RECOGNITION OF NASALIZED AND NON-NASALIZED VOWELS

Submitted By: BILAL A. RAJA
Advised by: Dr. Carol Espy Wilson and Tarun Pruthi
Speech Communication Lab, Dept of Electrical & Computer Engineering
University of Maryland, College Park

Maryland Engineering Research Internship Teams
(MERIT) Program 2006

TABLE OF CONTENTS

1 - ABSTRACT

When vowels are adjacent to nasal consonants (/m,n,ng/), they often become nasalized for at least some part of their duration. This nasalization is known to lead to changes in perceived vowel quality. The goal of this project is to verify if it is beneficial to first recognize nasalization in vowels and treat the three groups of vowels (those occurring before nasal consonants, those occurring after nasal consonants, and the oral vowels which are vowels that are not adjacent to nasal consonants) separately rather than collectively for recognizing the vowel identities. The standard Mel-Frequency Cepstral Coefficients (MFCCs) and the Hidden Markov Model (HMM) paradigm have been used for this purpose. The results show that when the system is trained on vowels in general, the recognition of nasalized vowels is 17% below that of oral vowels. This result suggests that automatic recognition of vowels would be greatly improved if we could first detect nasalization.

2 - WHAT IS NASALIZATION?

Nasalization is the production of the sound while the velum —that fleshy part of the palate near the back— is lowered, so that some air escapes through the nose during the production of the sound by the mouth. The effect is as if an [n] sound were produced simultaneously with the oral sound. Most common nasalized sounds are the nasalized vowels. When nasal consonants (/m,n,ng/)occur after the vowel in a word, the vowel is usually nasalized for at least part of its duration.

2.1 – Production of Nasal Sound

The following figure delineates the production of an oral vowel and a nasal vowel.

