homework #7

Recognition using ISIP Decoder

EE 8993: Fundamentals of Speech Recognition

December 6, 1998

submitted to:

Dr. Joseph Picone

submitted by:

Suresh Balakrishnama

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
MS 39762, USA
Email: balakris@isip.msstate.edu

1. Introduction

Speech to text is an important technology for extraction of information from audio data. Speech recognition finds an important application for automated telephone transactions, voice dictation for workstations. ISIP recognizer, also known as ISIP decoder, is free software available in ISIP's public domain that supports cross-word decoding of triphones and rescoring the trigram grammar lattices.

2. Problem Description

This assignment involves recognition of telephone digits using ISIP recognizer.Data should consist of four, seven and ten digits spoken with many constraints like spoken at slow or fast rate. The acoustic model will be a context-dependent phone model and the language model will be hand-tailored based on the input data. A non-real time demo consists of shell script wherein the user can enter the required number of frames of data for processing and obtain the output at every frame.
Mathematically, the aim of recognizer can be illustrated by equation

$$\hat{W} = \arg max P(W)P(A/W) \quad (1)$$

where $P(W/A) = \frac{P(W)P(A/W)}{P(A)}$ is Bayes' formula of probability theory. $P(W)$ is the probability that the word string W will be uttered, $P(A/W)$ is the probability that when the speaker says W the acoustic evidence A will be observed, and $P(A)$ is the average probability that A will be observed.

3. Components of Recognizer

A speech recognizer consists of three main subsystems: an acoustic model which converts the signal to a sequence of feature vectors, a language model which predicts the next word given a sequence of previously recognized words and a hypothesis search engine which finds the most probable sequence of words given a set of feature vectors. All subsystem are explained in detail in further sections:

