Continuous Speech Recognition for Punjabi Language

Wiqas Ghai
Khalsa College, Mohali, Punjab

Navdeep Singh
Mata Gujri College, Fatehgarh Sahib

ABSTRACT

Punjabi language is a tonal language belonging to an Indo-Aryan language family and has number of speakers all around the world. Punjabi language has gained acceptability in the media & communication and thereby deserves to get a pace in the growing field of automatic speech recognition which has been explored already for number of other Indian and foreign languages successfully. Some work has been done in the field of isolated word and connected word speech recognition for Punjabi language. Acoustic template matching and Vector quantization have been the supporting techniques. Continuous speech recognition is one area where no work has been done so far for Punjabi language. In this paper, an effort has been made to build automatic speech recognizer to recognize continuous speech sentences by using Tri-Phone based acoustic modeling approach on HTK 3.4.1 speech engine. Overall recognition accuracy has been found to be 82.18% at sentence level and 94.32% at word level.

Keywords

Tri-Phones, ASR, Hidden Markov Model, MLF, Acoustic Model, HTK, Gaussian Mixtures

INTRODUCTION

Automatic speech recognition [1] is a system shown in Fig 1 which performs automatic transformation of an acoustic signal of input speech to a text transcription. The real purpose behind ASR research is to allow a computer to recognize speech in real time with 100% accuracy irrespective of vocabulary size, noisy scenario, speaker characteristics, accents and channel conditions. The growing field of automatic speech recognition is facing various issues which are open with due regard to vocabulary size, mode of speech, speaker mode and above all the environmental robustness. These issues are being dealt by researchers so that performance of ASR systems can reach an optimum level. Automatic speech recognition is being explored to great extent in the application areas such as: Voice user interface, Voice interactive response, enhancing social interactive capability of handicapped people, learning a foreign language etc.

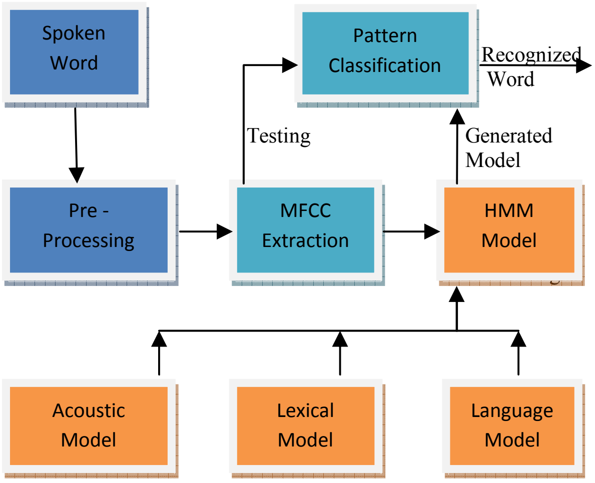

Fig 1 Automatic Speech Recognition System

1.1 Continuous Speech Recognition

Continuous speech is a speech in natural flow. It is comprised of connected words which are not separated by pauses. Practical implementation of automatic speech recognition requires the capabilities to handle continuous speech. Most impostant feature of continuous speech is that the pronunciations are highly dependent on the context. Recognition of continuous speech is complex due to the following reasons:

i. Unlike Isolated word and connected word speech, utterances in continuous speech utterances become overlapped. As a result, word boundaries become unclear and it becomes difficult to identify the start & end points of words.

ii. Co-articulation: Co-articulation [2] is a phenomenon which guides natural sounding speech and, brings into effect, the suprasegmental characteristics. It occurs at the boundary between the words. These suprasegmental characteristics degenerate the phonemic boundaries and thereby leads to acoustic variability for the initial & final parts of the word spoken in continuous speech. Finally errors are likely to occur in the recognition system.

iii. Rate of speech: Speech rate more specifically signifies the speaking rate and can be determined as number of words spoken per minute. Speech rate always has an impact on our fluency. Style of speaker and nature of text are two important properties of speech rate explained by Mathew [3]. Two parameters based on these properties are as:

Mean speech rate, $\mu = f$ (Speaker)
Variance, $\sigma^2 = g$ (Cognitive load)

Cognitive load is linked with the nature of text and describes the effort made & creativity required for selection of text to be spoken. It has been found that with change in the speech rate [4], duration of vowels show more changes than the consonants.

