PERFORMANCE OF HTK BASED AUTOMATIC SPEECH RECOGNITIONSYSTEM WITH CODED SPEECH

Mounika Jammula,
Assistant Professor,
Department of ECE,
CBIT,Hyderabad ,Telangana, India
Email:jmounika_ece@cbit.ac.in

ABSTRACT

The ASR systems are getting integrated into VoIP and wireless systems for Remote Speech Recognition (RSR). There are many network impediments that affect the accuracy of ASR over a communication network, but the most important happens to be the effects of speech coding. The speech will be transmitted to the recognition servers using different speech coding standards. It is known that the quality of speech will degrade, when it is being encoded/decoded in several phases over a transmission channel, especially when transmitting the voice through the packet data networks.The evaluation of the speech coding effects is done by applying codecson the speech data and testing them over a speech recognition system built using HTK. The evaluation wasperformed using narrowband and wideband codecs. The recognition accuracy for uncoded data and the data coded with wideband codecwas around 95%, whereas for the data coded with narrowband codecs,the accuracy ranged between 80% and 85%.

Index terms- ASR, HTK, TIMIT, MFCC, Baum-Welch re-estimation, Narrow band and wideband codecs.

1. Introduction

Automatic speech recognition (ASR) services are expected to increase on different communication channels because of the naturalness of the speech-based user interfaces. Due this fact, the interactive voice response systems will be replaced by speech, obsoleting the keypad entries and eventually the ASR technology shall be incorporated into the communication systems.

The communication systems such as wireless and VoIP systems use different narrowband and wideband speech codecs for voice compression with various bit rates. Thiscoding process on the speech data, effects the recognition rate at the server side. Though there are many network impediments that affect the accuracy of ASR over a communication network, the most important happens to be speech coding [2]·This paper deals with the evaluation of these speech coding effects on the speech recognition system.

The speech recognition system is built using HTK(HMM Tool kit). The speech data for the training and testing the speech recognition system is taken from TIMIT data base. In this paper, the recognition results of the uncoded test speech data against the results of coded test speech data withdifferent codecs is evaluated. The narrowband and wideband codecs are utilized in the process of testing. Finally the recognition result is compared to appreciate the best codec among the applied ones. The eventual objective is to understand the speech codec which gives the best accuracy.

The next sections in this paper are organized as follows: Section 2 details the Toolkit used for the ASR. Section 3 explains the TIMIT database. Section 4 explains the procedure for training and testing. Section 5 gives the implementation process flow. Section 6 details the speech codecs and section 7 about the experimental environment. Finally section 8 gives the ASR performance results with different speech coding standards.

2. HTK for ASR

The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models. It is developed by the Cambridge university Engineering department. HTK is primarily used for speech recognition research; it can also be used for numerous other applications including speech synthesis, character recognition and DNA sequencing [3].

Speech recognition process involves training and testing. The database, TIMIT is utilized for this purpose.

3. TIMIT Speech Database

The TIMIT database has a total of 6300 speech files (utterances) spoken by 630 people of United States of America. These people belong to different sexes and dialects. Among the 6300 files, 4620 are utilized for training and the 1344 utterances across remaining 1680 speech files are used as test speech data. In the making of corpus it is taken care that no speaker appears in both training and testing portions and also all the dialects regions are represented in both the subsets, with at least 1 male and 1 female speaker from each dialect[4]. The speech data utilized is sampled at 16 kHz and hence the models built are 16 kHz models.

4. Procedure for ASR Training and Testing

The speech recognition system as discussed earlier involves two processes namely training and testing. The training involves feature extraction from the training data and the corresponding model building. The testing phase is where test speech data is applied to the models and the recognition accuracy is obtained. All these processes are done with the help of individual modules available in the HTK [3].

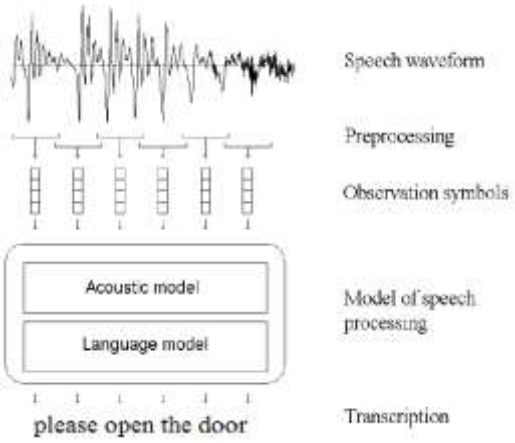

Figure 1: Overview of the speech recognition [1]

The process of building an HTK based speech recognition system is detailed in the sections below.

