KATHANA - Fluent Speech Recognition System based on Hidden Markov Model for Sinhala Language

S. Jayasena, K. Wimalawarne, T. R. Munasinghe, N. M. P. Cooray, H. R. Yatawatte and D.C.P. Rajapakse
Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka.

Abstract- **This paper discusses speech recognition based on Hidden Markov Model for Sinhala language. There are two main speech technology concepts in this scenario as speech synthesis and speech recognition. Speech synthesis is the artificial production of human speech. Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. This paper is based on KATHANA speech recognition system where the intention is to develop a system which is capable of converting human speech done in Sinhalese to text/ command.**

I. INTRODUCTION

Automatic speech recognition (ASR) is an active research area that enables easy to use communication between human and machine. Significant development has been happened during last two decades relating to this aspect and various adopting and perfecting techniques are happening based on Hidden Markov model (HMM) and artificial neural networks (ANN) (1).

Speech recognition systems can be characterized by many parameters, such as speaking model, speaking style, vocabulary. An isolated-word speech recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system does not. Spontaneous or extemporaneously generated, speech contains disfluencies, and is much more difficult to recognize than speech read from script.

Some systems require speaker enrollment, a user must provide samples of his or her speech before using them, whereas other systems are said to be speaker-independent, in that no enrollment is necessary. A speaker independent system is developed to operate for any speaker of a particular type, whereas a speaker adaptive system is developed to adapt its operation to the characteristics of new speakers. Some of the other parameters depend on the specific task.

Recognition is generally more difficult when vocabularies are large or have many similar-sounding words. When speech is produced in a sequence of words, language models or artificial grammars are used to restrict the combination of words.

Most speaker independent speech recognition systems consider the phoneme as the fundamental processing unit, thus represents words by concatenation of successive phonemes. Phonemes are represented using context dependent models such as triphone and demiphone where as words are represented using a pronunciation lexicon that contain words as a concatenation of phonemes (1)

A HMM in ASR output a sequence of n-dimensional real valued vectors such that each of this vector would consists of ceptral coefficients, which are obtained by taking a Fourier transform of a short time window of speech and de correlating the spectrum using a cosine transform, then taking the first (most significant) coefficients. In each state, the hidden Markov model tends to have a statistical distribution that is a mixture of diagonal covariance Gaussians which will give likelihood for each observed vector. Each word or each phoneme has a different output distribution. A hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes (2)

II. METHODOLOGY

The standard approach for implementing large vocabulary continuous speech recognition is to assume a simple probabilistic model of speech production. This approach is based on the fact that a specified word sequence, W produces an acoustic observation sequence Y with a probability $P(W,Y)$. Then the word sequence is decoded based on the acoustic observation sequence such that the decoded string has the maximum posterior (MAP) probability (3).

$$\hat{W} \ni P(\hat{W}|Y) = \max_{W} P(W|Y) \qquad (1)$$

By applying Bayes' Rule to the equation (1),

$$P(W|Y) = \frac{P(Y|W)P(W)}{P(Y)} \qquad (2)$$

The MAP decoding rule in the equation (1) can be rewritten as follows since $P(Y)$ is independent of W.

$$\hat{W} = \operatorname*{argmax}_{W} P(Y|W)P(W) \qquad (3)$$

$P(Y|W)$ in the equation (3) estimates the probability of a sequence of acoustic observations conditioned on the word string. Hence it is known as the acoustic model. $P(W)$ is generally known as the language model and it describes the probability associate with a postulated sequence of words (3).

When the recognizer starts up, it constructs the front end with noise reduction, the recognizer or decoder, and the linguist according to the configuration specified by the user. These components will in turn construct their own subcomponents. For example, the linguist will construct the acoustic model, the dictionary, and the language model. It will use the knowledge from these three components to construct a search graph that is appropriate for the task. Noise reduction techniques are also used for better performance in recognition. Here in this paper we briefly described about those sections.

of these speech segments. For efficient usage it is required to be trained. The training involves mapping models to acoustic examples obtained from training data. When creating an acoustic model the most important thing to do is selecting the model unit for the acoustic model. There are several possibilities such as words, sub-words include syllables and phonemes. Selecting the appropriate model unit depends on the context of which application use and the scale of the vocabulary.

