Structure of speech recognition system

This chapter presents a fundamental of continuous speech recognition with Hidden Markov Model. The first discussion describes the basic structure of continuous speech recognition system which included five stage; feature analysis, unit matching system, lexical decoding, syntactic analysis and semantic analysis. Then, we present three approaches to speech recognitions which are acoustic phonetic, pattern recognition and artificial intelligent approach. In pattern recognition approach, we only highlight the statistical method since this approach is adopted by Hidden Markov Model and used in this project. In HMM section, we present three basic problems that that are very important in HMM and that from the strength of HMM as a speech recognizer for different languages. In the last section, we discuss about the speech recognition in Standard Malay language and other languages.

Speech recognition process basically allowed human to speak with computer, having a computer to recognize what user is saying, and lastly, doing this in real time. This process fundamentally functions as a pipeline that converts Pulse Code Modulation (PCM) digital audio from a sound card into recognized speech. In Figure 3. 1, the basic structure of speech recognition system is presented. It contains front end unit, acoustic model unit, language model unit and search unit.

Input speech

Acoustic

Front-end

Acoustic Model

P(A| W)

Search

Language Model

P(W)

Recognized Utterance

Figure 3. 1: Block diagram of speech recognizer.

In the recognition process, the speech signal has to be digitized first, before computer can process the signal. Acoustic front ends are used for the purpose of feature extraction to transform the raw speech corpus into more usable information for both training and testing files. In feature extraction, the signal is converted to a sequence of feature vectors based on spectral and temporal measurements. In this project, Mel Frequency Cepstral Coefficients (MFCC) was used. It is the most commonly used acoustic front-end in field of speech recognition.

Two types of file to recognize speech are required by speech recognition engine which are acoustic model and language model. Acoustic model represent sub-word units or often called phonemes. It is created by compiling audio recordings of speech and their transcriptions into statistical representations of the sound that make up each word. The equation that represents acoustic model is P(A| W). Meanwhile, the language model is a file

containing the probabilities of set of words and controls which models are hypothesized. The equation that represents language model is. The following Bayesian formula for speech recognition is used with the objective to minimize the word error rate by maximizing , where refers to acoustic symbols and refers to word sequence.

The most crucial step in speech recognition system is the search step. In this step, many combinations of words must be investigated to find the most probable word sequence. Depending on the chosen criterion, the speech recognition classifications can be carried base on a tree structure as presented in Figure 3. 2

Speech Recognition

Speaking Style

Vocabulary Size

Speaker Mode

Speech Mode

Dictation

Medium

Small

Large

Speaker Independent

Isolated Speech

Spontaneous

Speaker Dependent

Continuous Speech

Speaker Adaptive

Figure 3. 2: Speech recognition classification

Types of Speech Recognition Approach

Basically, there are three approaches to speech recognition, namely the acoustic-phonetic approach, the pattern recognition approach and the artificial intelligence approach.

Acoustic-phonetic Approach

Acoustic phonetic is the earliest approaches to speech recognition based on finding speech sounds and providing appropriate labels to these sounds

(Anusuya and Katti, 2009). This approach draws on the distinctive feature theory first proposed by Jakobson, Fant and Halle in 1952 and later expanded by Chomsky and Hale in 1968. In between 1971 and 1976, this approach has been applied by Advanced Research Projects agency (ARPA) to develop continuous speech recognition system (Wilson, 1987).

The theory behind acoustic-phonetic approach is acoustic phonetics which it assumes that spoken language is divided into phonetic units that are finite and particular. This theory postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that the phonetic units are broadly characterized by a set of properties that are manifest in the speech signal, or its spectrum, over time (Rabiner and Juang, 1991).

The architecture of acoustic-phonetic approach is shown in Figure 3. 3. The first step is the speech analysis system, which provides an appropriate (spectral) representation of the characteristics of the time-varying speech signal. The most common techniques of spectral analysis are discrete Fourier Transform (DFT), Linear Predictive Coding (LPC), and Mel-Scaled Frequency Cepstral Coeffients (MFCC). The second step is the feature-detection stage. The spectral measurements are converted in a parallel fashion to a set of features describing the broad acoustic properties of the various phonetic units, e. g. nasality (nasal resonance), frication (random excitation), format locations (frequencies of the first three resonances), voiced/unvoiced classification (periodic or a periodic excitation), and energy ratios. The third step is the segmentation and labeling phase, in which the system tries to find feature stable regions and then label those regions according to how well the features within that region match those of individual phonetic units

(Ting, 2007). This approach is suitable for implementing applications with semantic and grammatical constraints, such as voice-dictation.

