A Comparative Study of Phoneme Recognition using GMM-HMM and ANN based Acoustic Modeling

Farheen Fauziya
Faculty of Engineering & Technology
MRIU, Faridabad

Geeta Nijhawan
Faculty of Engineering & Technology
MRIU, Faridabad

ABSTRACT

Phoneme is the smallest analogous unit of sound employed to form meaningful contrast between utterances. Hidden Markov Model (HMM), Gaussian Mixture model (GMM) and Artificial Neural Network (ANN) have been used in this paper to measure the accuracy and performance of recognition system using toolkits HTK, Sphinx3 and Quicknet, which are freely available for academic works. In this paper the performance of an ASR System based on Accuracy has been compared with TIMIT database.

Keywords:

Automatic Speech Recognition, MFCC, Hidden Markov Model

1. INTRODUCTION

Speech is the most natural form of communication. With the rapid development of communication technologies, a promising speech communication technique for human-to-machine interaction has become the need of the hour because the overall aim of processing speech is to comprehend and to act on spoken language. Automatic speech recognition (ASR) is the core challenge towards the natural human-to-machine communication technology. It is defined as computerized transcription of speech language into text.ASR technology has an immense potential to change the way we interact with machines. Scientists and engineers have been trying to build machines which can understand human voice commands (speech recognition) and vice versa by generating artificial voice which is close to natural human voice (speech synthesis). Although, ASR has been widely researched for more than 50 years, but due to lot of potential in phoneme recognition research area, researchers have been attracted towards ASR as phoneme recognition task and still trying to get better results in real time applications and scenarios. The reduction in performance of speech recognition system comes greatly due to mismatch of an environmental conditions in training and testing data.

The rest of the paper is organized as follows: Section 2 deals with the two modules of the recognition system i.e. feature extraction and feature recognition. Section 3 contains the Methodology of ASR system. Section 4 describes the Acoustic Modeling and Toolkits, Performance measures for an ASR system is discussed in Section 5. Results and conclusions are presented in Section 6 and 7 respectively.

2. THE RECOGNITION SYSTEM

A speech recognition system consists of four blocks as shown in Figure 1: - Feature extraction, Acoustic modeling, language Modeling and Decoder.

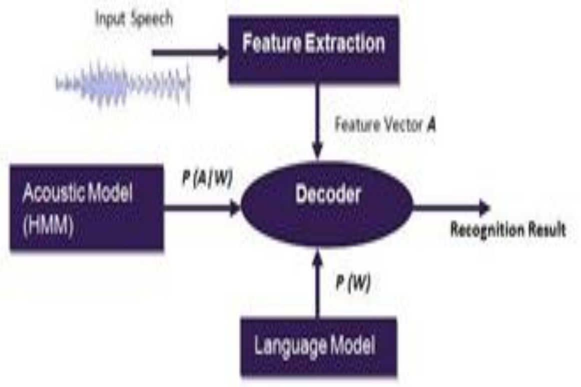

Fig 1: Architecture of an ASR System

In Recognition System mainly two modules can be considered based on the Figure 1. First module of the system is feature extraction which uses standard methods recommended in [2] that are Mel frequency cepstral coefficients (MFCC) based feature .The second module is the modeling. In practice, the modeling stage is subdivided in acoustical and language modeling, both based on HMMs as described in Figure 1. Recognition process starts with capturing the sound waves by a microphone. The electrical signals are converted into the digital signal to make them understandable by Speech system. Speech signal is then converted into discrete sequence of feature vectors, which is assumed to contain only the relevant information about given utterance that is important for its correct recognition. Finally recognition component finds the best match in the knowledge base, for the incoming feature vectors.

3. METHODOLOGY FOR ASR SYSTEM

3.1 Feature Extraction

The goal of the feature vector is to represent the underlying phonetic content of the speech that does not vary with time when same words are spoken. Feature extraction consists of computing representations of Speech signal that are robust to acoustic variation but sensitive to linguistic content. The features include formants, phase spectral information, pitch information and features based on the speech articulators. The features should ideally be compact, distinct and well represented by the acoustic model.An important property of feature extraction is the suppression of information irrelevant for correct classification such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone). Sometimes, however the information conveyed by these feature vectors may be correlated and less discriminative which may slow down the further processing.The most commonly used perception based approach for parameterizations of speech are Mel-frequency Cepstral coefficients (MFCCs) provides some way to get uncorrelated vectors by means of discrete cosine transforms (DCT).