A language model is used in speech recognition systems and speech commanding systems to improve the performance of such systems. Language models help a speech recognizer figure out how likely a word sequence is, independent of the acoustics. This lets the recognizer make the right guess when two different sentences sound the same.

A. Environment Robustness

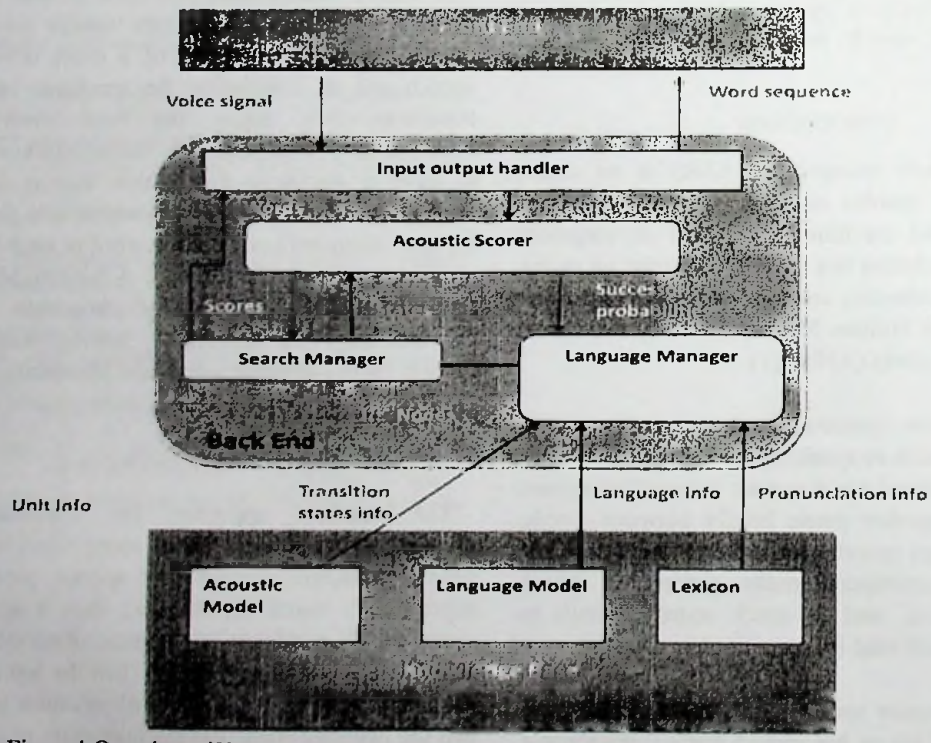

Figure 1 Overview of KATHANA speech recognition system. This consists three main components namely front end, back end and knowledge base.

Background noise can have a significant impact on the performance of a speech recognition system. When models trained in clean conditions are used in the real world, the mismatch between the training conditions and the test causes significant loss in recognition accuracy. Quality noise reduction improves the accuracy of Voice Activity detection and core recognition, both essential parts of a speech recognition system. We have used noise gating technique to improve application robustness in noisy environments.

An acoustic model reflects the way we pronounce a certain language. Acoustic utterance can be broken into phonetic segments. Acoustic Models are representations

Background noise can have a significant impact on the performance of a speech recognition system. When models trained in clean conditions are used in the real world, the mismatch between the training conditions and the test causes significant loss in recognition accuracy. Quality noise reduction improves the accuracy of Voice Activity detection and core recognition, both essential parts of a speech recognition system. KATHANA has used noise gating technique to improve application robustness in noisy environments.