Figure 3. 3: Block diagram of the acoustic-phonetic approach for ASR

(Rabiner & Juang 93)

Pattern Recognition Approach

Pattern recognition is the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns. The four best known approaches for pattern recognition are: (1) Template Matching, (2) Statistical classification, (3) Syntactic or structural matching, and (4) Neural network (Amit and Rama, 2007). In this section, the statistical classification approach will be discussed further.

Unlike the acoustic-phonetic approach, in the statistical pattern recognition approach, speech pattern is not segmented nor checked for its properties. In this approach, each pattern is represented by a set of d features or attributes and is viewed as a d-dimensional feature vector (Anil et al., 2000). Furthermore, the speech patterns are directly inputted into the system and compared with the patterns inputted in the system during training. The performance of the system is sensitive to the amount of training data, speaking environment and transmission characteristics of the medium used to create the speech (Rabiner and Juang, 1991).

test

pattern

Classification

Feature Measurement

Preprocessing

Classification

Training

training

Learning

Preprocessing

Feature Extraction/Selection

pattern

Figure 3. 4: Model for statistical pattern recognition (Anil et al., 2000)

As can be seen in Figure 3. 4, the recognition system is operated in two modes; training mode and classification mode. In the training mode, the feature extraction/selection module finds the appropriate features for representing the input patterns and the classifier is trained to partition the feature space. The feedback path allows a designer to optimize the preprocessing and feature extraction strategies. In the classification mode, the trained classifier assigns the input pattern to one of the pattern classes under consideration based on the measured features (Anil et al., 2000). One of the well-known statistical models in ASR research is the Hidden Markov Models (HMMs) which is used in this project.

Artificial Intelligence Approach.

The artificial intelligence (AI) approach exploits the concepts of both acoustic-phonetic approach and pattern recognition approach. This approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualising, analysing and finally making a decision on the measured acoustic features. The main idea of AI is to collect and employ knowledge from a number of sources for solving the problem in question (Rabiner and Juang, 1993).

There are two approaches to incorporate knowledge source to speech recognition. They are called bottom-up approach and top-down approach. Both approaches are different according to how they tackle the problems. The most commonly used approach is the bottom-up processor which the lowest level processes such as feature extraction or phonetic decoding, precede higher level processes such as lexical decoding or the language model as shown in Figure 3. 5. Meanwhile, in the top-down approach, processor integrates the word hypothesis matching, lexical decoding and syntactic analyses blocks into a consistent framework as shown in Figure 3. 6 (Ting, 2007). Expert systems are used widely with the AI approach.

Figure 3. 5: Bottom-up approach to knowledge integration (Rabiner and Juang 1993)

Figure 3. 6: Top-down approach to knowledge integration (Rabiner and Juang 1993)

Summary of Speech Recognition Approach

The acoustic-phonetic approach has not been widely used in most commercial applications (Rabiner and Juang, 1991). A limited success has been obtained because of the lack of good knowledge of the acoustics phonetics and the related area (Wilson, 1989). Statistical pattern recognition has been used successfully to design a number of commercial recognition systems (Anil et al., 2000). This approach is a popular choice for most ASR system nowadays because it is simple and is computationally feasible to use. As artificial intelligence approach is hybrid of acoustic phonetic approach and pattern recognition approach, it become the most difficult approach to use. This approach had only limited success largely due to the difficulty in quantifying expert knowledge. Another difficulty is the integration of many levels of human knowledge such as phonetics, phonotactics, lexical access, syntax, semantics and pragmatics (Anusuya and Katti, 2009). Three approaches that have been discussed above can be as in Table 3. 1.

