Journal of Speech Sciences 1(3):69-83.2013
Available at: hhtp://www.journalofspeechsciences.org

Phone based acoustic modeling for automatic speech recognition for punjabi language

GHAI, W.[1]; SINGH, N.[2]

[1]Department of Computer Science, Khalsa College (ASR) of Technology & Business Studies, Mohali

[2]Post Graduate Department of Computer Science, Mata-Gujri College, Fatehgarh Sahib

Abstract

Punjabi language is a tonal language belonging to an Indo-Aryan language family and has a number of speakers all around the world. Punjabi language has gained acceptability in the media & communication and therefore deserves to have a place in the growing field of automatic speech recognition which has been explored already for a number of other Indian and foreign languages successfully. Some work has been done in the field of isolated word speech recognition for Punjabi language, but only using whole word based acoustic models. A phone based approach has yet to be applied for Punjabi language speech recognition. This paper describes an automatic speech recognizer that recognizes isolated word speech and connected word speech using a triphone based acoustic model on the HTK 3.4.1 speech Engine and compares the performance with acoustic whole word model based ASR system. Word recognition accuracy of isolated word speech was 92.05% for acoustic whole word model based system and 97.14% for acoustic triphone model based system whereas word recognition accuracy of connected word speech was 87.75% for acoustic whole word model based system and 91.62% for acoustic triphone model based system.

Keywords: Acoustic Model, Phone, Triphones, Gemination, Speech Corpus, Pronunciation Dictionary.

1. Introduction

Speech is generated when vibrating vocal cords create puffs of air. These puffs result in air pressure variations and it is due to these variations that the sensation of hearing develops. Automatic speech recognition (1, 18) is a process of transforming a speech signal (Figure 1) to a text which closely matches the input speech signal. This technique is being used extensively in application areas such as: voice user interface, voice interactive response, enhancing social interactive capability of handicapped people, learning a foreign language etc.

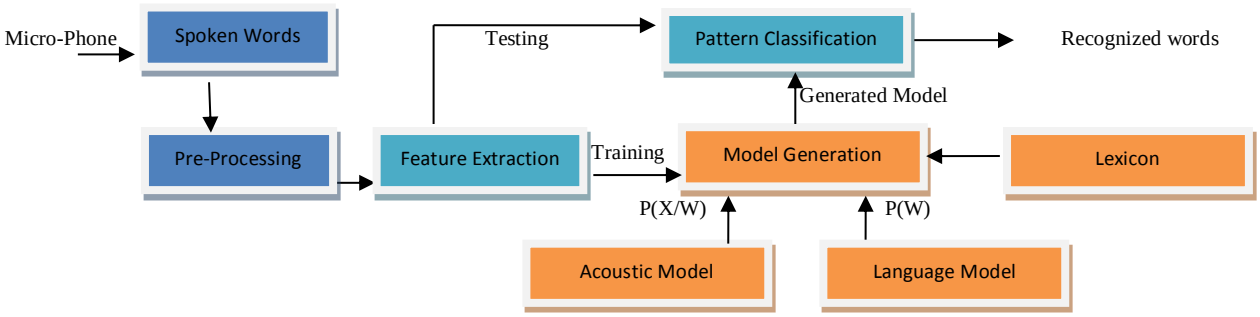

Figure 1: Automatic Speech Recognition System

1.1 Pre-Processing

The speech signal, acquired through a microphone, has to be pre-processed (1, 13) because in most applications, silence and background noise are undesirable. Pre-processing of the speech signal involves the following steps:

a. Background noise elimination: Fans, footsteps, opening and closing of doors, etc., are the real sources of background noise. Head mounted close speaking microphone helps in minimizing the background noise effect.

b. Pre-Emphasis Filtering: The signal is filtered using a simple high pass FIR filter H (z) to compensate for lip radiation and inherent attenuation of high frequencies in the sampling process. Here

$$H(z) = 1 - az^{-1} \qquad \text{where} \quad a \approx 1 \qquad 1.1$$

c. Framing: Here the pre-emphasized speech signal is blocked into frames of N samples. Adjacent frames are separated by M samples.

d. Windowing: Window means the portion of speech waveform to be processed. Each frame is multiplied by a window to reduce the edge effect of every frame segment. The Hamming window is common, and is represented mathematically by equation 1.2 as:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad \text{where} \quad 0 \le n \le N-1 \qquad 1.2$$

1.2 Feature Extraction

Feature extraction (1, 2, 14) is the process of extracting features such as power, pitch, and vocal tract configuration from the speech signal. There are two techniques for carrying out feature extraction: Temporal Analysis and Spectral Analysis. In case of temporal analysis, the speech waveform is analyzed directly, whereas in case of spectral analysis, features are computed from a Fourier transform of the signal.