A Noise Gate which is used for KATHANA is a software logic that is used to control the amplitude of samples of an audio signal in time domain. In most simple form, this noise gate allows a signal to pass through (or assign high

weight) only when it is above a set threshold: the gate is 'open'. If the signal falls below the threshold no signal (or less weighted sample) is allowed to pass: the gate is 'closed'. [1] The noise gate is used to extract the clean speech from the detected signal, when it is above the level of the background noise. The threshold is set above the level of the noise and so when there is no signal, the gate is closed. It uses a segment of audio that contains only static background noise, to determine the threshold level to be applied across the signal as a whole.

The spectrum on figure 2. depicts the spectrum for three sinhala words. Among the needful speech signals, uniform environment noise has added.

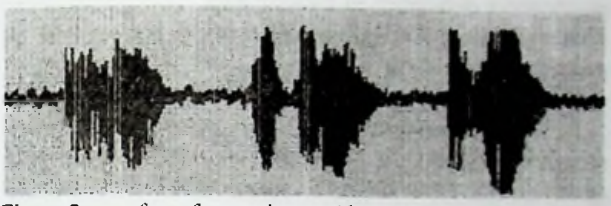

Figure 2 waveform for words තුන (thuna) හතර (hathara) පස (paha) in noisy enviorenment

Amplitude-smoothing is applied, so that the intensity of the sample signal is never suppressed or boosted in isolation. This smoothing procedure involves moving a window of fixed size over each sample in the signal data, applying a mathematical calculation using the sample amplitudes under that window, and replacing the central sample's amplitude with the new value. Mathematical calculation considers the neighborhood amplitudes and estimates a new value for the central sample. Mechanism looks out to preserve the shape of the wave, without having imperfections. Suppressed and spiked noise samples in acoustic data have eliminated.

Figure 3 Speech signal with reduced environment noise

The spectrum on figure 3 depicts the enhanced speech signal form for the considered noisy speech. Continuous background noise has reduced noticeably.

Volume normalization has done, due to different volume levels can cause effects at the core recognizing section. Volume normalization is the process of uniformly increasing (or decreasing) the amplitude of an entire audio signal so that the resulting peak amplitude matches a desired standard. Adjust the volume of audio signal to a standard level, when voice signal provide as a multiple voice files.

B. Recognition

The speech recognizer consists of a signal-processing unit, which transform speech data into an observation vector and a decoder, which find the highest scoring word string, given cepstral or other observation vectors. We are using the most common way of processing signal data for speech recognition, which is cepstral analysis. Cepstral analysis is done by taking the Fourier Transform of the observation , the log of the transformed observation, then taking the inverse Fourier Transform and finally windowed to retain the low order spectra.

Computing the $P(Y|W)$ which is the acoustic model in large vocabulary speech recognition, statistical model for sub word speech units are built. KATHANA builds sub word unit models based on the Hidden Markov Model (HMM).then using this set of sub word HMMs and the word lexicon, a set of word models are built. This is done by concatenating each of the sub word unit HMMs as specified in the word lexicon (3).

Then recognizer does a combined word level/ sentence level match to recognize the spoken utterance. To accomplish this task KATHANA performs two-pass search algorithm using word trellis index.

1. Search mechanism

Kathana performs two-pass algorithm in forward and backward directions using 2-gram and 3-gram models of the language (4). This is carried out using a lexicon structured as a tree with probabilities assigned from language model.

During the first pass 1-gram probabilities assign to intermediate nodes while 2-gram probabilities assign to word-end nodes. 1-gram factoring values are independent from the proceeding words; hence, they can be calculated statistically in a single tree lexicon. However, the values are not optimal theoretically with the true 2-gram probability, but errors occurring due to it can be recovered during second pass.

