.0	22.4	25.3
16	21.7	22.8	22.2	25.8
32	21.0	20.2	21.7	25.0
64	19.1	18.8	20.5	23.9

5 Results and discussions

The performance of the recognizer is evaluated and analyzed for parameters like Word Error Rate (WER), accuracy, and size of model. Different parameters used for performance evaluation are given below:

- **Word Error Rate**: It specifies how many insertions, deletions and substitutions have been done by the recognizer. The WER is calculated as:

$$WER = (I + S + D)/N \tag{3}$$

where N is the total number of words in the test transcription, I is the number of Insertions, S is the number of Substitutions and D is the number of Deletions done by the recognizer. Lower values of WER indicate higher accuracy of the model.
- **Accuracy**: It is the degree to which the developed model correctly recognizes the speech inputs. It is measured as percentage of correctly recognized words over total spoken words.
- **Size**: It is the amount of memory required to store the built model on the hardware. It is measured in Megabytes of data being stored.

Following results are obtained from the implementation of proposed ASR system:

5.1 Word error rate comparison for different acoustic models

The performance in terms of WER for trained CI and CD acoustic models at different GMM (=4, 8, 16, 32, 64) is shown in Table 6. It can be analyzed that WER improves with the increase in GMM for all acoustic models considered in this work (Figs. 4, 5). CI models are having lower WER than CD models at lower Gaussians (=8, 16), while CD models are showing lower WER on higher Gaussians (=64). Under CD models, Tied state models are having lower WER than untied and deleted interpolation models at lower GMM, while untied models give better results on higher GMM.

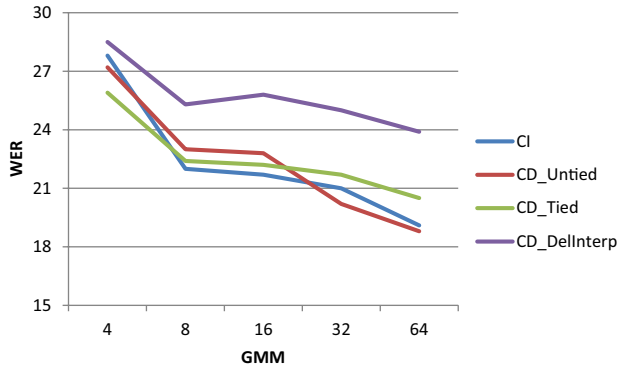

Fig. 5 Comparison of WER for different acoustic models

Table 7 Accuracy of different acoustic models

GMM	CI	CD_Untied	CD_Tied	CD_DelInterp
4	72.2	72.8	74.1	71.5
8	78	77	77.6	74.7
16	78.3	77.2	77.8	74.2
32	79	79.8	78.3	75
64	80.9	81.2	79.5	76.1

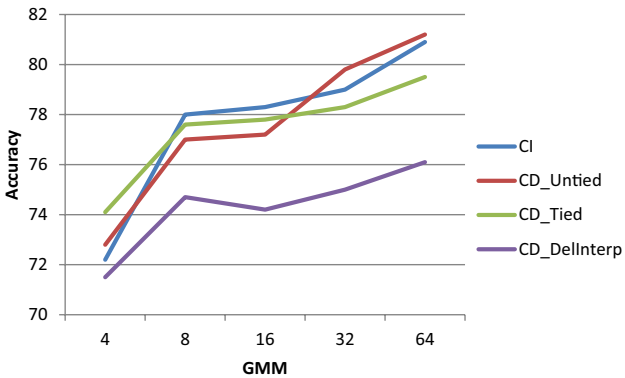

Fig. 6 Accuracy comparison of different acoustic models

5.2 Accuracy comparison for different acoustic models

The performance in terms of accuracy for trained CI and CD acoustic models at different GMM (=4, 8, 16, 32, 64) is shown in Table 7. The Deleted Interpolation models give lowest accuracy (Fig. 6) among all the models due to the negative effect of over-fitting, as iterative interpolation of mixture-weights fails to reduce the effect of over-fitting due to small vocabulary and training data used in this work. CD untied models are showing highest accuracy at 64 GMM.