1.3 Knowledge Models

An automatic speech recognition system requires the capability to know how the words sound. Knowledge models are meant for mapping a sound to a word/phrase. Acoustic models, language models and lexicon models are knowledge models. The acoustic model is used to represent the different ways a word of a particular language can sound. It makes use of audio recordings along with their transcriptions and compiles these two to produce statistical representations. Hidden Markov Models (HMM; 1, 2, 19) are one of the main techniques applied for acoustic modelling. Others are conditional random fields (22), neural networks and maximum entropy models. The language model provides the context information to a speech recognition system. It models the way the words are connected to form a sentence. For this purpose, it uses a vocabulary: a set of words to be recognized, and a grammar: set of rules for regulating the arrangement of words. Language model is essential for developing large vocabulary ASR. Acoustic and Language models are generated using training data. The lexicon model defines the mapping from vocabulary items to acoustic models, and may take the form of a pronunciation dictionary.

1.4 Pattern Classification

During the pattern classification stage, MFCC features of each test datum are compared with the model of each word learned during the training phase so that the best matched word can be given as the output of the speech recognition system. Mathematically, the pattern classification problem is expressed as:

$$P(W \mid X) = \frac{P(X \mid W).P(W)}{P(X)} \qquad 1.3$$

Accordingly, the classifier searches for a sequence of words W, given an acoustic observation X, so as to maximize the numerator of the equation 1.3. The prior probability of the word, i.e. P (W), is

provided by the language model, whereas the observation likelihood, i.e. P (X|W), is provided by the acoustic model.

2. Acoustic Phone Model

2.1 Need

Automatic speech recognition problem can be described as the comparison between a test pattern, in the form of a sequence of feature vectors and stored reference patterns. Acoustic word modelling attempts (1, 2, 3, 18, 19) to obtain a model for every word to be recognized. This type of modelling has a few limitations. First problem: Recognition can be done only for seen/known words. Second problem: As the vocabulary size increases, the need for storing a large number of reference patterns arises. Third problem: Pronunciation variations give rise to the need for a large number of reference patterns. As a result, the whole word approach becomes infeasible. Zipf's law (3) states that the number of events occuring often is small whereas, the number of events occuring very rarely is very large. This law is applicable to the field of speech recognition and, thereby, the frequency of a word is approximately inversely proportional to its rank in the frequency table. In other words, there are a small number of frequent words and a very large number of rare words. As a result, many words may not occur at all in a finite corpus of recorded speech.

2.2 Concept

Utterances of words in a language are usually composed of a set of sounds, i.e. phones, which may be considered as sub-word units. The words "ਸਾਡੇ" (saːɖeː; our), "ਕਰਦਾ" (krɖaː; does) and "ਤੁਸੀਂ" (t̪uːsiː; you) have individual sounds as ਸ (s), ਆ (aː), ਡ (ɖ), ਕ (k), ਰ (r), ਦ (ɖ̪), ਤ (t̪), ੁ (uː), ੀ (iː) and ੇ (eː). It is interesting that phones occurring for a few words can form new words i. e. words which were not considered earlier, e. g., "ਸਾਡਾ", "ਸਾਡੀ". In spite of the fact that the number of words in a language is very much greater than the number of phones, formation of more words through a limited number of phones is a useful phenomenon.