During the second pass in the reverse direction, rescoring is done using 3-gram language model and precise cross word context dependency. During this sentence dependent N-best score is calculated by connecting backward trellis in the result of first pass. In order to rescore for cross word context dependency and connecting the backward trellis speech input is scan again. Since first output candidate may not be the best one, we compute several candidates by continuing the search and sorting them.

C. Acoustic Model

The process of creation of acoustic models starts with preparation of the training and testing data. This data comprises of utterance recordings by multiple speakers and the corresponding transcripts encoded using the chosen phoneme set for the language. Then features are

extracted from these data and converted into Hidden Markov Models (HMMs). The HMM is a statistical model of characterizing data samples of a discrete time-series. An HMM is compromised of several states each of which essentially represents a certain part of target.

An acoustic model reflects the way we pronounce a certain language. Acoustic utterance can be broken into phonetic segments. Acoustic Models are representations of these speech segments. For efficient usage it is required to be trained. The training involves mapping models to acoustic examples obtained from training data.

1. Feature Extraction

A speech waveform cannot be delivered directly to a speech recognition system because they cannot recognize the raw speech waveform. The raw acoustic signal includes too much redundant and unnecessary information generated by speakers and recording channels and a wide range of background and random noise. Usually, speech signal is converted into a sequence of feature vectors by front-end processing to emphasize the characteristics of spoken words and suppress other irrelevant information. A feature vector is a parameterized representation of an acoustic signal containing the essential information of speech signal and stored in a compact way. In most speech recognition systems some kind of preprocessing techniques are used to obtain feature vectors.

To perform feature extraction initially the continuous speech signal should be divided into short overlapped time slices, which are called *frames* at a certain frame rate. The duration of a frame is typically about 10~20 ms based on the observation that speech signal is quasi-stationary within such a short interval which bring the assumption that, its statistical properties do not change too much. Each frame is then windowed and transformed into frequency domain via spectral analysis to obtain an acoustic feature vector as the representation of that time piece.

There are several methods of speech analysis exists and some of them are motivated by the nature of human hearing. Two of the most popular from the lot are Linear Predictive Coding (LPC) and Mel-Frequency Cepstral Coefficients (MFCC). [5]

2. Selecting Model Unit

When developing an acoustic model the most important question to be answered is what unit of language to use. There are several possibilities exist, such as: words, syllables or phonemes. Each of these possibilities has its own advantages as well as disadvantages. At a high level, following three requirements need to be followed to select an appropriate modeling unit,

- The unit should be accurate in representing the acoustic realization in different contexts.
- The unit should be trainable. Enough training data should exist to properly estimate unit parameters.
- The unit should be generalizable, so that any new word can be derived. [5]

But use of these units can be depending on the context which speech recognition system is used.

For a speech recognition system with small vocabulary such as voice dialing application, word-level speech modeling would be ideal to build an acoustic model. Since it has limited number of words, there is no need to worry about deficiency of training data. Word models are both accurate and trainable and there is no need to be generalized.

When considering a large vocabulary speech recognition system, use of word is not going to be the best option to build the acoustic model. There are so many words to be trained that make finding sufficient training data so difficult. If there are such data the word model become trainable and accurate. Moreover if new words to be added, the training process should restart all over again making the word model not generalizable.

In that scenario using a sub-word unit such as syllable or phoneme would be a better solution as they can share across words. Phoneme is widely accepted as model unit for speech recognition across the world since a language posses a limited number of phonemes. Such a small number makes it easier to find sufficient training data making it trainable. Furthermore, new word can be easily added to vocabulary by defining the pronunciation in terms of phonemes making it generalizable.

A one weakness involved with phonemic modeling approach is accuracy, as it ignores the influence of left and right context, which means hard to get an accurate description of acoustic signal. A solution proposed to address this weakness is *Triphone*, a context dependent model unit that considers not only the current phoneme but also its left and right neighbors, and is potentially capable to capture the co-articulation effect between adjacent speech units. [6]

But using that solution makes the phone model less trainable. As if a language include n number of phonemes there are n^3 number of Triphones. Therefore, these triphones have to be clustered together to form a smaller model set, in order to maintain the trainability of the model.