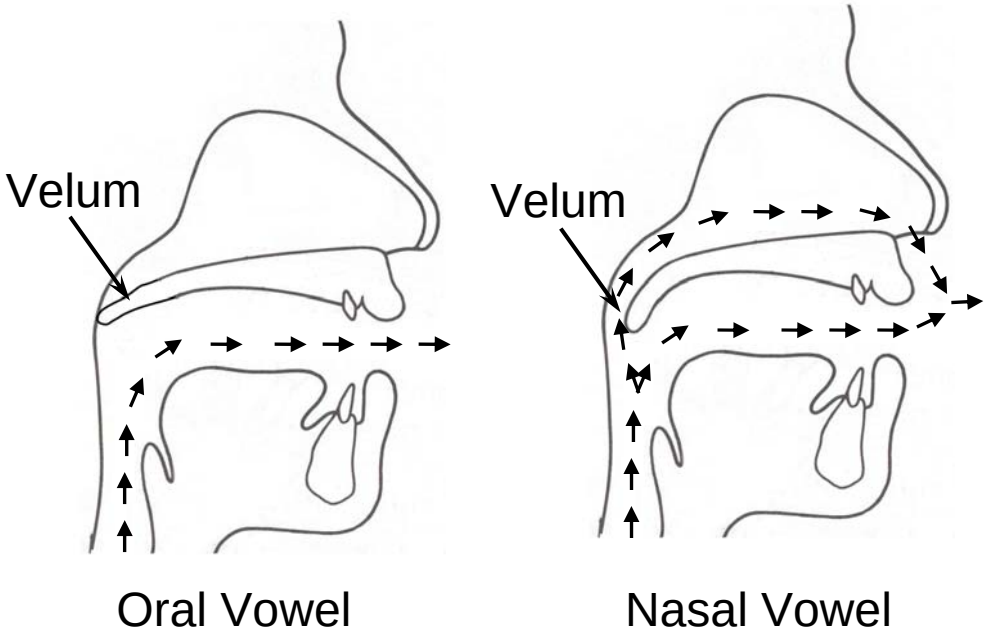

Fig-1 Production of Oral and Nasal Vowels

2.2 - <u>Common Spectral Characteristics of Nasalization</u>

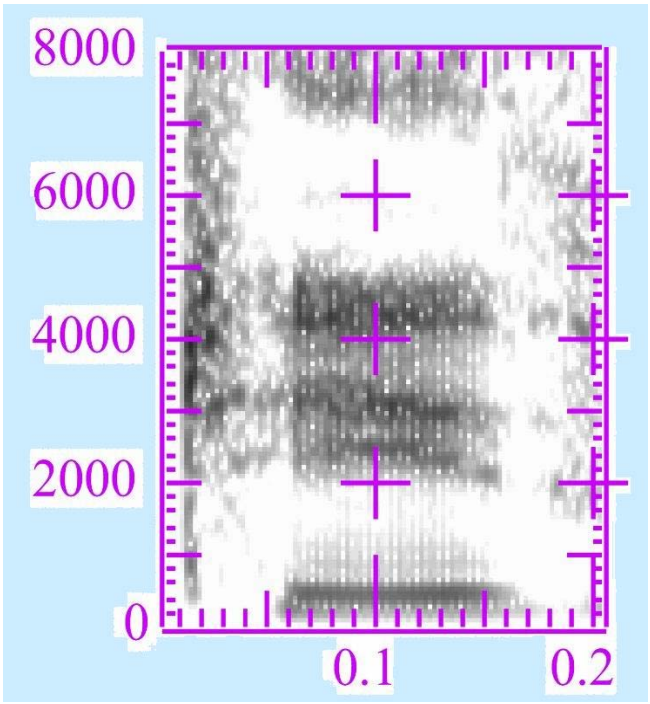

Fig 2(a) - Vowel “iy” in the word “Teeth”

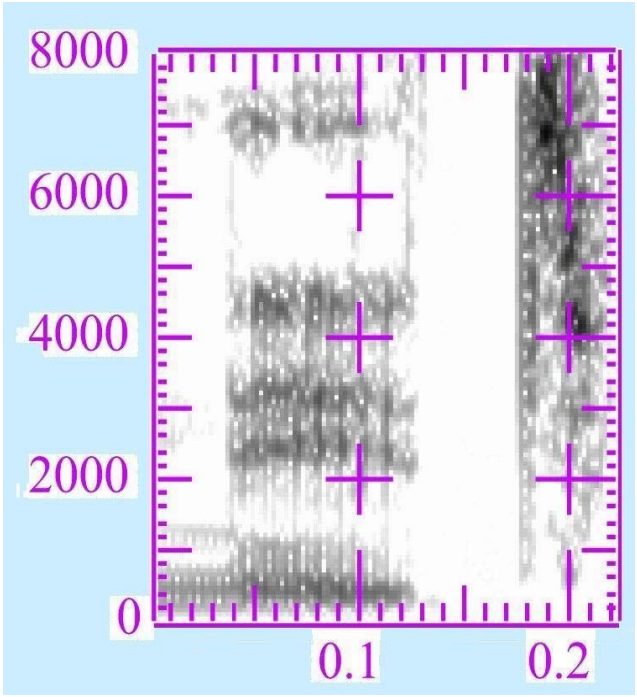

Fig 2(b) – Vowel “iy” in the word “Need”

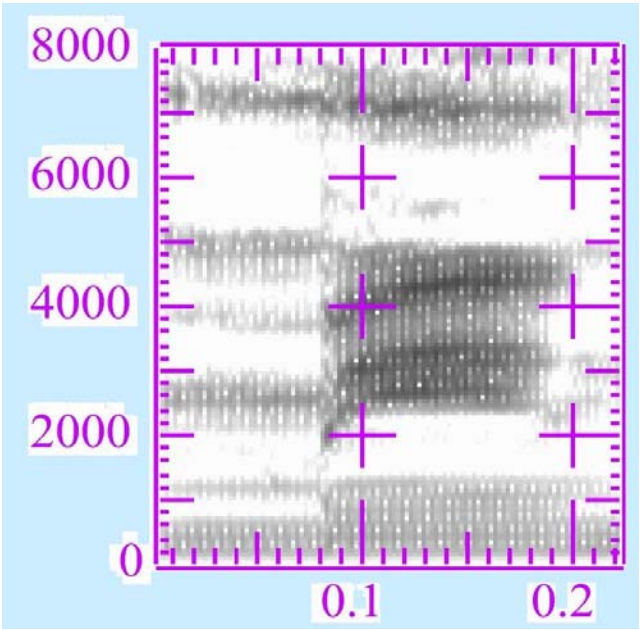

Fig 2(c) – Vowel “iy” in the word “Mean”

The above figures are showing the spectrograms of three different words with the same vowel (iy) occurring in them in three different contexts. All three words are uttered by the same speaker. The y-axis of the spectrogram represents the frequency whereas the x-axis represents the time.

Figure 2(a) delineates the spectrogram of the word “teeth” with the vowel “iy” occurring in it. We notice that we have high amplitudes at low frequencies. The darkness in the spectrum delineates high amplitude of the peaks. So the first resonance and the next resonances have high peaks and carry high energies; however, there is a deep valley between the first resonance and the second resonance.

Figure 2(b) depicts the spectrogram of the word “need” with the vowel “iy” occurring in it. The first resonance has a slighter lesser amplitude than in 2(a); however it has significant darkness till around 1000 Hz, which shows it carries slightly more energy than 2(a). It also has a big gap between first resonance and the next resonances.