2.3 Models

A phone has three phases (19) in its pronunciation: on-glide, pure phone, off-glide. This fact leads to the need of at least three states in a phone HMM. For an acoustic phone model, the words are

segmented into sequence of phones. A monophone HMM models only one phone. A monophone is context-independent which means that it does not change depending on its neighbouring phones. It is found that, due to coarticulation effects (18), a phone in fluent speech is always affected strongly by its neighbouring phones and thereby it produces different sounds depending on the phonetic context. To handle coarticulation effects, i.e., to incorporate context dependence, biphones and triphones (3, 4) can be used. A biphone HMM models a phone as being dependent on its left or right context where as a triphone HMM models a phone as being dependent on its right as well as left context. The pronunciation of the current word is also affected by its neighbouring words in continuous speech, therefore cross-word triphone HMMs are used to represent the co-articulation effect between words. When HMMs are trained for phones instead of words, enough data are found automatically to train all HMMs. HMMs for words are developed by concatenating HMMs of the individual phonemes forming the word. Unseen words also can be dealt with by this type of acoustic modelling.

3. Punjabi Language

3.1 Introduction

The Punjabi language (5, 15, 16) is a widely spoken famous Indo-Aryan language of the Indo-Iranian subgroup of Indo-European family of languages. The Punjabi language has a rich cultural heritage. Punjabi language in India is written using the Gurumukhi (16) alphabet which is a descendent of the Lahnda alphabet. Guru Angad Dev is the 2nd Guru of the Sikh religion and Guru Angad Dev Ji standardized Gurumukhi script in the 16th century. In some States of northern India, Devanagari script is also being used to write Punjabi. Just like Devanagari orthography, the Gurumukhi script too follows an Abugida writing system. Punjabi with Gurumukhi script is one of the 22 languages that is officially recognized in India and more specifically, the first official language in Indian Punjab. In contrast, there is no official status for Punjabi language in Pakistan Punjab. Punjabi is the 10th most widely spoken language in the world. It was first recognized as an independent language in the 11th century. The increasing acceptability of the Punjabi language in media and communication encourages an accelerated pace of ASR research & development. With regard to linguistic typology, Punjabi is an inflecting language and its word order is SOV, i.e., Subject Object Verb. Another important aspect of the Punjabi language is that it is a tonal and fusion language because it has a tendency to fuse morphemes. There are around 109 million speakers (21) in India & Pakistan (Table 1) who speak Punjabi as their 1st language. Spoken Punjabi in India relies more heavily on Sanskrit vocabulary through Hindi and has many dialects such as Majhi,

Pwadhi, Potwari, Dhani, Hindko, Malwi etc. Majhi is a prestigeous dialect of the Punjabi language. In our work, training and test speech samples are drawn from the Pwadhi dialect.

3.2 Phonetic Inventory

Phonological features of the Punjabi Language have been explored here with regard to automatic speech recognition. Unlike other Indian languages, Punjabi is a tonal language and its phoneme inventory (5, 15) contains 10 vowels, 25 consonants, 7 diphthongs and three tones whose production has neither frication nor stoppage of air in the mouth.

3.2.1 Vowels

Punjabi vowels (Table 2) are ten in number. Punjabi Vowels exist in pairs, i.e., a short vowel and a long vowel. Vowels in Punjabi are classified as Independent & dependent vowels. A dependent vowel or "matra" requires a consonant for its support. Each consonant of the Punjabi language is followed by an inherent '**ਅ**' sound but that can be altered by using a dependent vowel. In some cases, dependent vowels can't be used at the beginning of the word/syllable. This problem has given rise to the evolution of independent vowels which are used when there is no consonant to which a matra can be attached.

Table 1: Comparative statement of Speakers

Country	Number of Speakers (in millions)	%age of population
India	33 approx	3.00
Pakistan	76 approx	44.00

3.2.2 Consonants

The alphabet of the Punjabi language is comprised of 35 distinct letters and 6 special consonants which have been formed by placing bindi (◌̣) at the foot of consonants. Subscripted bindi of this kind is usually used to create a fricative from the closest corresponding stop, thus, subscripted bindi is used to distinguish the orthographic pairs s(ਸ)/ʃ(ਸ਼), p^h(ਫ)/f(ਫ਼), dʒ(ਜ)/z(ਜ਼), k^h(ਖ)/x(ਖ਼), and g(ਗ)/ɣ(ਗ਼); subscripted bindi is also used to distinguish the place of articulation in the pair l(ਲ)/ɭ(ਲ਼). Consonants of the Punjabi language (Table 3) with their Gurumukhi symbol, the IPA symbol and the phone created for this work. The 35 consonants of the Punjabi language have been divided into 7 groups (ਵਰਗਾਂ) each containing 5

consonants as described in Table 3. The velar nasal consonant /ŋ / can be added to the end of a syllable by adding a subscripted bindi (ੰ) and tippi (ੱ) to the preceeding vowel, e. g. **ਕਾਂ, ਅੰਬ.** All affricates are palatal.