3. Training

The process of creation of acoustic models starts with preparation of the training and testing data. This data comprises of utterance recordings by multiple speakers

and the corresponding transcripts encoded using the chosen phoneme set for the language. Then features are extracted from these data in the form of mel frequency cepstral coefficients and utilized to train Hidden Markov Models (HMMs). An HMM is compromised of several states each of which essentially represents a certain part of target. For an instance a 3 state HMM used to represent a phoneme the three states are correspond to the starting, the middle and the ending segments of phoneme respectively.

Then these extracted feature vectors are fed to the training module along with transcriptions. The training module calculates the transition probabilities of each feature vectors for an HMM state. It utilizes Baum-Welch Re-estimation algorithm to generate HMMs corresponds to all the phonemes occur in the training data. These HMMs store the observation likelihood for each state of the phoneme. Observation likelihood is the probability of a feature vector being generated by a hidden state (phone, etc.) P(features | phone). Usually the model topology is left to right as it enables to natural progression of the evolving signal and self transition can be used to model speech features belonging to the same state [7].

The next step of the training is to generate triphones. In order to generate triphones first monophone transcriptions are converted to equivalent set of triphone transcriptions. Then triphone HMMs are created by cloning monophones and then re-estimating using triphone transcriptions. In order to identify the context of each phoneme a decision tree is built based on the right and left context of the phoneme. The decision tree attempts to find those contexts which make the largest difference to the acoustics and then cluster similar models. Then tie states within triphone sets in order to share data and thus be able to make robust parameter estimation.

D. LANGUAGE MODEL

Language model contains dictionary, grammar and mapping from dictionary to symbols in grammar. This grammar helps to limit the search and make accurate the recognition.

In this research we worked on Phoneme separation of words in Language Model, Out Of Vocabulary words problem and N-Gram Language Model. These areas and content of the language model are thoroughly described in the paper.

1. Developing language model

Speech recognition is based on acoustics, although modern speech recognition systems heavily rely on models of the language too. In practice, all speech recognition systems do some kind of search, in which different sentences are hypothesized and their probability is computed using the acoustic models and language models. In the end, the hypothesis giving the highest probability is chosen as the recognition output. So in a speech recognition system it is necessary to introduce linguistic restrictions through the use of a language model in the recognition system.

Natural language technology in general and language models in particular are very brittle when moving from one domain to another. Current statistical language models are built from text specific to newspapers and TV or radio broadcasts which has little to do with the everyday use of language by a particular individual. We are investigating means of adapting a general-domain statistical language model to a new domain or user when we have access to limited amounts of sample data from the new domain or user.

The kind of texts that should be used to training a language model depends on the tasks where the model is going to be used. For example, if we want to create a language model to be used in a medical reports recognition task, it should be built using large amounts of medical reports or any other medicine related texts. If we want to create a more generic language model we should use more generic texts like newspaper texts. These generic models can also be used to generate task-adapted models when large amounts of related texts are not available, by interpolating a generic language model with a smaller model generated from specific text. [8]

2. Phoneme separation of words in LM

Phone separation of words in the language model dictionary is one of the main tasks in preparing the language model.

3. Content of LM

A language model comprises two main components: the vocabulary or the lexicon which is a set of words that can be recognized by the system and the grammar which is a set of rules that regulate the way the words of the vocabulary can be arranged into groups and form sentences. The grammar can be made of formal linguistic rules or can be a stochastic model. The linguistic models introduce strong restrictions in allowable sequences of words but can become computational demanding when incorporating in a speech recognition system. They also have the problem of not allowing the form of grammatically incorrect sentences that are often present in spontaneous speech.