3.2 Pre-emphasis

Many analysis methods focus on those parts of the speech spectrum with the highest intensity. If speech perception were to only need the strongest frequencies of speech, then frequencies above 1 kHz (which have much weaker intensity) would be largely ignored. It is thus clear that some aspects of weak energy at higher frequencies in the audio range. In particular, the center frequencies of the second and third formants are very important, and must be modeled well in most speech analysis methods. To assist feature extraction techniques to properly model formants of different intensities, a pre- processing technique called Pre-emphasis is often used as a first step in speech analysis. This pre-processing raises input speech energy by a variable amount that increases as frequency increases. The amount of pre-emphasis is usually specified by α (a constant).

$$S_{pre}[n] = S[n] + \alpha * S[n-1] \qquad \text{(i)}$$

Where, α=0.95, this is the form of differentiation that boosts the high frequency [5] while de-emphasis does integration.

3.3 MFCC (Mel frequency cepstral coefficients)

MFCC is the most common method for feature extraction in ASR system and can be considered a baseline for performance comparison of feature sets [1].Feature vectors are extracted from frequency spectra of windowed speech frames. This method modify the spectrum to model the frequency resolution of the human ear and Warp the frequency axis such that small differences between frequencies at lower frequencies are given the same importance as larger differences at higher frequencies. Because Human ear has non-uniform resolution i.e. we can detect small changes in frequency at low frequency whereas at high frequencies, only gross differences can be detected. Feature computation must be performed with similar resolution. Since the information in the speech signal is also distributed in a manner matched to human perception .Since the human auditory system becomes less frequency-selective as frequency increases above 1 kHz. This concept also has a direct effect on performance of ASR systems; therefore, the spectrum is warped using a logarithmic Mel scale. The non-linear frequency scale used an approximation to the Mel-frequency scale which is approximately linear frequency scale below 1 kHz and logarithmic and nonlinear for frequencies above 1 kHz [6].

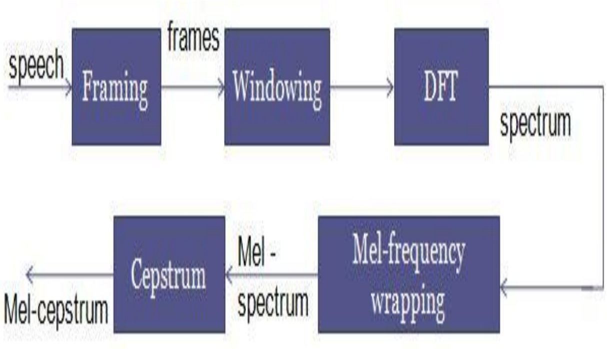

Fig 2: MFCC Workflow

Figure 2–MFCC workflow presents the entire procedure for extracting feature vectors. The input must be transformed into a sequence of acoustic feature vectors, each of which captures a small amount of information within the original waveform. Mathematically Mel scale is described as in equation (ii)

$$mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \ldots\ldots \text{(ii)}$$

The Mel frequency filter bank consist of triangular band pass filter in such a way that lower boundary of one filter is situated at the centre frequency of previous filter and the upper boundaries situated at the centre frequency of the next filter[7].

3.4 Statistical Framework

In the statistical framework, the sequence of words is selected by the recognizer that is more likely to be produced given the observed acoustic evidence [9] out of all valid sequences in the language L.

Let P (W|A) denote the probability that the words W were spoken given that the acoustic evidence A was observed. The recognizer should select the sequence of words W satisfying

$$\hat{W} = \arg\max_{W \in L} P(W \mid A) \ldots\ldots.. \quad \text{(iii)}$$

Since $P(W \mid A)$ is difficult to model directly, Bayes‘rule allows us to rewrite such probability as

$$\hat{W} = \arg\max_{W \in L} \frac{P(A \mid W)P(W)}{P(A)} \ldots\ldots \text{(iv)}$$

Where, P(W) is the probability that the sequence of words W will be uttered and determined by a language model. Since P(A) is independent of W, so it can be ignored and then the maximum a posterior probability (MAP) decoding rule of equation(1) will become

$$\hat{W} = \arg\max_{W \in L} P(A \mid W)P(W) \ldots\ldots \text{(v)}$$

The likelihood P(A|W), is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string .