4.1 Feature extraction (AcousticPreprocessing)

As shown in figure 1, the first step in ASR is the extraction of acoustic features. The aim of the feature extraction process is to translate the information contained in acoustic signals into a data representation that is suitable for statistical modeling and likelihood calculation. ASR requires acoustic features that represent reliable phonetic information consistently, i.e. features which describe the distinctive properties of speech sounds efficiently and that are reproducible over many tokens [1].

There are different methods that can be applied for parametrically representing the speech signal in the speech recognition task and here Mel Frequency CepstralCoefficients (MFCC) are utilized. Hence feature vectors are formed for the audio data in HTK and each of these feature vectors comprises 39 MFCCs (12 MFCCs, energy, deltas of the 12 MFCCs, delta of the energy, delta-deltas of the 12 MFCCs and delta-delta of the energy). Thus the total number of elements in the feature vector leads to a count of 39. A configuration file is given to the HTK which details the source format, output format, and other parameters required for the feature extraction process.

4.2 Acoustic modeling

In the course of building a speech recognition system, the primary task is the training of acoustic models. The training is done using the audio recordings of speech and their transcriptions (text scripts) and then compiling them into a statistical representation of sounds which make up words. Thus for all possible combinations of word strings **W** and observation sequences **O** (sequence of feature vectors)**,** the probability $P(\mathbf{O}|\mathbf{W})$ must be available. But, practically this number of combinations is too large to permit a look up and the combinations become even more massive in the case of continuous speech [1]. Hence continuous speech recognition will consider the phonemes as the basic unit for acoustic modeling.The most famously used modeling technique is the HMMs. Thus a unit of HMM represents a phoneme. There are only 40 to 70 unique phonemes in any language when stress levels are ignored and if the stress levels are considered, around 75 unique phonemes may occur. Hence HMMs based on phone models can be adequately trained.

While testing, these models are used to link the observed features of the speech signals with the expected phonemes of a sentence.

4.2.1 HMM Topology

A 5 state model for each phoneme unit is built where 3 states are emitting and the 2 remaining are starting and ending states, which are non-emitting. Each state is a representation of feature vectorof a 39-dimensional space belonging to that state. Now, in HMMs in ASR, each emitting state is characterized by a Gaussian Mixture component. To represent all these vectors by a single Gaussian mixture, we need a 39-dimensional mean vector and a 39-dimensional variance vector for every Gaussian; assuming cross covariance between the feature elements as zero. Now for a given feature vector, we can find out the probability of that feature vector belonging to a particular state (generally termed as observation probability) using GMM of that state. A square matrix called a transition probability matrix is represented with a size equivalent to the number of states, which defines the likelihood of staying in the same state or transition to any other state.

4.2.2Transcription

The transcription creation is the process of grammar preparation from the text data (utterances) corresponding to the data available in the audio format. The first task is to create the word list and then a phonetic-dictionaryis generated for the word list. Along with the dictionary, phonemes that make up the words are written to a separate text file. As there can be repeated phones, only unique ones are considered and they are hence referred as monophones.

4.2.3 Language model

Language model is the single largest component trained on billions of words developed for detecting the connections between the words in a sentence with the help of pronunciation dictionary [1]. ASR systems utilize *n*-gram language models to guide the search for correct word sequence by predicting the likelihood of the *n*th word on the basis of the $n-1$ preceding words.

There are two kinds of acoustic models that can be built using HTK, namely word internal and Cross word models. Correspondingly two kinds of language models are required; for the former type of acoustic model, a word network is

formed, while for the later an n-gram language model isnecessary.

5. Implementationof ASR System using HTK

After the feature extraction, flat start initialization is done and the estimation is performed for the first time, where the initial monophone models are trained. The re-estimation is performed using Baum-Welch algorithm [1].

Speech Codec	Sampling Frequency (kHz)	Bit Rate (kbps)
G.711	8	64
G.729	8	8
G.722	16	64

5.1 Baum-Welch re-estimation

The flat start initialization is where initially the parameters of a HMM are determined on a rough presumption. Once this is done, more accurate parameters are found by applying Baum-Welch re-estimation formulae. The essential problem is to estimate the means and variances of a HMM in which each state output distribution is a single component Gaussian. But there are multiple states and there is no direct assignment of observationvectors to individual states because of the underlying state sequence which is unknown. There is some approximate assignment of vectors to states.