1.2 Tri-Phones In Acoustic Modeling

Automatic Speech Recognition problem is represented with the help of Bayes' Rule [1] as:

$$W^{*} = \arg_{w} \max \frac{P(A/W).P(W)}{P(A)}$$

2.1

W: Observed word sequence

A: Sequence of acoustic observation vectors

Both Language Model & Acoustic Model are used to provide P (W) and P (A/W) respectively. As per the equation 2.1, many samples of each possible word sequence are to be obtained and then they are to be transformed to corresponding acoustic vector sequence. As the vocabulary size increases, the set of probable word sequences becomes very large. Segmentation of speech signal into fundamental acoustic units is another major aspect of automatic speech recognition and rather the continuous speech recognition. The benefit of using word, as acoustic unit, is that its acoustic representation is well defined. But using acoustic word model for continuous speech recognition [5] [6] poses problems such as individual training of words, absence of parameters sharing, linear growth of memory requirement etc. The solution to these problems requires a very large training set which is not an economical decision. So, acoustic word model is appropriate only for small vocabulary.

Incorporating the feature "Sharing of parameters" can help in saving the computing resources. Phone represents the basic sounds for an utterance. Phone acts as a sub-word acoustic unit which can bring this feature into action. Statistical correspondence between phones and acoustic vectors can be used to evaluate W^* defined in equation 2.1. It has been found that due to co-articulation effects, the beginning and end of a phone have been modified by the previous & next phone. Therefore, context dependence of phones [6] is a feature which brings down the efficiency of ASRs due to resulting acoustic variations of phonetic units. So the context dependent variations should be modeled. Tri-phones are the sequence of three phones L-X+R. This triphone structure describes the left context (L) and the right context (R) of the current phone (X) in the word under observation. e. g. triphones for word 'ਸੁਰਜ' are s-uu+r, uu-r+ax, r-ax+j. The final phone of that word is always modeled by a cross-word triphone having SIL as its right context or first phone of next word as its right context. In other words, tri-phone is used to establish the different contexts in which a phoneme can occur in an utterance. There are two ways to construct context dependent models. One is inter-word models [7] and the 2nd one is cross-word triphone. It too has been found [8] that cross-word tri-phones are more in number than word internal tri-phones. Since there are almost no clear pauses between words in continuous speech, so cross-word triphones becomes a better choice. HMMs can be used to determine the probability of correspondence between an acoustic vector and a specific tri-phone. A tri-phone HMM has been shown in the Fig 2.

A_{ij}: Probability of switching from state S_i to state S_j

P_i: Probability of producing acoustic vector V at time stamp t

V_t: Acoustic Vector corresponding to observation sequence V

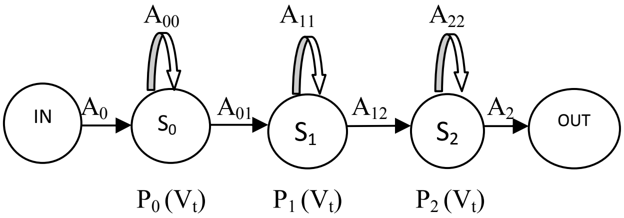

Fig 2 Tri-Phone HMM

Majority of tri-phones are rare and their robust training becomes impossible. Contextual information has to be constrained on the basis of context similarity. This is done by mapping rare phones to more frequent & much better trained tri-phones. A triphone mapped HMM is carried out as follows:

i. Monophones models training using single Gaussian mixtures

ii. Incrementing the number of mixture components in each state and thereby multiple GMMs are trained.