Table 2: Punjabi Vowel Set

Independent Vowel	ਅ	ਆ	ਇ	ਈ	ਏ	ਐ	ਉ	ਊ	ਓ	ਔ
Dependent Vowel/ Matra	◌	◌ਾ	◌ਿ	◌ੀ	◌ੇ	◌ੈ	◌ੁ	◌ੂ	◌ੋ	◌ੌ
IPA	ə (= a)	a: (= aa)	I(=iy)	i : (=yi)	e: (=ae)	ɛ: (=ay)	Ʊ(=u)	u: (=uu)	o: (=o)	ɔ: (=au)

3.2.3 Semi-Vowels

There are two semi-vowels in Punjabi language. 'ਵ' is labiodental and 'ਯ' is palatal.

3.2.4 Tonal Nature

Tonal nature (5, 15) is the most distinctive feature of Punjabi language with phonemic tones and as a result, words with identical spellings can be differentiated by changing tones. The contour tones used by the Punjabi language are realized over two successive syllables. In a contour tone system, each tone is implemented by a shift of the pitch. Patterns of pitch variation are employed in a tonal language to distinguish different meanings of words which have the same pattern of consonants and vowels. Low, level and high are three phonemically different tones of the Punjabi Language. These are indicated in writing using the voiced aspirate consonants **ਘ, ਝ, ਢ, ਧ, ਭ** and the inter-vocal **ਹ**.

3.2.5 Gemination

In Indo-Aryan Languages like Hindi or Urdu, the preceding vowel is lengthened for gemination whereas in Punjabi language, an addak (◌ੱ) is generally used as a diacritic to incorporate Gemination and thereby the subsequent consonant is found to be doubled or reinforced or lengthened. Addak is placed above the previous consonant in order to signal gemination of the following consonant. e.g. **ਸੱਤ, ਪੱਤਾ** Gemination leads to a change in the meaning of the word. As a result **ਸਤ,** which means truth, becomes **ਸੱਤ** which means seven.

Table 3: Punjabi Consonant Set

Phonetic Property Category	Primary Consonant [Unvoiced]		Secondary Consonant [Voiced]		Nasal(IPA)
	Unaspirated (IPA)	Aspirated (IPA)	Unaspirated(IPA)	Aspirated(IPA)	
Kavarg Toli i.e. Gutturals	ਕ (k=k)	ਖ (k^h=kh)	ਗ (g=g)	ਘ(kl=gh)	ਙ(ŋ)
Chavarg Toli i.e. Patatals	ਚ (tʃ=ch)	ਛ($tʃ^h$=chh)	ਜ(dʒ=j)	ਝ(tʃl)	ਞ(ɲ)
Tavarg Toli i.e. Cerebrals	ਟ(ʈ=tt)	ਠ ($ʈ^h$=tth)	ਡ(ɖ=dd)	ਢ(ʈl=ddh)	ਣ(ɳ=nn)
Tavarg Toli i.e. Dentals	ਤ(t̪=t)	ਥ($t̪^h$=th)	ਦ(d̪=d)	ਧ(t̪l=dh)	ਨ (n=n)
Pavarg Toli i.e. Labials	ਪ (p=p)	ਫ (p^h=f)	ਬ (b=b)	ਭ(pl=bh)	ਮ (m=m)
Semi-Vowels (IPA)	ਯ(j=y), ਵ(ʋ=v)				
Affricates (IPA)	ਚ, ਛ, ਜ				
Flap (IPA)	ੜ (ɽ=rh)				
Lateral	Dental & Alveolar		ਲ(l=l)		
	Retro-Flex		ਲ਼ (ɭ)		

3.2.6 Fricatives

Fricative consonants are speech produced, when air is forced through a narrow channel made by placing two articulators close together. Fricatives of Punjabi language along with their place of articulation are shown in the Table 4.