4. ACOUSTIC MODELING AND TOOLKITS

The process of establishing statistical representations for the feature vector sequences computed from the speech waveform is known as the Acoustic modeling of speech. Each of these statistical representations is assigned a label called a Phoneme. Gaussian Mixture Model (GMM)-Hidden Markov Model (HMM) is one most common type of acoustic models. Other acoustic models include segmental models, super-segmental models (including hidden dynamic models), neural networks, maximum entropy models, and (hidden) conditional random fields, etc.

Acoustic modeling also encompasses "pronunciation modeling", which describes how a sequence or multi-sequences of fundamental speech units (such as phones or phonetic feature) are used to represent larger speech units such as words or phrases which are the object of speech recognition. Acoustic modeling may also include the use of feedback information from the recognizer to reshape the feature vectors of speech in achieving noise robustness in speech recognition. In order to recognize speech, two basic components are usually required in Speech recognition engines. One component is an acoustic model, created by taking audio recordings of speech and their transcriptions and then compiling them into statistical representations of the sounds for words. The other component is called a language model or grammar component, which gives the probabilities of sequences of words. Language models are often used for dictation applications. A special type of language models is regular grammars containing sets of predefined combinations of words, which are used typically in desktop command and control or telephony IVR-type applications. The English language has about 40 distinct sounds that are useful for speech recognition, and thus we have 40 different phonemes. An acoustic model is created by taking a large database of speech (called a speech corpus) and using special training algorithms to create statistical representations for each phoneme in a language. Each phoneme has its own GMM-HMM.

4.1 GMM-HMM

HMM is a statistical modeling techniques with an under-lying doubly stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols [10]. GMM is parametric probability density function represented as a weighted sum of Gaussian component densities proposed by Zolfaghari and Robinson [13]. The technique assumes that a set of Gaussian components can represent a distribution based on the spectral envelope. The GMM parameters are iteratively estimated using the expectation maximization (EM) algorithm. The posterior probabilities of each bin being generated by the target mixtures are estimated, and then these values are used to calculate the Gaussian component parameters for HMM based classifier. The spatial variation is captured by GMM while HMM captures temporal variations.

HMM-GMM based systems can be attributed to the following: (a) its stochastic modeling can take care of the acoustic variations of speech, (b) it can handle time-sequences effectively and (c) it is computationally very efficient. However, HMMs have a number of limitations in modeling speech. Some of these are:(a)the conditional-independence assumption that prevents an HMM from taking the full advantage of the correlation that exists among the frames of a phonetic segment, (b) HMMs trained with maximum likelihood criterion lacks discriminative power, and (c) the awkwardness with which the contextual information are incorporated into HMM systems[14]. The Hidden Markov Models assume a Gaussian Mixture model (with a variable number of clusters) in each of the states of the HMM.A HMM models temporal data in as a sequence of states. States are usually defined as separate GMMs, and their successive usage across time is governed by a transition matrix. The transition matrix is learned from training data and defines the probabilities of moving from one state to another ensuring that the data are optimally explained. Ultimately, what the HMM does is create a sequence of GMM models to explain the input data, thus being sensitive to temporal changes. The parameters of acoustic model in HMM based speech recognition system usually estimated using maximum likelihood estimation. The weakness of MLE lies in that it cannot directly optimize word or phone recognition error rates due to its strong assumption of its sufficient training data and model correctness [15]

4.2 Artificial neural network

The ANN consists of a Multilayer Perceptron (MLP) network whose frame-based outputs represents posterior probabilities of phoneme occurrences and is used as state occupancy probabilities in HMMs.This technology is capable of solving much more complicated recognition tasks, and can handle low quality, noisy data, and speaker independence. The artificial neural networks (ANN), on the other hand, though poor in handling time-sequences, have good pattern discriminative power and can incorporate contextual information rather easily. In the recent year, to overcome the limitation of HMM as mentioned above, researchers are thinking of a hybrid HMM-ANN system with ANN.