Table 3. 1: Speech recognition approaches

Approach	Representation	Recognition Function	Typical Criterion
Acoustic phonetic Approach			
	Phonemes / Segmentation and Labeling.		
		Probabilistic lexical access procedure.	
			Log likelihood ratio.
Statistical Pattern Recognition Approach			
	Features vector		
		Clustering functions (code book).	
		Discriminate functions.	
			Classification error.
Artificial Intelligence Approach			
	Knowledge based		
			Word error probability.

HMM for Speech Recognition

As mentioned in the previous section, statistical pattern recognition is widely used in development of speech recognition.

Three Basic Problems of HMMs

In the development of the HMMs methodology, there are three fundamental problems of interest. The first one is the evaluation, the second one is the decoding and the third one is the learning.

Problem 1 : The Evaluation Problem

Given HMM Î» and a sequence of observations , what is the probability that the observations are generated by the model, ?

In the first problem, given a sequence of observations, the goal is to compute the likeliness of this sequence to be produced by a given HMM. We need to find a way to compare HMMs that best fits the observations (Germain, 2009). This problem can be solved using dynamic programming such as the forward algorithm. It can be used to solve isolated recognition.

Problem 2: The Decoding Problem

Given HMM Î» and a sequence of observations , what is the most likely state sequence in the model that produced the observations?

In the second problem, given a sequence of observations, the goal is to find the more likely sequence of states that could have been generated in the HMM. Thus, we “ recognize” the original information of the underlying Markov process (Germain, 2009). This problem can be solved using Viterbi

algorithm. This problem is related to the continuous recognition and the segmentation.

Problem 3: The Learning Problem

Given HMM Î» and a sequence of observations , how should we adjust the model parameter in order to maximize ?.

In the third problem, given sequences of observations, the goal is to find the optimal HMM where these sequences are the more likely to be produced (Germain, 2009). This problem can be approximated using Baum-Welch algorithm or also known as forward-backward algorithm. The learning problem is used if we want to train HMM for the subsequent use of recognition task by solving the problem first.

According to Germain (2009), these three problems are related to efficient algorithms since they were optimized by using dynamic programming. The two algorithms widely used in recognition are the Viterbi algorithm and the Forward Backward algorithm.

The Strength of HMMs for Speech Recognition

HMM has been widely used in speech recognition due to the strength of the models themselves. Rabiner and Juang (1991) stated that the strength in HMM models lie in two abroad areas, which are its mathematical framework and its implementation structure. They can be summarized as follow:

The inherent statistical (mathematical precise) framework which the ease and availability of training algorithms for estimating the parameters of the models from finite training sets of speech data,

The flexibility of the resulting recognition system in which one can easily change the size, type, or architecture of the models to suit particular words, sounds and so forth, and

The ease of implementation of the overall recognition system.

Another possible reason why HMMs are used in speech recognition is that a speech signal could be viewed as a piece-wise stationary signal or a short-time stationary signal. The HMMs also can be trained automatically and are simple and computationally feasible to use (Holmes and Huckvale, 1994).

The Summary of HMMs

Since the 1970, the idea of machine might talk with humans has been flawed in speech recognition area. The statistical pattern recognition approaches via HMM has become dominant in development of ASR system. Based on three basic problems in HMM as explained in section 3. 4. 1, the new development system can be applied for real-world applications. This is because of the strength of HMM which include the availability of its mathematical framework, simple architecture, feasible to use and the high accuracy for its performance made this model as ideal solution for ASR system. As a result there are a lot of applications that use speech recognition such as machine translation and bio-medicine.

Standard Malay ASRs

According to Ting (2007), the research of speech recognition in Malaysia is still in its infancy stage due limited to small vocabulary, isolated word application and lack of speaker-independency. However, there is a great potential for the application of the speech technology in Malaysia especially

in the context of Malay speech. Currently, Malay language has been applied in speech recognizer commercial product such as teliSpeech Recognizer 2. 0. It shows Malay language has great potential to be applied in many more speech recognizer applications. In this section, several researches on Malay ASRs are presented.