Table 4: Fricatives

Labio-Dental	Dental & Alvelor	Palatal	Velar	Glottal
ਫ਼(f=ff)	ਸ (s=s), ਜ਼ (z=zz)	ਸ਼ (ʃ=sh)	ਖ਼ (x=khh), ਗ਼ (ɣ=gg)	ਹ (ɦ=hh)

3.2.7 Syllables

There are 7 syllables types of Punjabi language including V, VC, CV, VCC, CVC, CVCC and CCVC. There are around two million syllables including nasal and non nasal vowels for the Punjabi language, not considering tone distinctions.

4. Previous Work

Kumar (5) initiated work on ASR for the Punjabi language by creating an experimental, speaker dependent, real-time, isolated word recognizer for Punjabi. Whole word models were the basis of their

speech recognition task. The scope of this work was intended as a comparison of the performance of ASR on small vocabulary speaker dependent isolated spoken words using HMM and Dynamic Time Warping (DTW) techniques. A template-based recognizer using linear predictive coding with dynamic programming computation and vector quantization with HMM recognizers in isolated word recognition tasks were the two approaches used in this work.

Dua et al. (7) showed automatic speech recognition for isolated words of the Punjabi language by using Hidden Markov Model toolkit (HTK) 3.4v installed on Linux environment Ubuntu 11.10. They used whole word models for their speech recognition task. Type of speakers and nature of environments, i.e., room environment vs open space, were used as parameters for the performance evaluation of developed ASR systems. To build an interactive system, a user interface was developed in JAVA.

Kuldeep et al. (9) attempted to develop an ASR for recognizing isolated word speech for the Hindi language using acoustic word models. Vocabulary size was just 30 words of the Hindi language. HTK v3.4 was used on linux to develop the system.

Kuldeep et al. (8) developed an ASR system for recognizing connected words of the Hindi language. The developed ASR is based on acoustic word models. A grammar-based language model was created using extended Backus-Naur form specifications. The system was trained on a task with a vocabulary size of 102 words with the speech of 12 speakers of age group 18-23 years. The test module contained 190 speech samples of 3 male + 2 females = 5 speakers of age group 17-25 years. The HTK v3.4 was used to implement the system.

Mishra et al. (10) developed a speaker independent connected Hindi digits recognition system for clean as well as noisy environments using robust feature extraction techniques such as revised Perceptual linear prediction, Bark frequency cepstral coefficients and mel frequency perceptual linear prediction. Except MFCC features which were extracted using HTK tool, all other features were computed using MATLAB and later saved in HTK format. A 5-state HMMs with 9-Gaussian mixtures were used for all the experiments. For recognition in noisy environments, four types of noises were considered i.e., babble, white, pink and F-16 noise. Results obtained in their experiments showed that the MF-PLP feature extraction technique gave the best results in comparison to all other techniques.

5. Implementation

Depending upon the type of utterances, speech (17) is classified as isolated words, connected words, continuous speech and spontaneous speech. In this paper, an ASR system for recognizing isolated word and connected word speech of the Punjabi language has been developed using the HTK toolkit on the Linux platform. The latest version of HTK i.e., HTK 3.4.1 (11) is an open source platform comprising a set of modules based on ANSI C. This kit has been used for developing the system. The HTK training

tools (11, 20) have been used to estimate the parameters of a set of HMMs using training utterances and their transcriptions.

5.1 Task Grammar & Network

The HTK utility HParse is used to compile a task grammar and subsequently obtain the task network. This task grammar basically sets the platform for defining the structure of test samples. Another HTK tool, HSGen, is used to ensure the accuracy of writing the task grammar in the presence of a task dictionary by generating a predefined number of grammatically confirmed sentences.

ANGAD	[ANGAD]	ax nng g ah d	BHEIN	[BHEIN]	bh ey nn
ARJUN	[ARJUN]	ax r j ah n	CHANGI	[CHANGI]	ch ah nng g ii
ASEEN	[ASEEN]	ax s ii ng	DOU	[DOU]	d o
BHEIN	[BHEIN]	bh ey nn	HUUNDAA	[HUUNDAA]	hh uh nng d aa
BHEINAA	[BHEINAA]	bh ey nn aa	KIYON	[KIYON]	k ih oh ng
BHRAA	[BHRAA]	bh r aa	VICH	[VICH]	v ih ad ch
ATTH	[ATTH]	ah tth	HUKK	[HUKK]	hh ah ad k

Figure 2: Example enteries from the dictionary

These sentences can be used as test prompts for obtaing test samples during performance evaluation.