Table 8 Size (in Mb) of different acoustic models

GMM	CI	CD_Untied	CD_Tied	CD_DelInterp
4	88.8	239	303	303
8	172	444	405	405
16	339	853	608	608
32	674	1669.12	1013.76	1013.76
64	1341.44	3307.52	1822.72	1822.72

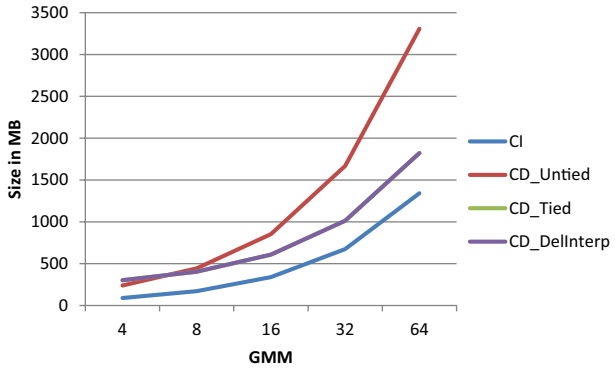

Fig. 7 Size comparison of different acoustic models

5.3 Size comparison for different acoustic models

Results obtained for size comparison are given in Table 8. The developed CD tied and CD deleted interpolation models are having same size, overlapping each other as shown in Fig. 7. The deleted interpolated models seem unsuitable for the proposed ASR system due to their high WER, low accuracy and large size. So, it is suggested that deleted interpolated models should not be used for developing ASR system having small vocabulary and training data. The size of CI models is smallest among all models for all GMM values due to less number of states to be trained, while size of CD Untied models is largest among all models due to large number of triphones being trained. As size and accuracy are relative to each other at 64 GMM, the model having better performance than others may be considered suitable for

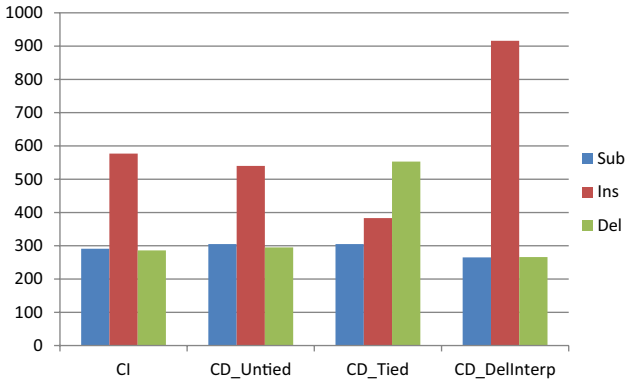

Fig. 8 WER comparison of different acoustic models at 64 GMM

development of ASR system in general. The model definition file(MDEF) generated for CD untied models is smaller in size than the MDEF file generated for tied models, as number of triphones considered in tied models are far more than untied models. But mixture weight file generated for untied models is relatively large in size than the file generated for tied models, as states are tied in CD tied models. It results in larger size of CD untied models.

5.4 Performance comparison of acoustic models at 64 GMM

Detailed performance of all models developed at 64 GMM is given in Table 9. Both CD untied and CI models show comparable performance. CD untied model recognizes maximum number of words correctly. Figure 8 shows comparison of various types of errors (substitution, insertion and deletion) in all acoustic models at 64 GMM. No. of substitution and Deletion errors are least in Deleted Interpolation model due to robustness of the model, but number of insertions are high.