In brief the process goes like this; the training observation vectors are first divided equally amongst the model states and some initial values are given for the mean and variance of each state. It then finds the maximum likelihood state sequence using the Viterbi algorithm,which reassigns the observation vectors to states and re-estimatedagain to get better initial values. This process is repeated until the estimates do not change.

5.2 Context dependent models

The monophone models built can be utilized for testing but in order to increase the recognition accuracy, context dependent triphone HMMs are trained.Given a set of monophone HMMs, the final stage of model building is to create context-dependent triphone HMMs. This is done in two steps. Firstly, the monophone transcriptions are converted totriphone transcriptions and a set of triphone models are created by copying the monophonesandthen,re-estimated. Secondly, similar acoustic states of these triphones are tied to ensure that all statedistributions can be robustly estimated. Context-dependent triphones can be made by simply cloning monophones and then re-estimating using triphonetranscriptions.At this juncture i.e. while building the context dependent models, a decision is made whether to build word internal models or cross-word models.

If word internal triphones are to be trained, then word boundaries in the training transcriptions are marked explicitly by a short pause which is later deleted. If cross-word models are to be built, then word boundaries in the training data can be ignored and all monophone labels can be converted to triphones. The word internal models are tested using the Viterbi decoder along with the dictionary and the word network. A module/command HDecode is used to test the cross-word models and here the n-gram language model is applied along with the acoustic models.

6. Standard Speech Codecs Considered

The following popularly used speech codecs in VoIP and GSM wireless networks standardized by ITU-T [5] are considered for the analysis. Summary of the different sampling and bit rates for these codecs is given in Table 1.

TABLE1: NB AND WB SPEECH CODECS WITH DIFFERENT SAMPLING AND BIT RATES

7. Experimental Environment

The following section gives the details of the speech codecs, speech database, ASR toolkits and the HMM creation and recognition setups for the testing.

7.1. NB and WB Speech Codecs

The c-source code for all the above speech codecs referred in the table 1,is downloaded from ITU- website. The c-code is compiled in Linux environment to create the executables for encoders and decoders separately for analysis.

All the above speech codecs are evaluated with the HTK version 3.4.1, where the HMMs are generated using 16-kHz speech data. Using HTK, two kinds of Context Dependent (CD) Tri-phone HMMs can be built; Word Internal models and Cross word models. In the present paper, each model is created with 5-states per HMM including 2 null states, with each state modeled by 8 Gaussians and 16 Gaussians for word Internal and crossword models respectively. Language Model is created as Bi-gram Models (word network) and also trigram models, where the former was used in testing word internal models and the later for cross word models.

The feature extraction process is performed, where 39 MFCC features (1-Energy and 12-MFCC, and their DELTA and double DELTA values), are generated (for training and testing) per frame with 40 Mel Filter bands with 50Hz-6800Hz for wideband and 130Hz-3400Hz for narrowband frequencies.

The original speech files (4620 speech files for training) in the TIMIT database are sampled at 16-kHz. Using this speech database, during ASR training phase, HMMs are created for the 16-kHz with corresponding ASR configuration parameters.

7.2. Speech Recognition Setup

Testing of WB Data with 16-kHz HMMs:
For wideband codec analysis, 16-kHz speech based HMMs are used [2]. The encoded-decoded data (which simulates the channel) from the wideband codecs is directly given to the

ASR system with 16 kHz HMMs (uncoded) for recognition as shown in Figure 2.

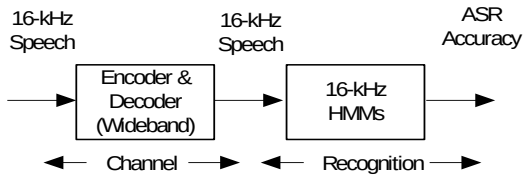

Figure 2: Recognition with 16-kHz trained models (HMM) for wideband codecs

Testing of NB Data with 16-kHz HMMs:
All the narrowband codecs work with 8-kHz sampling rates only. Figure3 shows the procedure for testing narrowband codecs with 16-kHz trained models (16-kHz HMMs).The original 16-kHz speech data is down sampled to 8-kHz first and then this data is encoded-decoded with the respective narrowband codec. The decoded data is up-sampled back to 16-kHz before the recognition analysis with 16-kHz speech based HMMs.