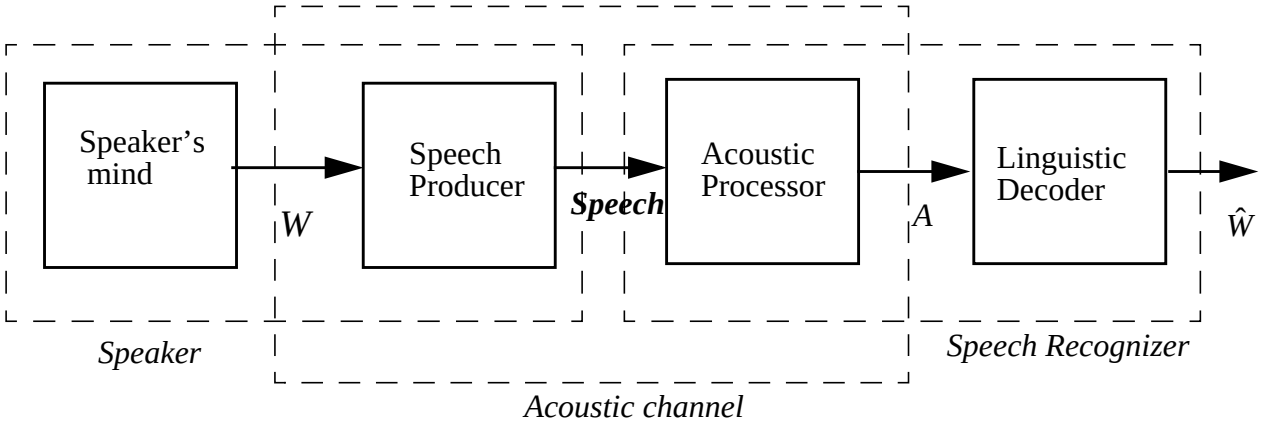

Figure 1 : Block diagram showing components of a speech recognizer

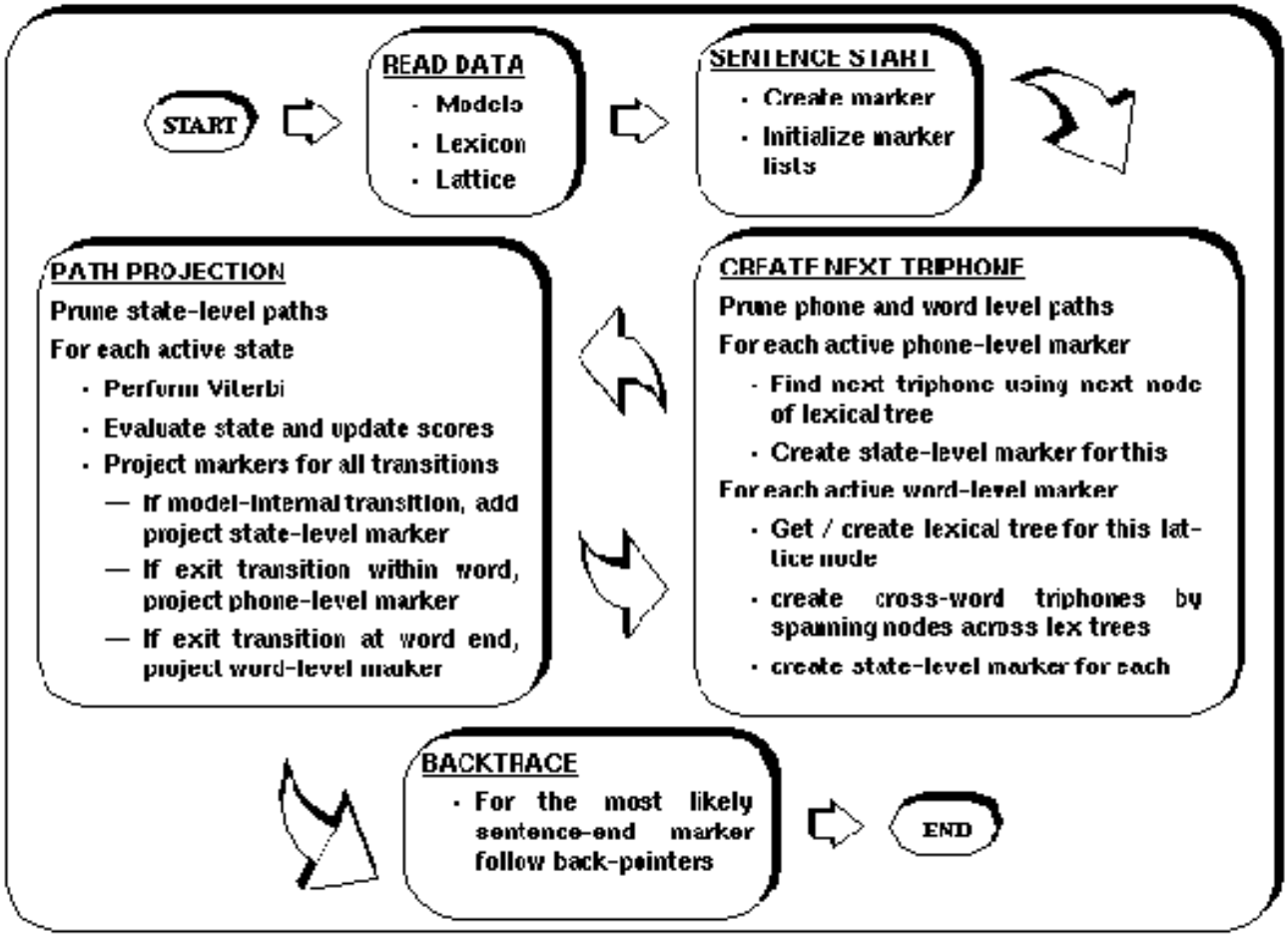

Figure 2 : Flowchart exhibiting processes involved in trace projection class of ISIPs decoder

3.1. Acoustic Model

When the speech recognition system is presented with the acoustic signal, the system uses various signal processing and feature extraction techniques to transform the input speech into a time-series of feature vectors that is a suitable representation for the subsequent statistical phonetic evaluation of data. This is often called the acoustic front-end of the speech recognition system. The speech samples are usually processed in frames of typically 10-15 ms duration and overlapping windows usually 25-30 ms long. The most popularly used features are the mel-frequency cepstral coefficients, delta, delta-cepstrum coefficients and energy, along with their first and second order temporal derivatives.

In most of large vocabulary speech recognition systems, comparator is omitted and the signal processor outputs σ_i are directly handled by the recognizer. The signal processor outputs constitute the observable symbols a_i. In equation (1) the recognizer needs to determine the value of $P(A/W)$, the probability that when the speaker uttered the word sequence W the acoustic processor produced the