iii. Clonning of state output distributions of Monophones for all triphones

iv. Training triphone tied system

1.2.1 Triphone Clustering & Decision Tree

In contrast to monophone modelling involving k words [7] [9] for successful training, triphone models require k^3 words for their training. This much training data even if becomes available, some possible triphone may not occur in training data and some triphones may occur with such a frequency which can not provide optimum estimation. To overcome this problem, clustering of triphones is performed and criterion for clustering is phonetic similarity. There are two mechanisms for clustering. First one is data driven which allows the states to be clustered by using a similarity measure between the states and then second one uses decision trees, as shown in Fig 4. Construction of decision/phonetic tree for each phone requires phone set of the target language, design of phonetic questions and annotated training data set.

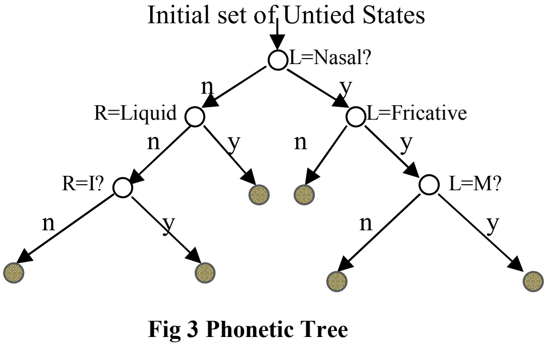

Fig 3 Phonetic Tree

Each node of tree corresponds to a question on a specific question regarding the phonetic structure of phones in context of the current phone. States in each leaf node are tied. In contrast to Data-driven approach, Decision tree algorithm deals with unseen triphones effectively and has the ability to cluster the entire model.

1.2.2 Phonetic Questions

Phonetic questions form the basis for phonetic tree. The phonetic properties well defined for each language are used to design the list of questions. There are mainly three phonetic properties: Place, Manner and Type of excitation. Phones are classified with due regard to these properties. This classification sets the stage for generating phonetic questions. Some of the main components in the list of questions prepared for Punjabi language, defined about left & right contexts of each triphone, have been shown in Fig 4. All the contexts, required for optimum acoustic realization of each phone, have been included in the questions.

```
………………………………
QS "R_Nasal"    { *+m, *+n, *+nn, *+ng, *+nng}
QS "R_Vowel" { *+aa, *+ad, *+ae, *+au, *+uu, *+uh, *+ah, *+ax,
*+ay, *+ey, *+ii,*+ih,*+ou,*+o,*+oh,*+iy,*+yi }
QS "R_Unvoiced-Cons" { *+p, *+pp, *+t, *+tt, *+k, *+kk, *+kh,
*+s, *+ss, *+sh, *+f, *+th, *+tth, *+hh, *+ch, *+chh }
QS "R_Voiced-Cons"     { *+j, *+jh, *+b, *+d, *+dd, *+dh, *+g,
*+gg, *+y, *+l, *+ll, *+m, *+n, *+ng, *+nn, *+nng, *+r, *+v,
*+rh, *+z }
QS "R_Unvoiced-All"    { *+p, *+pp, *+t, *+th, *+tt, *+k, *+kk,
*+kh, *+s, *+ss,*+sh,*+f,*+tth,*+h,*+hh,*+ch,*+chh,*+sil }
QS "R_Affricate"       { *+ch,*+chh,+*j,*+jh }
QS "R_Fricatives"      { *+s,*+ss,*+h,*+hh,*+sh,*+z,*+f,*+x}
QS "R_Laterals"        { *+l,*+ll }
QS "R_Trill"           { *+r }
QS "R_Flap"            { *+rh }
………………………………
```

Fig 4 List of Questions

The extra questions created will be ignored whenever they will be found irrelevant to the data. Use of Phonetic questions and tree topology is meant for maximizing the likelihood of training data. A well designed list of phonetic questions and then decision trees for all phones lead to better generation of tied state triphone models.