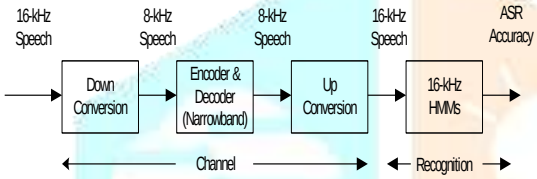

Figure 3: Recognition with 16-kHz trained models (HMM) for narrowband codecs

8. RESULTS AND ANALYSIS

ASR Accuracy Rate is measured using the following approach [3]:

- The percentage correct is given as

$$\text{Percentage correct} = \frac{(N - D - S)}{N} \times 100\ \%$$

  - Where N represents the total number of labels if the sentences are considered for calculation of percentage correct, and it represents words if percentage correct for words is to be calculated.
  - D is the number of deletion errors
  - S is the number of substitution errors.

- The percentage accuracy is given as

$$\text{Percentage accuracy} = \frac{(N - D - S - I)}{N} \times 100\ \% \qquad (3.2)$$

Where I is the number of Insertion errorsPercentage accuracy is calculated considering the Insertion errors.

8.1. Performance of ASR

Table-2 shows the ASR word recognition accuracies for the different codecs with different bit rates for 16-kHz HMMs.

TABLE 2: ASR RECOGNITION ACCURACIES

Test speech data	Result (%)
Uncoded(WI)	95.39
G.711(WI)	85.96
G.729(WI)	80.04
G.722(WI)	95.32
Uncoded (CW)	54.45

- WI – Word Internal Models
- CW – Cross Word Models

9. CONCLUSIONS

The speech recognition system was developed using HTK for 16-kHz ASR models. The uncoded data when tested using a bigram language model (word network), has given a word recognition accuracy of 95.39%.

The same 16-kHz ASR models were used for testing the data coded by narrowband and wideband speech codecs. The narrow band speech codecs, G.711 and G.729, and the wideband speech codec G.722 were tested for performance analysis of the ASR. The ASR performance for both the narrow band codecs was closer, despite the fact where the G.729 codec has a compression rate of 8 kbps when compared to the G.711 codec @ 64kbps. The recognition accuracy for the wideband codecs was better when compared to the narrowband ones. The reason behind this observation happens to be the 8kHz sampling rate and the lower bit rate in narrow band codecs.Furthermore, the recognition accuracy of the wideband codecs almost equaled to that of the uncoded speech data, where the former one turned out to be 95.32% and the later one, 95.39%.

The results observed for CW models are poor when compared to WI models with HTK. The recognition accuracy employing the CW models was evaluated only for uncoded data and it turned to be 54.45%.

10. FUTURE SCOPE

Word internal Models:

The HTK implementation is limited to the bigram language models while applying for the word internal models. The ASR system can have increased recognition accuracies if this limitation is overcome. The testing process is quite slow when these models are employed; hence the reason behind this can be examined further such that the process can be speeded up.

Cross word models:

The recognition accuracy obtained while employing the cross-word models is less when compared to the Word internal models with both bigram and trigram language

models. Therefore, the reason behind this can be investigated further and also the cross-word models can further be improved for getting recognition accuracy on par with the word internal models.

The work can be further extended to verify the ASR performance for more codecs and also considering other network impediments like packet drop and noise conditions. The MOS values can be calculated under the effects of these impediments and the observations can be made for the same. The delay variations, tandeming and trans-coding effects of different wireline and wireless codecs in communication networks can also be studied further.

11. REFERENCES

[1] Ir. P. Wiggers and Dr. drs. L.J.M. Rothkrantz, "Automatic Speech Recognition Using Hidden Markov Models", Delft University of Technology,MS Course work, September 2003.
[2] A.V.Ramana, P. Laxminarayana and P. Mythilisharan, "Investigation of ASR Recognition Performance and Mean Opinion Scores for Different Standard Speech and Audio Codecs", IETE Journal of Research, March-April 2012, Volume 58, Issue 2, pp. 121-129
[3] Steve Young,Gunnar Evermann, "HTK 3.4.1 Tutorial", March 2009.
[4] TIMIT speech database, http://www.ldc.upenn.edu/
[5] ITU-T Recommendations, http://www.itu.int