4. Out of Vocabulary words

Typically the size of the lexicon is something between 10,000 and 60,000 words. Restricting the recognizer to certain words naturally poses the problem that the words outside the lexicon cannot be recognized correctly. These words are called out-of-vocabulary (OOV) words in

speech recognition literature. Several approaches like expand the vocabulary, split words into smaller word fragments, statistical morphs, etc have been proposed to tackle the problem. [9]

The statistical morphs are found using the Recursive MDL algorithm, which learns a model inspired by the Minimum Description Length (MDL) principle. The basic idea is to run the algorithm on a large text corpus, and the algorithm tries to find a morph lexicon that encodes the corpus efficiently, but is still compact itself. [9] This principle splits words in fragments if the fragments are useful in building other common words. The rarest words end up being split in many fragments, while very common words remain unsplit.

The error rate of a speech recognizer is no less than the percentage of spoken words that are not in its vocabulary (OOV words). So a major part of building a language model is to select a vocabulary that will have maximal coverage on new text spoken to the recognizer. This remains a human intensive effort. A corpus of text is used in conjunction with dictionaries to determine appropriate vocabularies. A tokenizer which is a system that segments text into words is needed. Then a unigram count for all of the spellings that occur in a corpus is determined. Those words that also occur in the dictionary are included. In addition a human screens the most frequent subset of new spellings to determine if they are words.

5. N-Gram LM

An N-Gram model is a type of probabilistic model for predicting the next item in a sequence. N-grams are used in various areas of statistical natural language processing and genetic sequence analysis. An n-gram is a subsequence of n items from a given sequence. The items in sequence can be phonemes, syllables, letters, words or base pairs according to the application. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or "digram"); size 3 is a "trigram"; and size 4 or more is simply called an "n-gram". Some language models built from n-grams are "(n − 1)-order Markov models". Previously, weighted mixtures of word n-gram language models have been used to provide a topic adaptation method for large vocabulary speech recognition [10]. A disadvantage of this method, however, is that these models require large numbers of parameters per topic, which in turn necessitates a large quantity of training data for each topic and associated storage space per topic in the resulting language model. [11]

III EXPERIMENTS

To make the speech recognition experiments we used HTK as our baseline recognizer which is a toolkit for building Hidden Markov Models (HMMs). [12] It was a test of a limited set of words simply a dialing a call using digits and commands. And it gave us expectable results while doing the experiment. It is important to maintain the same microphone volume level, environment while it was in training, for better performance.

IV. RESULTS AND DISCUSSION

This section explains the experiments done and results got for those. To make the speech recognition experiments we used HTK as our baseline recognizer which is a toolkit for building Hidden Markov Models (HMMs). It was a test of a limited set of words simply digits from zero to nine. And it gave us expectable results while doing the experiment. It is important to maintain the same microphone volume level, environment while it was in training, for better performance. For the testing purposes we used the pre-recorded audio files as the input to speech recognition system. It shows around 75% accuracy level with our acoustic model.

Figure 4 Test result with HTK

We defined three criteria for test the system.

1. Isolated word
2. Connected word
3. Continuous speech

We test the system under these criteria and a sample test result can be shown as follows,

Utterance	WER	
	Wav	Microphone
හය බිංදුව	0	29
අට නවය නවය බිංදුව දෙක එක	33.33	31.42
තුන තුන හත නවය	0	17
බිංදුව හය අට හත හතර	32	33.33
හතර අට දෙක හතර හය හත	33.33	29.52
නවය හය නවය අට හතර	20	24
අට එක දෙක තුන දෙක	40	40
නවය හත පහ හතර	0	15
තුන	0	0
දෙක හත එක හතර හත අට නවය දෙක හය	14.29	22.5