Malay ASR – Research Paper 1

Title: Malay Continuous Speech Recognition Using Density Hidden Markov Model

A study by Ting (2007), aims to solve the constraints of current Malay speech recognizers which are speaker-dependent, small vocabulary and isolated words. A basis study and research on developing a medium vocabulary, speaker independent and Malay continuous speech recognition system are also provided. This study used Continuous Density Hidden Markov Modeling (CDHMM) with mixture densities, which is more capable in modeling inter-speaker acoustic variability compared to other alternative techniques such as Discrete Markov Hidden Model (DHMM), Neural Network and Dynamic Time Warping. A word-based Malay connected word recognition system was being designed and developed by extending the existing Malay isolated word recognition system to Malay continuous speech recognition tasks. Meanwhile, Malay phonetic segmentation and classification experiments were included as a preliminary research in using sub-word model as modeling unit which is needed in developing large vocabulary system. This task will provide basis on developing sub-word unit based Malay medium and large vocabulary continuous speech recognition system.

As the study focus on three major limitations in current Malay speech recognition which are speaker independency, continuous speech and medium and large vocabulary, the following experiments were done.

Speaker Independency

In this experiment, the performances of the use of DHMM and CDHMM with different training algorithms were carried out. The evaluation was performed on Speaker Dependency-multi speaker Malay isolated digit recognition task. The result showed that the recognition accuracy of CDHMM is higher the DHMM as shown in Table 3. 2

Table 3. 2: Comparison of DHMM and CDHMM recognizers with different training algorithm

Ting (2007) states the better result by CDHMM in Speaker Dependency-multi speaker test motivates its usage in Speaker Independency task in the future.

Continuous Speech

In this experiment, the whole-word based CDHMM connected word recognition system for Malay connected digit recognition task was developed. Ting (2007) stated the specification of the system is as follows,

MFCC feature extraction.

CDHMM with multivariate Gaussian mixture densities.

Five state left-to-right whole word models.

Segmental K-mean connected word training procedure with Viterbi/segmental kmean as word model re-estimation algorithm.

Unigram and bi-gram language modeling

Full Viterbi Search for decoding.

For this experiment, the use of segmental K-mean and manual segmentation training procedure on recognition accuracy for Single speaker test and Multi speaker test were carried out. The result shows the word accuracy achieved from using segmental K-mean algorithm is higher than manual segmentation for both tests. Ting (2007) explained that this is because the segmental K-mean algorithm can converge at a more optimum string segmentation than the manual segmentation, and thus generate more reliable and robust model. Meanwhile, the bi-gram model performs better than unigram model in multi speaker test with segmental K-mean training.

Medium and Large Vocabulary

The experiment attempted the use of sub-word unit modeling in Malay phonetic classification and segmentation task on medium vocabulary continuous speech database. Phone model was chosen because the small sizes of such units enable the model being well trained on limited training data. Meanwhile, CDHMM mixture density was used that several maxima in the probability density function can model the contextual variability of the same phoneme. A set of 35 Malay phones was chosen. Experimentations on Malay phoneme classification showed the following results (Ting, 2007):

Incorporation of energy and differential cepstrum increases the classification rates.

Increasing the number of Gaussian mixture components further improve the accuracy.

The accuracy for intra-phone classification is higher than the all phoneme classification.

Greater improvement for consonants compared to vowel category.

Furthermore, HMM based Viterbi forced alignment was used in phonetic segmentation. It used to investigate the effect of using different feature sets and varying number of mixture components. The following results are observed (Ting, 2007):

For small tolerance (5-10ms), HMMs with fewer Gaussians perform better.

For large tolerances (> 20ms) HMMs with more Gaussians perform better generally.

For medium tolerances (15ms), the result shows an intermediate change in segmentation result between small and large tolerances where increasing the Gaussian start to perform better.

As conclusion, Ting (2007) says, the reasonable good phoneme classification accuracy and phonetic segmentation performance enable the extension to Malay phoneme recognition and finally lead to large vocabulary Malay continuous speech recognition.