4.3 Toolkits

HTK: The Hidden Markov Modeling Toolkit (HTK) is well established portable framework primarily designed for building and manipulating Hidden Markov model (Young S.et al., 2002) [11] and to model time series[8].HTK is primarily designed to build HMM-based systems used for speech processing and speech recognition tools.

Sphinx3: Sphinx is developed by Carnegue Mellon University (CMU). CMU Sphinx is a large vocabulary, speaker independent speech recognition code base and suite of tools.

Quicknet: Quicknet is open source software developed in the Speech Group at the International Computer Science Institute by David Johnson. It is primarily designed for use in speech processing and has been used for tasks other than ASR, including handwriting recognition. It implements most commonly used algorithm that is Multi-layer Perceptron with few layers in statistical pattern recognition system [9].\

5. EVALUATION OF PERFORMANCE OF ASR SYSTEM

To evaluate the performance of ASR systems the word error rate (WER) is very important metrics. For simple recognition systems (e.g., isolated words), the performance is simply the percentage of misrecognized words. However, in continuous speech recognition systems, such measure is not efficient because the sequence of recognized words can contain three types of errors. The first error, known as word substitution, happens when an incorrect word is recognized in place of the correctly spoken word. The second error, known as word deletion, happens when a spoken word is not recognized (i.e., the recognized sentence does not have the spoken word). Finally, the third error, known as word insertion, happens when extra words are estimated by the recognizer (i.e., the recognized sentence contains more words than what actually was spoken). In the following example, the substitutions are bold insertions are underlined, and deletions are denoted as #.

Correct sentence: “Can you bring me a glass of water, please?”
Recognized sentence: “Can you bring # a glass of cold water, **police?”**

To estimate the word error rate (WER), the correct and the recognized sentence must be first aligned. Then the number of substitutions (S), deletions (D), and insertions (I) can be estimated. The WER is defined as

$$WER = \frac{(S+D+I)}{(N)} * 100 \qquad \text{(vi)}$$

Where S =Number of substitutions=deletions and I=insertions and N=Total no of words in the reference.

$$W_{accuracy} = 1 - WER \qquad \text{(vii)}$$

$$W_{accuracy} = \frac{(N-S-D-I)}{(N)} = \left(\frac{H-I}{N}\right) \qquad \text{(viii)}$$

Where, H=the number of correctly recognized words

6. RESULTS

The experiments in this paper rely on the Texas Instruments and Massachusetts Institute of Technology (TIMIT) corpus. TIMIT is a standard data set that is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of ASR systems (John S. Garofolo, 1993).TIMIT corpus is recorded with a high-fidelity microphone in a noise-free environment with 6300 utterances from 630 different speakers of American English. Each utterance is recorded as a 16-bit waveform file sampled at 16 KHz. The entire data set is divided into a set of training speakers and a set of test speakers[4], marked TRAIN to be used to generate an ASR baseline, and TEST that should be unseen by the experiment until the final evaluation. The phoneme recognition accuracy with different Toolkits is presented in table 1 and all the results are in percentage.

Table 1. Phoneme Recognition Accuracy with Different Toolkits

TOOLKIT	GMM2	GMM4	GMM8	GMM16	GMM32
HTK	52.58	57.10	60.88	63.64	66.00
Sphinx3	48.20	52.80	56.70	60.00	62.30
Quicknet	68.39				
Parameters in Quicknet	Num_of_layers=3, Nodes in Hidden layer=1000, Nodes in input=feature dimension,Num_of_outputs=40.				