2. PUNJABI LANGUAGE

Punjabi language has been a famous Indo-Aryan language with increasing acceptability in media and communication which has raises its prospects for accelerated pace in the field of ASR research & development. Punjabi language has 32 different dialects. Punjabi is based on Gurumukhi Script [10] standardized by 2nd Guru of Sikh religion Guru Angad Dev Ji in 16th century. Punjabi language has earned an official status among 22 official languages of India. Its phonetic inventory has 10 vowels, 25 consonants, 7 diphthongs and three tones. Two important features of Punjabi language are being explored here.

2.1 Tonal nature

Emotional information, emphasis and differentiation among words are the main aspects which are mostly present in the speech. Tone is a feature which is capable of conveying all these information. Tone in speech makes use of pitch which is governed by frequency of vocal chords vibrations. As a result, tone changes the meaning of a word. Few famous tonal languages are Chinese, Vietnames, Bodo and Punjabi. Tonal nature of Punjabi language is its distinctive feature among all other Indo-Aryan Languages. In Punjabi language, there are five aspirated sounds ਘ, ਝ, ਢ, ਧ, ਭ which act as tonal characters. They are considered as multi-function phonemes because their pronunciation varies as per their position in the pronounced word and accordingly three phonemically different tones: High, Low and Level are produced. e. g.

ਘੋੜਾ: Low Tone due to placement at 1st Position

ਲਾਭ: High Tone due to placement at End

In addition to all this, there is a glottal sound 'ਹ'which is also being used to generate tones when placed in middle and end position of word.

e.g. ਚਾਹ: High Tone => ਚ ਆ'

2.2 Gemination of Consonants

In Punjabi language, gemination is being incorporated with the help of a symbol called addak (◌ੱ). This is placed above the preceeding consonant. For example, in Punjabi word "ਅੱਗ" (which means "Fire") addak has been placed on the top of 'ਅ' to geminate 'ਗ'. As a result, in pronunciation of "ਅੱਗ", 'ਗ'is elongated or lengthened and thereby, pronunciation of "ਅੱਗ" becomes "ਅਗ੍ਗ".

3. PREVIOUS WORK

Work on ASR for Punjabi language was initiated by [11] creating an experimental, speaker dependent, real time, isolated word recognizer for Punjabi Language. Whole word model has been the basis of their speech recognition task. Scope of this work was extended to comparison of the performance of ASR for small vocabulary of speaker dependent isolated spoken words using the Hidden Markov Model (HMM) and Dynamic Time Warp (DTW) technique. Template-based recognizer using linear predictive coding with dynamic programming computation and vector quantization with Hidden Markov Model based recognizers in isolated word recognition tasks were the two approaches used in this work.

Dua et al. [12] worked for automatic speech recognition for isolated words of Punjabi language by using HTK 3.4 installed on Linux environment Ubuntu 11.10. They used whole word model for their speech recognition task. Type of speakers and nature of environments i.e. room environment & Open space, have been used as parameters for performance evaluation of developed ASRs. To build an interactive system, a user interface was developed with the help of JAVA.

Sinhala, being National language, is mother tongue of Sinahalese people in Sri-Lanka. It is linked closely to Punjabi Language as it belongs to the family of Indo-Aryan languages. Nadungodage & Weerasinghe [13] developed a continuous speech recognizer using written Sinhala vocabulary. Sinhala sentences were recorded with the help of 'Praat' tool at sample frequency 16 KHz using a mono channel. A bi-gram language model was created. HTK was used for developing continuous speech recognizer. System was trained from only a single female. Sentence recognition accuracy obtained was 75% and word recognition accuracy obtained was 96%. It was observed that most of the incorrectly identified utterances differed from the correct utterances only by 1 or 2 syllables.

Bhaskar et al [14] have made an attempt to develop automatic speech recognition for Telgu language using HTK. 29 context dependent phonemes of Telgu language were considered for training models. Triphone modeled by left to right 5 state HMMs has been used as basic acoustic unit.