Malay ASR – Research Paper 2

Title: Isolated Malay Digit Recognition Using Pattern Recognition Fusion of Dynamic Time Warping and Hidden Markov Models

The project by Al-Hadad (2008) presents a pattern recognition fusion method for isolated Malay digit recognition using Dynamic Time Warping (DTW) and HMM. DTW is used to detect the nearest recorded voice while HMM is used to emit new feature vector for each frame according to an emission probability density function associated with that state. The goals of this project are: (1) to increase the accuracy percentage of Malay speech recognition, (2) to develop patterns of reference for the Malay digit in the recognition database using DTW and HMM, and (3) to fuse DTW and HMM using weight mean vector for improving the recognition.

The algorithm is tested on Malay digits from 0 to 9 which are ‘ KOSONG’, ‘ SATU’, ‘ DUA’, ‘ TIGA’, ‘ EMPAT’, ‘ LIMA’, ‘ ENAM’, ‘ TUJUH’, ‘ LAPAN’, and ‘ SEMBILAN’. Each digit is repeated 10 times by 15 male and 15 female speakers. While recording, a speaker will pause for 1 second between each digit. The system begins with input speech, followed by end point detection, framing, normalization, filtering, MFCC, time normalization, and using DTW and HMM to calculate the reference patterns as can be seen in Figure 3. 7. In the last stage, weighted mean vector is used with results from DTW and HMM to get the final decision output. The weight mean vector equation used as follow,

where,

= query recognition rate in HMM test phase,

= query recognition rate in DTW test phase,

= the real time value of recorded speeches and

= weight mean vector.

Figure 3. 7: Block diagram for decision fusion on Malay isolated digit recognition using DTW and HMM

The result shows that the fusion technique can be used to fuse the pattern recognition outputs of DTW and HMM. The refinement normalization by using weight mean vector give better performance with accuracy of 94% compared to accuracy for DTW and HMM, which is 80. 5% and 90. 7% respectively. The results obtained are shown in Table 3. 3.

Table 3. 3: Comparison digit recognition accuracy test result

Summary of Standard Malay ASRs

The Table 3. 4 summarizes the studies in speech recognition for Malay language in Malaysia. All these studies utilized HMM techniques that used to recognize isolated words and continuous speech.

Table 3. 4: Several Malay speech recognition

Developer

Research Title

Overview of Research

Results

Ting Chee-Ming, Sheikh Hussain Shaikh Salleh and A. K. Ariff

2009

Malay Continuous Speech Recognition Using Fast HMM Match algorithm

The research describes the implementation of fast HMM match algorithm in a phoneme-based Malay continuous speech recognition system.

The strategy is decouples the phone HMM state-likelihood computations from the main search which is bounded by syntactical and lexical constraints.

The strategy they used avoids redundant state-likelihood computations of same phone HMM across word models in conventional search.

The result shows, fast decoding algorithm yields a speedup factor of 31. 8 compared to without decoupling, while maintaining word accuracy without loss at 91. 9% for a test-set perplexity of 15. 45 in speaker-dependent mode.

Noraini Seman and Kamaruzaman Jusoff

2008

Acoustic Pronunciation Variations Modeling for Standard Malay Speech Recognition

In this study, two types of pronunciation variations are defined which are complete or phone change and partial or sound change

SM speech database was built using HTK version 3. 2. It contains 4 hours utterance from news broadcast.

Two different approaches are evaluated which are; probabilistic pronunciation variation dictionary to augment the base form lexicon and pronunciation variation information is introduced during the decoding process.

The experiment shows there is not clear boundary that separates a phone change and sound change.

The proposed techniques of handling sound change are not as effective as the methods for handling phone change.

Fadhilah Rosdi and Raja N. Ainon.

2008

Isolated Malay Speech Recognition Using Hidden Markov Models

Automatic isolated word speech recognition for Malay language was developed using HMM.

It focuses on 5 isolated phonemes word structure such as empat (four), lapan (eight), rekod (record), tidak (no), tujuh

(seven) and tutup (close).

The experiment can identify a spoken word at an average rate of 88. 67% which is acceptable for speech recognition.

Tian-Swee Tan, Helbin-Liboh, A. K. Ariff, Chee-Ming Ting and Sheikh Hussain Shaikh Salleh.

2007

Application of Malay Speech Technology in Malay Speech Therapy Assistance Tools

The project is design the training software which assists the therapist to diagnose Malaysian children with stuttering problem.

The speech recognition system utilizes the HMM techniques.