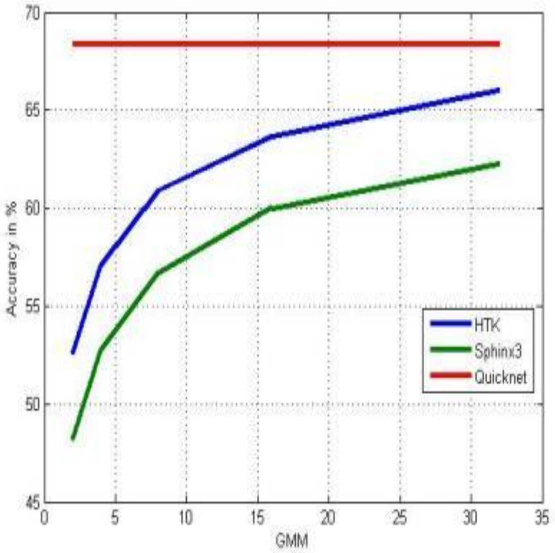

Fig 3: Performance of ASR system based on Accuracy

Fig.3 shows the performance of HTK, Sphinx3 and Quicknet. With the above figure it is clear that HTK gives the better results as compare to Sphinx with the increase of Gaussian value. Both frameworks can be used to develop, train, and test a speech model from TIMIT corpus speech utterance by using Hidden Markov modeling techniques. The uniform line in the above graph is for Quicknet toolkit. Since Quicknet is not a statistical approach so there is no Gaussian. Just to compare the results we have plotted. Table 1 provides performance overview of different toolkits which may help readers for analysis or for choosing the best toolkit for their work (research development) among the techniques and toolkits considered and can be seen that Quicknet is providing better accuracy than HTK and Sphinx.

7. CONCLUSION

Hidden Markov Models(HMM) are the most successful and widely used tool (with the exception of some ANN architectures) for phonetic, syllable and word tokenization, that is, the translation from sampled speech into a time-aligned sequence of linguistic units of the three toolkits used Quicknet provides better result than HTK and Sphinx 3.

In the future work, to improve the performance with real data, more investigations are required on the proper number of mixtures on Gaussian model and on the proper parameter sets. Hybrid method of this both models will be experimented in future for more extensively speaker recognition.

8. REFERENCES

[1] S.B. Davis and P Mermelstein. ”Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions Acoustics Speech and Signal Processing”, 28:357–366, 1980.

[2] Deller J.R., Hansen J.H.L. & Proakis J.G.: Discrete-Time Processing of Speech Signals. IEEE Press, 2000

[3] H. Bourlard and C. J. Wellekens, “Links between Markov models and multilayer Perceptron” *IEEE* Trans. on Pattern Analysis and Machine Intelligence, vol. 12, no. 12, pp. 1167– 1178,1990. Forman, G. 2003.

[4] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of IEEE, 1989

[5] Gold, B., Morgan, N.: Speech and Audio Signal Processing: Processing and Perception of Speech and Music. John Wiley, New York (2000)

[6] Steven,S.Volkmann,J. and Newmann,E., ”A scale for Measurement of Psychological Magnitude Pitch.” Journal of the Acoustical Society of America 8:185-190,1935

[7] Utpal Bhattacharjee, ”A comparative study of LPCC and MFCC feature for the recognition of Assamese phonemes”IJERT,ISSN:2278-0181,Vol.2 issue 1,January-2013

[8] Anant G. Veeravalli,W.D.pan,Reza Adhami,paul G Cox,”Phoneme recognition using Hidden Markov Model,Huntsville Simulation conference,October 2004

[9] International computer science Institute, Speech Group, http://www.icsi.berkeley.edu/groups/speech,

[10] An introduction to Hidden Makov Model,”L.R.Rabiner & B.H.Juang”IEEE ASSP MAGAZINE JANUARY 1986

[11] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland. The HTK Book. Revised for HTK Version 3.2 Dec. 2002. http://htk.eng.cam.ac.uk/

[12] Phil Blunsom, ”Hidden Markov Models”, August 19, 2004

[13] P Zolfaghari and A.J. Robinson. Formant analysis using mixtures of Gaussians. In ProceeSLP, pages 904-907,1996

[14] Anjan Basu and Torbjgrn Svendsen ,”A Time-Frequency Segmental Neural Networks for Phoneme Recognition ”Acoustic Speech and Signal processing, IEEE 1993.

[15] Zaihu PANG et al. Discriminative training of GMM-HMM acoustic model by RPCL learning, Front. Electr. Electron. Eng. China 2011, 6(2): 283–290