According to theory, deleted interpolation models are smoothed models which improve the accuracy of model. From Table 9, we can analyze that this model is having

Table 9 Performance of different acoustic models at 64 GMM

Details	CI	CD_Untied	CD_Tied	CD_DelIntpl
Sentences in the Test set	975	975	975	975
Words in Test Set	6051	6051	6051	6051
Words correctly recognized	4898	4911	4810	4605
No. of errors and type	1153 (Sub: 291, Ins: 577, Del: 286)	1140 (Sub: 305, Ins: 540, Del: 295)	1241 (Sub: 305, Ins: 383, Del: 553)	1447 (Sub: 265, Ins: 916, Del: 266)
WER	19.1	18.8	20.5	23.9
Decoding speed	0.06×RT	0.06×RT	0.06×RT	0.07×RT

least substitution and deletion errors. CD tied models are showing least number of insertions but number of deletions are quite more. Decoding speed or average time taken by the acoustic model to recognize a word is same for all models except deleted interpolation, requiring more time.

6 Discussion

The Context Independent, Context Dependent Untied, CD Tied and CD Deleted Interpolated models are compared in terms of WER, accuracy and size. A common pronunciation dictionary is used for all models to give an unbiased view. In experiments, it is found that CD untied models have given higher accuracy and lower WER than all other models making it most promising model among four. In addition, results at higher GMM levels (i.e. at 64 GMM) seem to be better than lower levels. But, from required storage point of view, the CD untied models require much larger storage space as compared to CI models which have little less accuracy than the CD untied models. For applications which are resource constrained in terms of space, CD untied models may not prove to be a favorable choice, while CI models having very low storage requirement seems fruitful despite little less accuracy. At 64 GMMs, it is observed that CI models and CD Deleted Interpolation models have low substitution and deletion errors. On the other hand, CD deleted Interpolated models have very high insertion error. As substitution and deletion errors affect the perceived accuracy more than the insertion errors, the insertion errors may be ignored in CD deleted Interpolation models making them one of the suitable choice.

7 Conclusion

An automatic speech recognition system for mobile phone applications has been proposed and implemented for four different character-based acoustic models- context independent, context dependent tied, context dependent untied, and context dependent deleted interpolation, to find most suitable and fruitful alternative for the proposed system. Out of four acoustic models, context independent models require less space as compared to others, while having less accuracy. On the other hand, context dependent untied model gives better accuracy than others, while having more space requirements. The major limitation of context dependent untied model is the requirement of higher memory space, which may slower down the mobile phones having limited small memory. Context independent models are good choice if slight decrease in accuracy is acceptable. Small vocabulary has been considered in this case. Corpus size can be increased to accommodate more words but the size of models will increase. With high recognition accuracy at low processing and storage requirements, the developed ASR system may be utilized to build different mobile phone applications. The work can be extended by taking into consideration syllable-based and hybrid acoustic modeling. Further, morpheme based speech recognition can be investigated where pronunciation dictionary and language model are built based on morphemes.

References

Acoustic Model Types – CMUSphinx Open Source Speech Recognition. (n.d.). Retrieved March 16, 2018 from https://cmusphinx.github.io/wiki/acousticmodeltypes/.

Adda-Decker, M., Adda, G., Gauvain, J., & Lamel, L. (1999). Large vocabulary speech recognition in French. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)* (pp. 45–48 vol.1). IEEE. https://doi.org/10.1109/ICASSP.1999.758058.

Aggarwal, R. K., & Dave, M. (2011). *Discriminative techniques for hindi speech recognition system* (pp. 261–266). Berlin: Springer. https://doi.org/10.1007/978-3-642-19403-0_45.

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5*(2), 179–190. https://doi.org/10.1109/TPAMI.1983.4767370.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics, 41*(1), 164–171. https://doi.org/10.2307/2239727.

Beaufays, F., & Weintraub, M. & Yochai Konig. (1999). Discriminative mixture weight estimation for large Gaussian mixture models. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)* (pp. 337–340 vol.1). IEEE. https://doi.org/10.1109/ICASSP.1999.758131.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56*, 85–100. https://doi.org/10.1016/J.SPECOM.2013.07.008.

