homework solutions for:

Homework #7: Phone Number Recognition

submitted to:

Dr. Joseph Picone
ECE 8993 Fundamentals of Speech Recognition

June 3, 1998

submitted by:

Jonathan Hamaker

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
Box 9571, 216 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-8335, Fax: 601-325-3149
Email: hamaker@isip.msstate.edu

Introduction

In previous assignments we have tackled topics dealing with front-ends, HMM modeling, and classification. In this assignment we put a good number of those concepts together to experiment with a full speech recognition system (absent the training process). We integrate an audio input system, a feature extraction system, a decoder, and a grammar-specific post-processing system for the recognition of phone numbers.

Problem Statement

Using the ISIP recognizer [1], build a system that recognizes spoken telephone numbers. The system must accommodate 4, 7, and 10 digit strings. The system must use as many constraints about telephone numbers as possible. For acoustic models, use the ISIP context-dependent phone models currently packaged as part of the ISIP recognizer demo. The system must also have its own language model and an interface to an audio system.

Methodology

The flowchart for the phone number system demo is shown in Figure 1. We use the DAT machine audio interface (*narecord*) and a signal detection system for recording the input data. The signal detector is optimized for a certain type of speech and is prone to failure so the user is also limited to 20 seconds for input of the telephone number. This provides more than enough time for the speaker to input a ten digit number. The raw data is converted to MFCC format files using the *cparam* and *cview* programs. These MFCC files are created using a 10 msec frame and 25 msec window. In all, there are 12 mel-scaled cepstral coefficients and log energy plus delta features, and delta-delta features for each frame of data. These steps constitute what is commonly referred to as the front-end of a speech recognition system.

Next is the actual recognition portion of the process. In this phase, we use the ISIP recognizer [1] to decode the utterance and a post processor to determine if the string of numbers spoken is a valid phone number. For this task, we use a set of crossword triphone models trained on the OGI Alphadigit corpus. We use a simple bigram grammar for the decoding process as shown in Figure 2 as this is what the current version of the decoder is limited to. With a finished version of the decoder we would have used a compiled grammar which had the rules for telephone numbers compiled into it. Since this type of grammar decoding is not available to us at this time, we use a utility to post-process the decoder output to determine if it is a valid phone number. The algorithm for determining a successful phone number is shown in Figure 3.

Results

The results for this experiment are abysmal. The sentence error rate is 100% and the word error rate is well over 100% using standard NIST scoring. The insertion rate for words is the key issue causing the poor results. For a sentence, “EIGHT THREE ONE ZERO”, spoken at a normal pace, the recognizer hypothesized “SIX ZERO OH ONE OH THREE EIGHT EIGHT FIVE SIX”. There are a few reasons that could obviously cause this problem. The first is a mismatch in speaking and channel conditions. The models were trained on telephone quality data while the test samples were taken over a high quality audio system. The second, and perhaps most