5.2 Pronunciation Dictionary

A pronunciation dictionary was created so that speech audio & transcriptions can be efficiently compiled into acoustic models. Bindi, Addak and Tippi are denoted in the dictionary (Figure 2) using the phone symbols ‘ng’, ‘ad’ and ‘nng’. As previously discussed in section 3.2 that gemination in Punjabi Language is being applied with the help of Addak. The vocabulary for speech recognition training and testing experiments comprises 200 words, selected in a phonetically balanced way so that all the phones of the Punjabi language can be covered corresponding to the correct pronunciations. From this vocabulary, a prompts file was created, and validated to confirm the inclusion of all phones with an average frequency of 6 occurrences per phone (minimum 4 occurrences of phone, maximum 8 occurences of phone). The HTK tool HDMan was used to check the prompts word list against the dictionary to confirm the presence of a pronunciation for each word in the prompts file.

5.3 Phone Transcription File

A perl script was used to create word level transcriptions in the form of an MLF (Master label file) which contains a label for each word in the prompts file. Later on, HLed tool was used to transform

word level transcriptions to phone level transcriptions. The result is a phone level MLF which provides phone level transcriptions for each utterance in the prompt file.

5.4 Speech Corpus

ASR is trained on a pre-determined set of 200 words of the Punjabi language recorded in .wav format using a unidirectional microphone in a closed room environment with Audacity 2.0.0 (12). Six = 3 male & 3 female speakers were invited to record training data samples. 5 utterances for each word from each speaker were recorded. A sampling rate was fixed at 48 KHz. A total of 6*5*200 = 6000 speech files were recorded.

5.5 Acoustic Analysis

Recorded speech data are parameterized into a sequence of features with the help of the HCopy tool of HTK. MFCC features extraction was applied. A sampling rate of 16 kHz & a frame rate of 10 ms with Hamming window of 25 ms are used for feature extraction. 39 MFCC features, comprising 12 mel cepstra plus log energy and their first & second order derivatives, are extracted.

```
QS 'R_NonBoundary' { "*+*" }
QS 'R_Velar' { "*+gh","*+gg","*+g","*+kh,"*+k" }
QS 'R_Palatal' { "*+sh","*+vh","*+jh","*+j","*+chh","*+ch" }
QS 'R_Labial' { "*+m","*+b","*+ph","*+pp","*+p" }
QS 'R_Dental' { "*+z","*+r","*+l","*+s","*+n","*+d","*+th","*+t" }
QS 'R_Stop' { "*+chh", "*+ch", "*+gh", "*+gg", "*+g", "*+kh", "*+k", "*+dh", "*+dd", "*+d",
"*+tth", "*+tt", "*+th", "*+t", "*+bh", "*+b", "*+ph", "*+p" }
QS 'R_Nasal' { "*+nng","*+ng","*+nn","*+n","*+m" }
QS 'R_Vowel' { "*+yi", "*+iy", "*+oh", "*+o", "*+ou", "*+ih", "*+ii", "*+ey", "*+ay", "*+ax",
"*+ah", "*+uh","*+uu","*+au","*+ae","*+ad","*+aa"
```

Figure 3: Few Questions

5.6 Monophone HMMs & Re-Alignment

HMM training is based on the definition of a prototype model which provides a model structure. In our work, a topology of 3-state left-right with no skips is used for its definition. Flat start monophones are created with the help of this prototype model. The HCompV tool helps in this regard. HERest is an embedded re-estimation tool and has been used for re-estimating flat start monophones at least thrice. A wider pruning beam has been applied to fix poor acoustic matching, which may occur when there are few training files. Robustness of the model is another issue to be addressed e.g. obtained by fixing the silence models. The HHEd tool is used to fix the silence models by tying the emitting state of the sp (short pause) model to the centre state of the silence models. The resultant phone transcriptions are used to re-estimate

monophones using the HERest tool. The training data are realigned using the latest phone models with the help of the HTK tool HVite. All the pronunciations for each word are reconsidered at this step and a pronunciation best matching the acoustic data is chosen, to be used for further processing.