Bodo language is a tonal language which belongs to Sino-Tibeton family and enjoys official status in Indian constitution. Bhattacharjee [18] made an attempt to handle tonal nature of Bodo language by developing an ASR for recognizing its Tonal words. A baseline ASR, developed using Mel Frequency Cepstral Coefficient as feature vector & Recurrent Neural Network as recognizer, has been enhanced for the task with the help of two approaches. First was to combine MFCC features with prosodic features and in the second approach, two separate recognizers were used to recognize the base-syllable and tone respectively. There has been an increase of almost 8% in the first approach and then 16% increase in the recognition accuracy of developed ASR.

Banerjee et al [15] have worked for the Bengali language continuous speech recognition. 48 phones of Bengali language were used for the ASR. They conducted training for both monophone & triphone models on same training data. Their results have shown that tied-state triphone based models outperformed monophone based models. In addition to this, they found that tied-state triphone models give better results even in a poor resource scenario. Average recognition rate and percentage accuracy obtained for male speakers were better than the results obtained for the female speakers.

4. EXPERIMENT

Automatic Speech recognition system for continuous speech of Punjabi language has been developed using HTK toolkit on the Linux platform. Latest version of HTK i.e. HTK 3.4.1 [16] has been used for developing the system. HTK training tools have been used to estimate the parameters of a set of HMMs using training utterances and their transcriptions. Phonetic inventory of Punjabi language has been utilized to create proper pronunciations of Punjabi language. Both Seen as well as unseen data have been included during the performance analysis.

4.1 Pronunciation Dictionary

A phonetically balanced pronunciation dictionary is required to compile speech audio & transcriptions into acoustic model. Lexicon preparation has also been given due care so that all the phones of Punjabi language can be covered corresponding to the correct pronunciations. Our lexicon contains 383 words. Creating prompts file involves a serious check on the inclusion of all phones & their frequency of occurrence. This well developed prompts file helped us to create an almost complete words list file. HDMan tool has been used to obtain dictionary and a file containing 48 monophones as shown in the Table 1. Average frequency of occurrence for each phone has been kept at 6.

Table 1: Some Phones for Punjabi Language

Phone	Alphabet		Phone	Alphabet
A	ਅ		AA	ਆ
AD	◌ੱ		AE	◌ੇ
EY	◌ੈ		II	◌ੀ
IH	ਿ◌		J	ਜ
CH	ਚ		CHH	ਛ
NG	◌ਂ		NNG	◌ੰ
TH	ਥ		TTH	ਠ
T	ਤ		TT	ਟ
N	ਨ		NN	ਣ
S	ਸ		SH	ਸ਼

4.2 Speech Corpus

In-house creation of speech corpus has been carried out. 100 sentences of Punjabi language have been designed to record continuous speech in .wav format using unidirectional microphone in a quiet room environment with Audacity 2.0.0 tool [17]. Average number of words per sentence is 7.

ਤੇਰੇ ਪਿਤਾ ਜੀ ਦਾ ਨਾਂ ਕੀ ਹੈ

(What is the name of your father)

ਗੁਰੂ ਨਾਨਕ ਦੇਵ ਜੀ ਸਾਡੇ ਪਹਿਲੇ ਗੁਰੂ ਹਨ

(Guru Naanak Dev Ji is our First Guru)

ਸੋਮਵਾਰ ਹਫਤੇ ਦਾ ਪਹਿਲਾ ਦਿਨ ਹੂੰਦਾ ਹੈ

(Monday is the first day of week)

ਤੂਸੀ ਹੱਕ ਨਾਲ ਇਥੇ ਆਇਆ ਕਰੋ

(You shall come here with right)

ਅਸੀ ਸਾਰੇ ਸ਼ਾਮ ਨੂੰ ਖਾਣਾ ਖਾ ਰਹੇ ਸੀ

(We all were taking meals in the evening)

ਸੱਚ ਨੂੰ ਕੋਈ ਵੀ ਫਰਕ ਨਹੀਂ ਪੈਂਦਾ ਹੈ

(Truth remains unaffected)