Beulen, K., Bransch, E., & Ney, H. (1997). State tying for context dependent phoneme models. In *European Conference on Speech Comnumicution and Technology* (pp. 1179–1182).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.

Dey, N., & Ashour, A. S. (2018). Sources localization and DOAE techniques of moving multiple sources. In *Direction of arrival estimation and localization of multi-speech sources* (pp. 23–34). Cham: Springer. https://doi.org/10.1007/978-3-319-73059-2.

Dey, N., & Ashour, A. S. (2018). Applied examples and applications of localization and tracking problem of multiple speech sources. In *Direction of arrival estimation and localization of multi-speech sources* (pp. 35–48). Cham: Springer. https://doi.org/10.1007/978-3-319-73059-2.

Dey, N., & Ashour, A. S. (2018). Challenges and future perspectives in speech-sources direction of arrival estimation and localization. In *Direction of arrival estimation and localization of multi-speech*

sources (pp. 49–52). Cham: Springer. https://doi.org/10.1007/978-3-319-73059-2.

Dua, M., Kadyan, V., Aggarwal, R. K., & Dua, S. (2012). Punjabi speech to text system for connected words. In *Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012)* (pp. 206–209). Institution of Engineering and Technology. https://doi.org/10.1049/cp.2012.2528.

Ferreiros, J., & Pardo, J. M. (1999). Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations. *Speech Communication, 29*(1), 65–76. https://doi.org/10.1016/S0167-6393(99)00013-8.

Hasnat, M. A., Mowla, J., & Khan, M. (n.d.). Isolated and continuous bangla speech recognition: implementation, performance and application perspective. Retrieved January 3, 2018 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.173.372&rep=rep1&type=pdf.

History of Punjabi Language & Gurmukhi Alphabet | Trumbull, CT Patch. (n.d.). Retrieved January 4, 2018 from https://patch.com/connecticut/trumbull/history-of-punjabi-language--gurmukhi-alphabet.

Huang, X. D., Hwang, M.-Y., Li, J., & Mahajan, M. (n.d.). Deleted interpolation and density sharing for continuous hidden Markov models. In 1996 *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (Vol. 2, pp. 885–888). IEEE. https://doi.org/10.1109/ICASSP.1996.543263.

Huang, X. D., & Jack, M. A. (1988). Hidden Markov modelling of speech based on a semicontinuous model. *Electronics Letters, 24*(1), 6–7.

Huang, X. D., & Jack, M. A. (1990). *Semi-continuous hidden Markov models for speech signals. Readings in speech recognition*. San Francisco: Morgan Kaufmann Publishers Inc. Retrieved January 4, 2018 from https://dl.acm.org/citation.cfm?id=108259.

Kumar, Y., & Singh, N. (2017). An automatic speech recognition system for spontaneous Punjabi speech corpus. *International Journal of Speech Technology, 20*(2), 297–303. https://doi.org/10.1007/s10772-017-9408-2.

Liu, Y., & Fung, P. (2004). State-dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition. *IEEE Transactions on Speech and Audio Processing, 12*(4), 351–364. https://doi.org/10.1109/TSA.2004.828638.

Lučić, B., Ostrogonac, S., Vujnović Sedlar, N., & Sečujski, M. (2015). Educational applications for blind and partially sighted pupils based on speech technologies for Serbian. *The Scientific World Journal. 2015*. https://doi.org/10.1155/2015/839252.

Nkosi, M., Manamela, M., & Gasela, N. (n.d.). Creating a pronunciation dictionary for automatic speech recognition -a morphological approach. Retrieved January 3, 2018 from http://www.satnac.org.za/proceedings/2011/papers/Network_Services/176.pdf.

Patel, H. N., & Virparia, P. V. (2011). A Small Vocabulary Speech Recognition for Gujarati. *International Journal of Advanced Research in Computer Science, 2*(1), 208–210.

Persian Influence on Punjabi (Shahmukhi and Gurumukhi) Language | Universal Urdu Post. (n.d.). Retrieved March 16, 2018 from http://universalurdupost.com/english-articles/12-01-2016/33581.