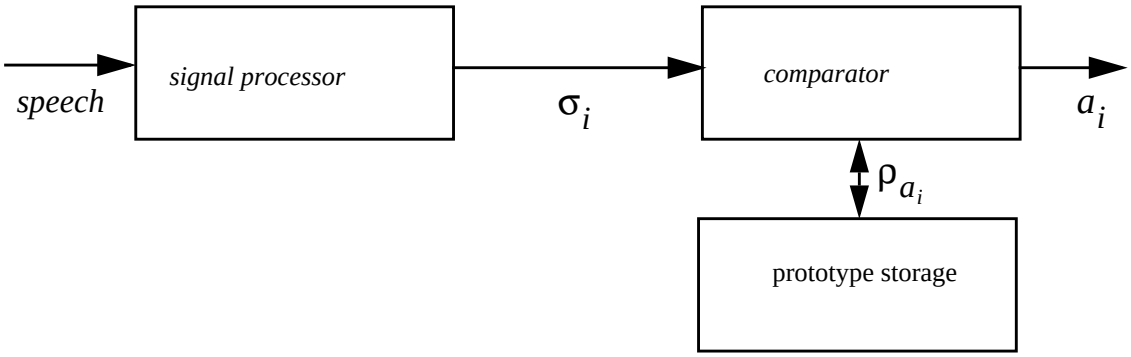

Figure 3 :Schematic diagram of a acoustic processor

data A. The number of different possible values of pairing W with A is large and complex for computational purposes. The total process for modeling involves the way the speaker pronounces the words of W, along with other environmental characteristics and the acoustic processing performed by the front end. The acoustic model most widely applied in speech recognizer is the Hidden Markov Model(HMM). HMM involves phonetic classification wherein the feature vectors are evaluated using weighted mixtures of multivariate Gaussian probability density functions. These density functions are conditioned by the context-dependent phonemes i.e given the forward and backward phonemes what is the probability that a given phoneme would occur.

3.2. Language Model

The selection of a language model has a significant influence on the performance of a speech recognition system. Language model is an important subsystem of speech recognizer since it provides adequate information about the occurrence of words and constraints occurring for different sequence of words. Design of language model should be incorporated to accommodate all grammatical constraints, accent and style of the speaker, and also, the size of language model is very important for real time computational purposes.

In equation (1) the priori probability $P(W)$ can be computed using Bayes' formula for decomposition i.e., we need to compute the probability for every word string W. Mathematically, this can be expressed as:

$$P(W) = \left(\prod_{i=1}^{n} P((w_i / w_1) \ldots\ldots w_{i-1}) \right) \quad (2)$$

The recognizer thus determines estimates of the probabilities $P((w_i / w_1) \ldots\ldots w_{i-1})$. The choice of a word w_i depends on the equivalence classes $\varnothing(w_1 \ldots\ldots w_{i-1})$. Thus, the priori

probability becomes:

$$P(W) = \left(\prod_{i=1}^{n} P((w_i / w_1) \ldots\ldots w_{i-1}) \right) \quad (3)$$

The language model comprises of determining the appropriate equivalence classification $\varnothing$ and a method of estimating the probabilities $P((w_i / w_1) \ldots\ldots w_{i-1})$.

3.3. Hypothesis search

The decoder strategy involves finding the most likely word given the language model, acoustic model and the a spoken utterance. Hypothesis search involves implementation of a search engine to determine the most probable word to follow another word. Two most common search algorithms are - Viterbi search and N-Best search. The main practical approach of any type of search algorithms should be reuse of computations, dynamic memory allocation and as the computations proceed pruning to be applied to shorten the partial hypotheses and extract out the hypotheses with extremely low probability values.

4. Implementation

The main task of this experiment was to use the ISIP decoder to recognize the four, seven and ten telephone digits recorded from a DAT machine. The software is available from ISIPs public domain.

1. Obtain the speech file. The audio to be recognized were obtained from recording on a DAT machine in raw format at a sampling rate of 8 KHz. Speech files for four, seven and ten digits were obtained. The unix command narecord was used to obtain the speech file in raw format.

   narecord -s 8000 9611_four.raw

2. As a second step, features had to be extracted from the speech file. The feature extraction program available as freeware [1] was used for this purpose. The software had utilities for extraction of different types of features. Since the mel-cepstrum, delta, delta-delta and energy features are most widely used as input data for speech recognizer, the utility programs corresponding to these features were used. These utility programs required speech file to be in linear and accepts in wav format. A shell script was used to convert the raw file to NIST wav format. Utility program *cparam* was used to extract features for all available frames and in required dimensions.

   cparam -m -w 25 -p 12 -d -g -e -H NIST 9611_four.wav 9611_four.mfcc

3. cview utility program was used to view the feature values and print them in format accepted by the recognizer. ISIP recognizer accepts input values in 39 dimensions with all values in a single line and every frame of data separated by a newline.