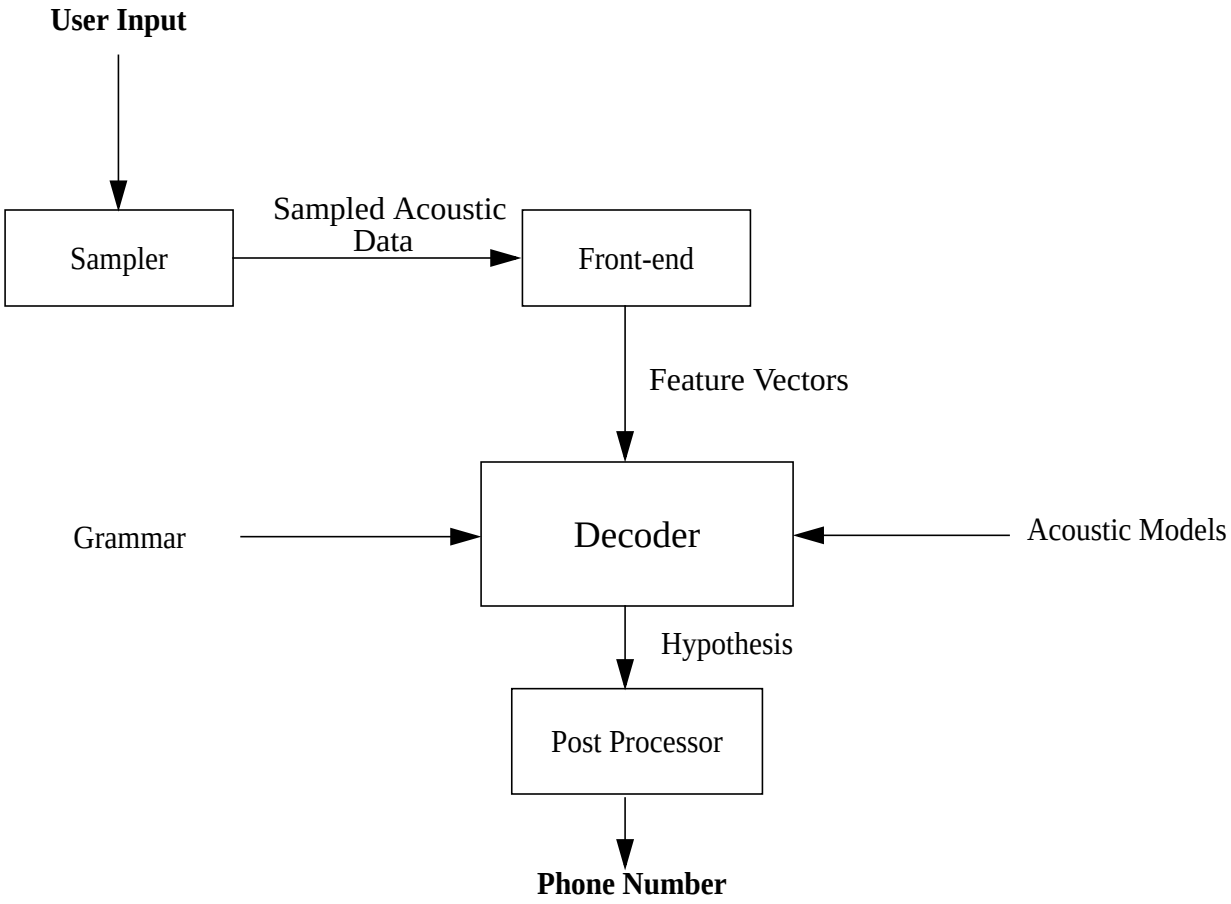

Figure 1. Data flow for telephone recognition experiments

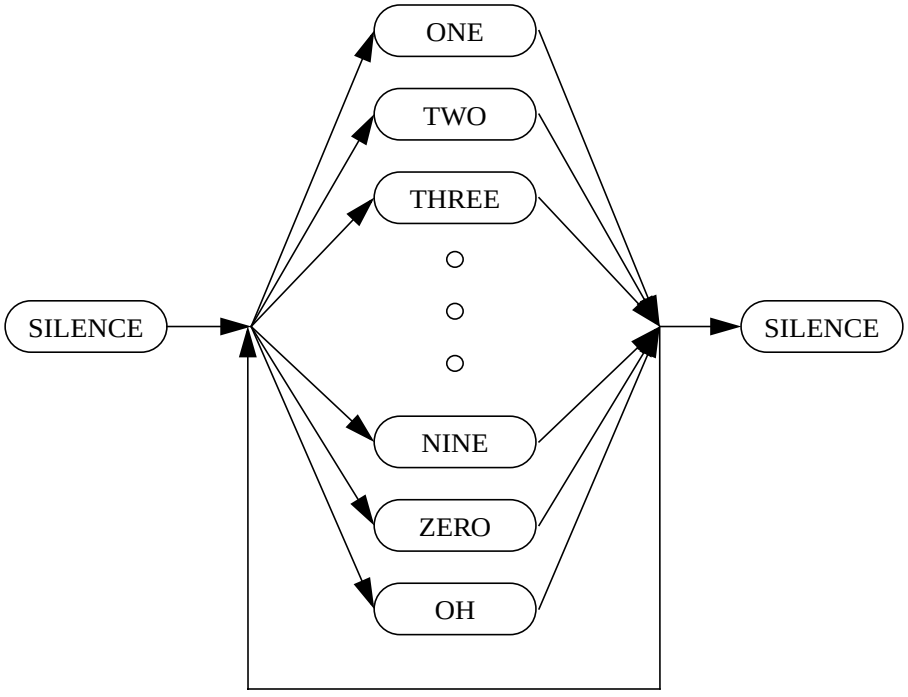

Figure 2. A simple connected digit grammar for telephone number recognition

important reason is that this version of the decoder did not allow a word insertion penalty to be applied. A similar type of error was seen when running alphadigit experiments with the same model set. One had to set the word insertion penalty to around -100 to get reasonable results. Otherwise, the recognizer would hypothesize as many as three times as many words as was actually spoken. A last possible source of the errors is a mismatch between the MFCC files used

Spoken String	Speech Rate	Hypothesized String
8310	fast	6012
8310	medium	—
8310	slow	6001038856
8335	fast	—
8335	medium	6533856
8335	slow	—
9338310	fast	—
9338310	medium	6013833956
9338310	slow	—
3258335	fast	6023302
3258335	medium	6033352356
3258335	slow	—
2059338310	fast	2013339922
2059338310	medium	—
2059338310	slow	—
6013258335	fast	—
6013258335	medium	—
6013258335	slow	—

Table 1: Results of test cases for telephone number recognition system. Those marked with a “—” are those which did not produce a valid phone number.

to train the models and the MFCC files generated. We have found that the delta features we generated are suspect. A listing of the results for all test cases are shown in Table 1.

References

[1] N. Deshmukh, A. Ganapathiraju, J. Hamaker, and J. Picone, "A Public Domain Decoder for Large Vocabulary Conversational Speech Recognition," submitted to the *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, USA, May 1999.

```
# check lengths of numbers
#
if ((length(hypothesis) != 4) && (length(hypothesis) != 7) && (length(hypothesis) != 10)) {

    output error
    exit
}

# loop over phone number possibilities
#
if (length(hyp) == 4) {
   if ((hypothesis[0] eq “NINE”) || (hypothesis[0] eq “ZERO”) || (hypothesis[0] eq “OH”)) {
        output error
        exit
   }
   else {
        output hypothesis
   }
}
elsif (length(hypothesis) == 7) {
   if ((hypothesis[0] eq “ONE”) || (hypothesis[0] eq “ZERO”) || (hypothesis[0] eq “OH”)) {
       output error
       exit
   }
   else {
       output hypothesis
   }
}
elsif (length(hypothesis) == 10) {
   if ((hypothesis[0] eq “ONE”) || (hypothesis[0] eq “ZERO”) || (hypothesis[0] eq “OH”)) {
       output error
       exit
   }
   else {
       output hypothesis
   }
}
```

Figure 3. Algorithm for validation of telephone number hypothesis