Pronunciation guide for English and Academic English Dictionaries at OxfordLearnersDictionaries.com. (n.d.). Retrieved March 16, 2018 from https://www.oxfordlearnersdictionaries.com/about/pronunciation_english.html.

Punjabi/Phonetics - Wikibooks, open books for an open world. (n.d.). Retrieved March 16, 2018 from https://en.wikibooks.org/wiki/Punjabi/Phonetics.

Radeck-Arneth, S., Milde, B., Lange, A., Gouvêa, E., Radomski, S., Mühlhäuser, M., & Biemann, C. (2015). *Open source german distant speech recognition: corpus and acoustic model* (pp. 480–488). Cham: Springer. https://doi.org/10.1007/978-3-319-24033-6_54.

Ruan, S., Wobbrock, J. O., Liou, K., Ng, A., & Landay, J. (2016). Speech is 3 × faster than typing for english and mandarin text entry on mobile devices. Retrieved January 3, 2018 from http://arxiv.org/abs/1608.07323.

Sarma, H., Saharia, N., & Sharma, U. (2017). Development and analysis of speech recognition systems for assamese language using HTK. *ACM Transactions on Asian and Low-Resource Language Information Processing, 17*(1), 1–14. https://doi.org/10.1145/3137055.

Satori, H., & ElHaoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. *International Journal of Speech Technology, 17*(3), 235–243. https://doi.org/10.1007/s10772-014-9223-y.

Schmitt, A., Zaykovskiy, D., & Minker, W. (2008). Speech recognition for mobile devices. *International Journal of Speech Technology, 11*(2), 63–72. https://doi.org/10.1007/s10772-009-9036-6.

Shackle, C. (n.d.). Punjabi language | Britannica.com. Retrieved March 16, 2018 from https://www.britannica.com/topic/Punjabi-language.

Smart Voice Recorder for Android - Download. (n.d.). Retrieved January 4, 2018 from https://smart-voice-recorder.en.softonic.com/android.

Taylor, S. (2010). "Striking a healthy balance": speech technology in the mobile ecosystem. In A. Neustein (Ed.), *Advances in speech recognition* (pp. 19–30). Boston: Springer US. https://doi.org/10.1007/978-1-4419-5951-5_2.

Thalengala, A., & Shama, K. (2016). Study of sub-word acoustical models for Kannada isolated word recognition system. *International Journal of Speech Technology, 19*(4), 817–826. https://doi.org/10.1007/s10772-016-9374-0.

Thangarajan, R., Natarajan, A. M., & Selvam, M. (2009). Syllable modeling in continuous speech recognition for Tamil language. *International Journal of Speech Technology, 12*, 47–57. https://doi.org/10.1007/s10772-009-9058-0.

The World Factbook — Central Intelligence Agency. (n.d.). Retrieved March 16, 2018 from https://www.cia.gov/library/publications/the-worldfactbook/fields/2098.html.

Training an acoustic model for CMUSphinx – CMUSphinx Open Source Speech Recognition. (n.d.). Retrieved March 16, 2018 from https://cmusphinx.github.io/wiki/tutorialam/.

Walha, R., Drira, F., El-Abed, H., and A. M. A (2012). On developing an automatic speech recognition system for standard arabic language. *International Journal of Electrical and Computer Engineering, 6*(10), 1138–1143.

Why your smartphone won't be your next PC | Digital Trends. (n.d.). Retrieved January 4, 2018 from https://www.digitaltrends.com/computing/why-your-smartphone-wont-be-your-next-pc/.

Yang, H., Oehlke, C., & Meinel, C. (2011). German Speech Recognition: A Solution for the Analysis and Processing of Lecture Recordings. In *2011 10th IEEE/ACIS International Conference on Computer and Information Science* (pp. 201–206). IEEE. https://doi.org/10.1109/ICIS.2011.38.