   cview -h -n 39 9611_four.mfcc > 9611_four.dat

4. Next step involves applying decoder's trace projection class for recognizing the spoken telephonic data. The primary data structure used by the ISIP decoder to propagate path information along the frames is called a *trace*. The decoder uses two sets of traces - one at the phone level and the other at state level. The trace projection version utilized for this experiment accepts grammar as input and can be run in demo mode wherein the user can key in number of frames to be processed and the output can be viewed. The decoder is supplied a list of files holding the various system data through a parameter file as one of the command line arguments.The parameter file consists of following model files:

monophones	a list of the basic monophones that constitute the acoustic context
lexicon	a list of all the words and pronunciations used for the current application
transitions	an indexed list of various state transition matrices for the HMMs
states	state parameters for each HMM state
models	a list of HMMs with index pointers to the states
triphones	this is a list of all the possible triphones mapped tothe corresponding constituent monophones, and an index to the correct HMM.

The decoder reads the data from these files provided using the parameter file. Apart from these, required input mfcc file and grammar file are also inputted through the parameter file to the decoder. Input and grammar files are changed based on the speech file (four, seven or ten digits) to be recognized. Command used for obtaining the text output is:

nice-19 /ftp/pub/resources/courses/ece_8993_speech/homework/1998/utilities/decoder/trace_projection/bin/i386_SunOS_5.6/trace_projector -p data/inpatients/params.text -n 5 -c 3 -g 2 -demo

After this command the desired number of frames can be keyed in and the output can be accordingly viewed. Keying the total number of frames would give the entire text output. The real time computation usually involves less than a minute per frame.

5. Results

Most of the software was written in C and C++. Shell scripts were also used for manipulating the speech data file to different formats. All the software and utility programs along with the output can be obtained from ISIPs ftp directory: */ftp/pub/resources/courses/ece_8993_speech/homework/1998/problem_07/balakrishnama/*

Output from speech recognizer is obtained in form of text wherein the number of frames computed, starting and ending frame, duration taken and other parameters can be observed.

```
Time = 1080 New = 554 Deleted = 434 Total = 126989
# ZERO 1080 2 -53945.222763 : [SIL] 2 0 -129.430745
1080 1077 -53945.222763 r-ow+sil
1077 0009 -53745.586899 iy-r+ow
0009 0006 -569.370379 z-iy+r
0006 0002 -374.795671 sil-z+iy
0002 0000 -129.430745 sil

# ZERO1 1080 2 -53945.222763 : [SIL] 2 0 -129.430745
1080 1077 -53945.222763 r-ow+sil
1077 0009 -53745.586899 iy-r+ow
0009 0006 -569.370379 z-iy+r
0006 0002 -374.795671 sil-z+iy
0002 0000 -129.430745 sil

# ZERO2 1080 2 -53945.222763 : [SIL] 2 0 -129.430745
1080 1077 -53945.222763 r-ow+sil
1077 0009 -53745.586899 iy-r+ow
0009 0006 -569.370379 z-iy+r
0006 0002 -374.795671 sil-z+iy
0002 0000 -129.430745 sil
```

Figure 4 : Output obtained from a speech recognizer for spoken data of four digit telephone numbers

Figure 4 shows the output (text) obtained from the evaluation using ISIP decoder. The output shows the time factor which illustrates number of frames of data processed and the number of traces deleted and created for forward and backward search. The alphabetic parameters are the required output, read from left to right. Following the alphabetic parameters are the starting and ending frame of each trace and the corresponding triphones obtained during each search. The four digit telephone number recorded was 9611 with silence in between every digit, though the recognizer did not output the exact digits but the silences were correctly recognized. The reason behind this inaccuracy is that the decoder is trained on telephonic speech data and the parameters are set up according to the telephonic data. Since our input was recorded from a DAT machine the system is not trained for this type of data and hence the model values are quite different which causes the inefficiency.

6. Acknowledgments

I wish to thank Aravind Ganapathiraju and Neeraj Deshmukh, ISIP Decoder team for extending constant support on this experiment, by providing required software and guidance throughout the experiment.

7. References

[1] P.Loizou, “Feature Extraction program for Speech Recognition”, University of Arkansas at Little Rock, May 1998.

[2] N.Deshmukh, A.Ganapathiraju, J.Hamaker and J.Picone, “DoD Decoder Report”, Institute for Signal and Information Processing, Mississippi State University, August 1998.