The voice pattern of the normal and stutter children are used to train the HMM model for classifying the problem of speech stuttering.

Malay-Text-to-Speech system and Talking Head are also utilized in this project.

The study shows that the average percentage of correct recognition rate for normal speech is 96% while for the artificial stutter speech is 90%.

Existing Speech Recognition Using HMM for Other Languages

In this section, several existing HMM based speech recognition system for languages other than English are presented. We choose the continuous speech recognizer for language that under resource and highlight the objectives, methods and performance for each of them.

Tamil Speech Recognition

Tamil is a Dravidian language spoken predominantly in the state of Tamilnadu in India and Sri Lanka. A study by Thangarajan et al. (2008), aims to build a small vocabulary word based and a medium vocabulary triphone

based speech recognizers for Tamil language. The Tamil speech recognition was developed in three modules which are dictionary, language model, and acoustic model. They are built on CMU Sphinx-4, the fourth version Sphinx software from the Carnegie Mellon University. This application is a state-of-art HMM based speech recognition which featured feasibility of continuous speech and speaker-independent large-vocabulary recognition. The program is entirely written in Java programming language.

In this study, statistical tri-gram language model were built for 341 words and 1700 phonemes. A total of 22. 5 hours of continuous speech for training and 7. 5 hours of continuous speech for testing were recorded by 25 speakers. For training stage, the acoustic model trainer has been employed by CMU, SphinxTrain for word and triphone based models. Triphone based model follows a generic strategy for training with the following process; (1) Flat-start monophones training, (2) Baum-Welch training of monophones, (3) triphone, (4) creation, (5) training context dependent untied models, (6) building decision tress and parameter sharing, and (7) mixture generation. Meanwhile word based model follows only the first two steps in its training procedure. Table 3. 5 shows, the training parameter for both models. Notice that, for word model, the number of states in HMM is 20 since duration of words is longer than phones.

Table 3. 5: Training parameter

After the end of training stage, SphinxTrain generates the parameter files of the HMM namely, the probability distributions and transitions matrices of all the HMM models. Then, the language model, dictionary and acoustic model

were deployed on Sphinx-4 decoder configured with the components shown in Table 3. 6.

Table 3. 6: Components of ASR

The performance of Tamil speech recognition is tested for both word and triphone models in batch mode with three trails. They are; (1) trained voice on trained sentences, (2) trained voice on test sentences and (3) new voice on test sentences. The result is generated in term of word error and word accuracy. Word errors are categorized into number of insertion, substitutions and deletions and the word accuracy is computed by the following equation:

The accuracy of both models is very high for the trained voice with trained sentence which are 94. 55% for word model and 93. 70% for triphone model. For trained voice on test sentences the results show 71. 05% for word model and 88. 82% for triphone model. Meanwhile the results for new voice on test sentences are 70. 08% for word model and 92. 06% for triphone model. From the observation, Thangarajan (2008) pointed out that, the word error shows a majority of deletions errors in word based model and substitution errors in triphone model. As conclusion, he stated, for medium and large vocabulary in Tamil language, a triphone based approach is best suited for Tamil language.

Uyghur Speech Recognition

Uyghur is an agglutinative language belongs to the Turkic language family, which is grouped to the Altaic languages system. A study by Silamu and Tursun (2009) aimed to develop HMMbased continuous speech recognition system called UASRS using HTK 3. 3 and MS Visual C++ 8. 0.

In this study, the text corpus was collected from various sources which original text corpus includes 30000 sentences. However, according to the phonetic context, they used Greedy algorithm to choose the sentences from original text corpus. From this process 1018 sentences were selected where each sentence mostly includes 5 to 10 words. According to the characteristic of Uyghur language, 5 states HMM topology for monophone or triphone, 3 states HMM topology for silence and short pause, and training acoustic model for each unit by using 54 speakers data were used. The 2gram language model by using the text corpus of Uyghur including original text was built.

The development of speech recognizer consisted of two parts. The first part included acoustic modeling, language modeling and the recognizer based on HTK 3. 3, while the second part included user interface based on Microsoft Visual Studio 2005. Further, the user interface included the input speech data, speech reco