5.7 Tied State Triphones

The use of triphones instead of monophones can help in improving recognition accuracy. Monophone transcriptions were transformed to context dependent triphone transcriptions with the help of the HLEd tool. Cloning of models was carried out with the help of the HHEd tool. Cloned models were re-estimated using the triphone-transcribed speech data. Triphone states were then tied as necessary to provide robust parameter estimates. Left and right contexts defined by phonetic context questions were selected from a question list. The list of questions was created manually; example entries are shown in the Figure 3. The HHEd tool performs decision tree state tying.

6. Perfomance Evaluation

ASR performance was evaluated in three phases. For performance evaluation, percentage accuracy metric was used. The results are shown in the Figure 4. With I: Insertion Error, D: Deletion Error, S: Substitution Error and N: Total Words

$$\text{Percentage accuracy} = \frac{N-I-S-D}{N} * 100$$

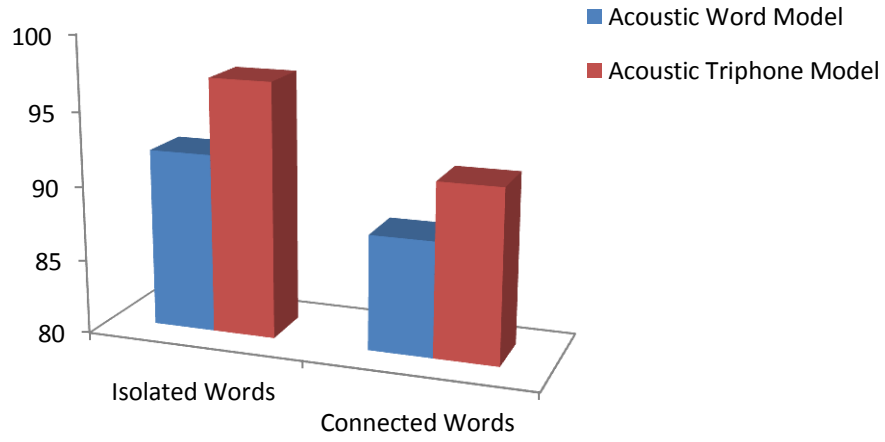

Figure 4: ASR PERFORMANCE

6.1 Phase 1: Isolated word speech recognition

In this phase, ASR performance for isolated words speech was evaluated for the system based on acoustic word model and acoustic triphone model respectively. Their performances were compared and it was found that acoustic phone model outperforms acoustic phone model. Test speech samples for 70 words were recorded to 5 speakers including 3 males & 2 females. Percentage accuracy was 92.05% for acoustic word model and 97.14% for acoustic triphone model respectively.

6 Phase 2: Connected word speech recognition

In this phase, ASR performance for connected words was evaluated for the system based acoustic word model and acoustic triphone model respectively. Their performances were compared and it was found that acoustic triphone model outperforms acoustic monophone model. Test speech samples for 50 sequences of words were recorded to 5 speakers including 3 males & 3 females. Each sequence of speech sample contained 6 words. Percentage accuracy was 87.75% for acoustic word model and 91.62% for acoustic triphone model respectively.

7 Conclusion & Future Work

The phone based acoustic model approach is new to the Punjabi language automatic speech recognition. This paper focuses on implementing an ASR for recognizing isolated word and connected word speech in the Punjabi Language. The HTK 3.4.1 speech engine was used for implementing our ASR systems. Two approaches of acoustic modelling were used: whole word models and triphone models. Performance evaluation has shown that triphone based acoustic modelling outperforms acoustic whole modelling. The word recognition accuracy of isolated word speech was 92.05% for acoustic whole word model based system and 97.14% for acoustic triphone model based system. The word recognition accuracy of connected word speech was 87.75% for acoustic whole word model based system and 91.62% for acoustic triphone model based system. Future work will include the development of continuous speech recognition, the development of faster training and testing and advanced user interface, e.g., using the speech recognition tool Sphinx, the improvement of robustness by applying noise reduction, speech enhancement techniques for better results, and the development of large vocabulary continuous speech recognition (LVCSR) system for the Punjabi language.