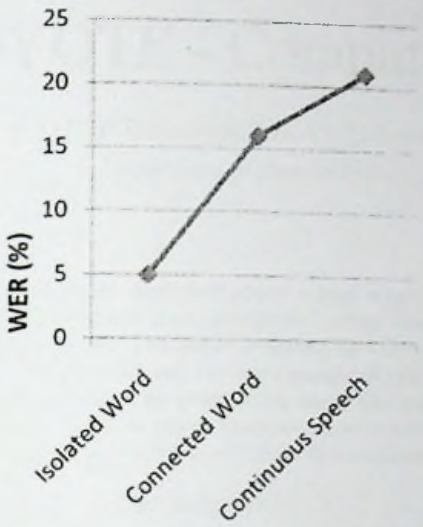

Figure 5 Result comparison

As described in above methodology, we have done the quality enhancement of the pre-recorded voice samples before feeding to the recognizer. There we have promising outcome for the recognition accuracy. Still we are doing furthermore testing on the subject and by the paper doing discussion more on it.

We have come up with a large vocabulary continuous speech recognition system for Sinhala language which is speaker independent and noise robust. To accomplish this we integrated number of state of the art speech technologies pioneered in areas such as acoustic and language model building, searching through lexicon tree and noise filtering. We have initiated a dictionary file, lexicon file and a grammar file for Sinhala language that can be used extensively for speech recognition process.

REFERENCES

[1] Sorin Dusan, Larry R. Rabiner, "On Intergrating Insights from Human Speech Perception into Automatic Speech Recognition", Piscataway, U.S.A .

[2] Jeff Blimes, (2002, January),"What HMMs can do", [Online].Available:https://www.ee.washington.edu/techsite/papers/documents/uweetr-2002-0003.pdf. Accessed:September 2009.

[3] Lawrence Rabiner, Biing-Hwang Juang, (2003), "Fundamentals of Speech Recognition", Prentice-Hall.

[4] Akinobu Lee,Tatsuya Kawahara.Shuji Doshita, "An efficient two-pass search algorithm using word trellis index",Available:http://winnie.kuis.kyoto-u.ac.jp/lab5/bib-e/icslp98-1.pdf. Accessed: August 2009.

[5]. Hwang, M. Y., "Subphonetic Acoustic Modeling for Speaker-IndependentnContinuous Speech Recognition", Ph.D. thesis, Carnegie Mellon University, 1993, Available: http://handle.dtic.mil/100.2/ADA275217 Accessed: August 2009.

[6] On Developing Acoustic Models Using HTK L.J.M. Rothkrantz et al. Available: http://www.kbs.twi.tudelft.nl/docs/MSc/2004/Spaans_Mike/thesis.pdf Accessed: December 2009.

[7] Application of Triphone Clustering in Acoustic Modeling for Continuous Speech Recognition in Bengali Pratyush Banerjee et al. Available: http://www.facweb.iitkgp.ernet.in/~pabitra/paper/icpr08_speech.pdf Accessed: January 2010.

[8] Nuno Souto, Hugo Meinedo & João P. Neto, "Building language models for continuous speech recognition systems". [Online]. Available: http://www.inesc-id.pt/pt/indicadores/Ficheiros/191.pdf Accessed: October 2009

[9] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola & Mikko Kurimo,"Morphologically Motivated Language Models in Speech Recognition" . [Online]. Available: http://www.cis.hut.fi/mcreutz/papers/Hirsimaki05akrr.pdf Accessed: Aug 2009

[10] R. Kneser & V. Steinbiss,(1993)"On the Dynamic Adaptation of Stochastic Language Models" . [Online]. Available: http://www.computer.org/portal/web/csdl/doi/10.1109/ICASSP.1993.319375 Accessed: September 2009

[11] Gareth Moore & Steve Young, "Class-based language model adaptation using mixtures of word-class weights". [Online]. Available: ftp://svr-ftp.eng.cam.ac.uk/pub/reports/moore_icslp00.pdf Accessed: September 2009

[12] Hidden Markov Model Toolkit (HTK) web site. [Online]. Available: http://htk.eng.cam.ac.uk/ Accessed: Oct 2009