ਦੁਜੇ ਗੁਰੂ ਦਾ ਨਾਂ ਗੁਰੂ ਅੰਗਦ ਦੇਵ ਹੈ

(Name of 2nd Guru is Guru Angad Dev)

ਤੇਰੀ ਕੂੜੀ ਦਾ ਨਾਂ ਕੀ ਹੈ

(What is the name of your daughter)

Fig 5 Few Punjabi Sentences for training data

5 male + 4 female = 9 speakers were invited to record continuous speech samples of training data. Few sentences have been shown in the Fig 5. Sampling rate has been fixed at 48 KHz.

4.3 Perfomance Evaluation

There are two parameters which are generally used for the evaluating ASR performance. These are: Sentence recognition accuracy and word recognition accuracy. Evaluation performance of developed ASR has been conducted in three phases. First two phases contains the seen data where as in 3rd phase unseen data has been introduced in test data.

4.3.1 Phase 1

In first phase, Speakers used for getting training samples were used to get test samples on 102 sentences and out of that 86 sentences were recognized correctly. 665 words out of 698 words, contained in 102 sentences, were correctly recognized

4.3.2 Phase 2

In the second phase, Speakers used for recording training samples were used for getting test samples on 90 test sentences and 74 sentences have been recognized correctly. Test sentences in this case were involving seen data i.e. words which have been used for recording training speech sentences but these sentences were different from the training sentences. 522 words out of 558 words, contained in 90 sentences, were correctly recognized.

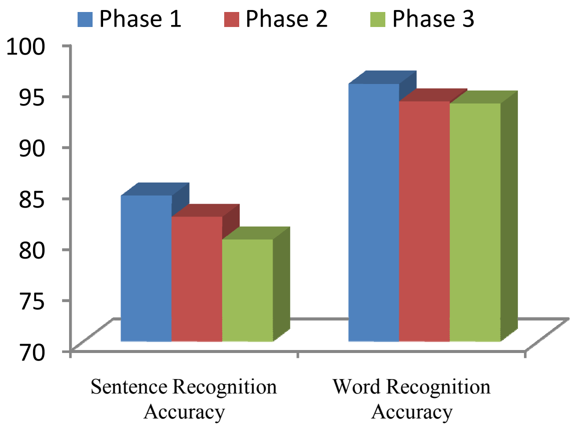

Fig 6: Performance Analysis

4.3.3 Phase 3

Third phase of our work is involved with the use of unseen data in 40 test sentences of continuous speech i.e. use of the words which have not been involved in the training sentences. 32 sentences have been recognized correctly. Some of the unseen words which have been used in the sentences are “ਬੱਚੇ”, “ਅਮਨ”, “ਹਿਸਾਬ”, “ਸੱਚਾ”, “ਡਾਂਗ”, “ਦਿਨਾਂ” and “ਉਸਨੇ”. 224 words out of 240 words, contained in 40 sentences, were correctly recognized.

The performance of developed ASR for three phases has been depicted in the Fig 6 where percentage sentence recognition accuracy and percentage word recognition accuracy has been compared for the three phases.

Overall sentence recognition accuracy has been found to be 82.18% and word recognition accuracy has been found to be 94.32%.

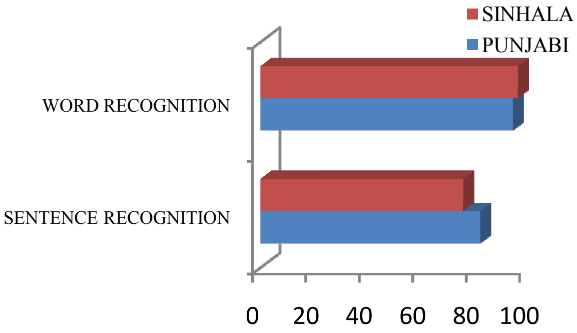

Fig 7: Performance comparison

As mentioned earlier in section 3, Sinhala also belongs to Indo-Aryan family of languages. Peformance of our ASR has been compared with that Sinhala continuous speech recognizer [13] and the analysis has been shown in Fig 7. Sentence recognition accuracy is better for Punjabi language with a margin of 6.44% whereas word recognition is better for Sinhala language with a margin of 1.82%.