8 References

1. Rabiner L, Juang BH, Yegnanarayana B. Fundamentals of Speech Recognition. Pearson Publishers; 2010.
2. Rabiner LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE; 1989. Vol.77, No.2, pp. 257-286.
3. Livescu K, Lussier EF, Metze F. Sub-word modeling for automatic speech recognition: Past, Present, & Emerging Approaches. Signal Processing Magazine, IEEE; Nov 2012. Volume: 29, Issue: 6, pp. 44-57. ISSN: 1053-5888
4. Thangarajan R, Natarajan AM, Selvam M. Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language; 2008 March. WSEAS TRANSACTIONS on SIGNAL PROCESSING. Available from http://www.wseas.us/e-library/transactions/signal/2008/30-649.pdf.
5. Singh PP.; 2010. *Sidhantak Bhasha Vigiyaan*, Madaan Publication, Patiala.
6. Kumar R. Comparison of HMM and DTW for Isolated Word Recognition of Punjabi Language. In Proceedings of Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Sao Paulo, Brazil. Springer Verlag; 2010 Nov 8-11. Vol. 6419 of Lecture Notes in Computer Science (LNCS), pp. 244– 252. Available from: http://link.springer.com/chapter/10.1007%2F978-3-642-16687-7_35.
7. Dua M, Aggarwal RK, Kadyan V, Dua S. Punjabi Automatic Speech Recognition Using HTK. International Journal of Computer Science Issues. Vol. 9, Issue 4, No 1; Jul 2012. Available from: http://ijcsi.org/papers/IJCSI-9-4-1-359-364.pdf
8. Kumar K, Aggarwal RK, Jain A. A Hindi speech recognition system for connected words using HTK. International Journal of Computational Systems Engineering; 2012 Vol.1, No.1, pp.25 – 32. Available from: http: //www. inderscience.com /info/inarticle.php? artid=44740.
9. Kumar K, Aggarwal RK. Hindi Speech Recognition System Using HTK. International Journal of Computing & Business Research. ISSN (online): 2229-6166. Vol. 2 Issue 2; 2011 May.
10. Mishra AN, Biswas A, Chandra M, Sharan SN. Robust Hindi connected digits recognition. International Journal of Signal Processing, Image Processing and Pattern Recognition. Vol. 4, No. 2; 2011 Jun. Available from: http://www.sersc.org/journals/IJSIP/vol4_no2/8.pdf
11. HTK-3.4.1 tool kit retrieved Jul 7, 2012 from http://htk.eng.cam.ac.uk.
12. Audacity 2.0.0, retrieved; 2012 Jul 15 from http://download.cnet.com/Audacity/
13. Martin JH, Jurafsky. Speech & Language Processing. Pearson Education; 2000.
14. Kesarkar MP. Feature extraction for speech recognition. M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay; 2003 Nov. Available from: http://www.ee.iitb.ac.in/~esgroup/es_mtech03_sem/ sem03_paper_03307003.pdf
15. Lata S. Challenges for Design of Pronunciation Lexicon Specification (PLS) for Punjabi Language. Available from: http://hnk.ffzg.hr/bibl/ltc2011/book/papers/MPLRL-4.pdf; 2011.
16. An Introduction to Gurmukhi. Available from: http:// guca.sourceforge. net/resources /introductiontogurmukhi/ an.introduction .to. gurmukhi.pdf; 2005.
17. Anusuya MA, Katii SA. Speech Recognition by Machine. International Journal of Computer Science & Information Security; 2009. Vol. 6, No. 3.

18. Yook D. Introduction to Speech Recognition. Department of Computer Science, Korea University; 2003. Available from: http://ai.korea.ac.kr/data/readings/intro/doc/yook.lecture-note-asr-1.pdf

19. Rothkrantz LJM. Automatic Speech Recognition Using Hidden Markov Model. TUDelft. IN4012TU, Real-time AI & Automatische Spraakherkenning; 2003. Available from: http://www.kbs.twi.tudelft.nl/docs/syllabi/speech.pdf.

20. HTK Book. Retrieved on Mar 18, 2012 from http://htk.eng.cam.ac.uk.

21. http://simple.wikipedia.org/wiki/Punjabi_language

22. Morris JJ. A STUDY ON THE USE OF CONDITIONAL RANDOM FIELDS FOR AUTOMATIC SPEECH RECOGNITION. Dissertation, Ohio State University, USA. Available from http://www.cse.ohio-state.edu/~morrijer/Publications/DissertationMorris.pdf; 2010