5. CONCLUSION & FUTURE WORK

This paper has been focused on the work for implementing ASR for recognizing the continuous speech of Punjabi Language using triphone based acoustic models. HTK 3.4.1 speech engine has been used for implementing ASR. Sentence recognition accuracy for all the three phases of experiment has been found in the range of 80-87% and word recognition accuracy has been in the range of 93-96%. Improvement in the sentence recognition accuracy is very much required. Another important aspect which is to be explored for Punjabi continuous speech recognition is its tonal nature. These two aspects will be useful in development of LVCSR for Punjabi language.

6. REFERENCES

[1] Rabiner, L. Juang, B. H., Yegnanarayana, B. 2010. Fundamentals of speech recognition, Pearson publishers.

[2] Dang, J., Honda, M., Honda, K. 2004 Investigation of Co-articulation in Continuous Speech of Japanese Acoustical Science and Technology Acoustical Society of Japan. Volume 25, No. 5.

[3] Mathew, A. S. 1995. Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition. Carnegie Mellon University, Pittsburgh.

[4] Gay, T. 1978 Effect of Speaking Rate on Vowel Format Movements. JASA, Vol. 63, pp. 223 – 230.

[5] Thangarajan, R., Natarajan A. M., Selvam, M. 2008 Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language; WSEAS Transactions on Signal Processing.

[6] Schwartz, R M., Chow, Y L., Rucos, S., Krasner, M., Makhoul, J. 1984 Improved Hidden Markov Modeling Phonemes for continuous speech recognition, presented at IEEE Int. Conf. Acoustics, Speech, Signal Processing. Vol 9, Pages: 21-24.

[7] Lee, C. H., Giachin, E., Rabiner, L. R., Pieraccini, R., and Rosenberg, A. E. 1990 Improved acoustic

modelling for continuous speech recognition. Speech Research Department AT&T Bell Laboratories

[8] Ursin, M. 2002 Triphone clustering in Finnish continuous speech recognition .Master's Thesis, Helsinki University of Technology.

[9] Zeljkovic, I., Narayanan, S. 1993 Improved HMM Phone and Triphone Models for Realtime ASR Telephony Applications AT&T-Laboratories.

[10] Singh, P. P. 2010. Sidhantak Bhasha Vigiyaan, Madaan Publication, Patiala.

[11] Kumar, R. 2010. 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. Lecture Notes in Computer Science (LNCS), (Vol. 6419, pp. 244 – 252), Springer Verlag.

[12] Dua, M., Aggarwal, R. K., Kadyan, V., Dua, S. 2012. Punjabi automatic speech recognition using HTK. International journal of computer science issues, Vol. 9, Issue 4, No. 1.

[13] Nadungodage, T., Weerasinghe, R. 2011 Continuous Sinhala Speech Recognizer Conference on Human Language Technology for Development, Alexandria, Egypt

[14] Bhaskar, P. V., Rao, S. R. M., Gopi, A. 2012 HTK Based Telgu Speech Recognition. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 12, ISSN: 2277 128X.

[15] Banerjee, P., Garg, G., Mitra, P., Basu, A. 2006 Application of Triphone Clustering in Acoustic Modelling for Continuous Speech Recognition in Bengali. Communication Empowerment Lab, IIT Kharagpur.

[16] HTK-3.4.1 retrieved July 7, 2012 from http://htk.eng.cam.ac.uk

[17] Audacity 2.0.0, retrieved June 15, 2012 from http:/ /download. cnet .com/Audacity/

[18] Bhattacharjee, U 2013. Recognition of the Tonal Words of Bodo Language. International Journal of Recent Technology & Engineering. ISSN: 2277-3878, Volume-1, Issue-6, January 2013