Figure 2(c) shows the spectrogram of the word “mean” with the vowel “iy” occurring in it. We notice that it does not have high amplitudes at low frequencies. The first resonance has significantly low energy. However, in this case we do not see a deep valley between the first resonance and the next resonances. The first resonance and the next resonances have almost the same amplitudes. So over all we get a flatter spectrum at low frequencies. All this characterizes nasalization.

3- EXPECTED RESULTS

3.1 – Related Previous Researches

Two previous studies have obtained results that are relevant for the work in this study:

1- Bond (1975) suggested that the recognition rates for vowels are worse when the vowels are excerpted from a context in which the vowels are affected by nasality.[1]

2- According to Bell-Berti (1993), carry over coarticulation is usually much smaller than the anticipatory coarticulation. [2]

3.2 – Hypothesis

From the above mentioned research, we concluded two things:

1 – Vowels with the nasal consonants before them are not nasalized to a large extent. Therefore, their identifications rates should be comparable to oral vowels. The nasal consonant is not carried over to the vowel to the extent that it significantly affects the vowel recognition. This is also evident from the two spectrograms shown in figure 2(a) and figure 2(b). The spectrogram pattern is almost the same.

2 – Vowels with nasal consonants after them are nasalized strongly. When we are about to utter a nasal consonant we anticipate it and start to lower the velum much earlier. As a consequence, the vowel gets nasalized. The spectrogram in Figure 2(c) is clearly distinguishable from the other two spectrograms and characterizes nasalization.

4- METHOD

4.1 - The Task

The task was to get twenty different vowels (aa, ae, ah, ao, aw, ax, axh, axr, ay, eh, er, ey, ih, ix, iy, oy, ow, uh, uw, ux) from the TIMIT database [3] and make twenty different models; one for each vowel. These models were then tested and trained using HMM Tool Kit (HTK).

4.2- The Technique Used

Hidden Markov Models (HMMS) is a widely used pattern recognition system. So we used Hidden Markov Models to create models for all the vowels. Mel-Frequency Cepstral Coefficients (MFCCs) were used in an HMM framework for all the experiments. Using TIMIT transcriptions all the vowels were extracted from the speech signals and were divided into three categories:

1) Oral Vowels (OV)
2) Nasal Before Vowel (NV)
3) Vowel Before Nasal (VN)

4.3 HMM Tool Kit (HTK)

HTK is a standard tool kit for HMM based pattern recognition. The following block diagram gives the over view of the processing stages of HTK from the speech signal to the analysis of the results.

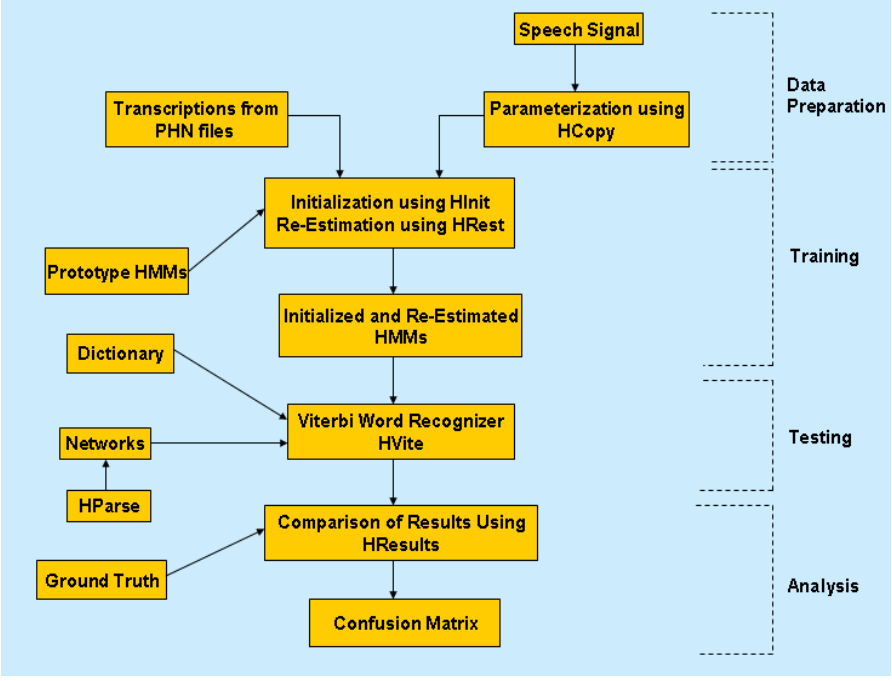

Figure 3 – HTK Processing Stages

From the above block diagram, we see that HTK has four main processing stages:

1) Data Preparation
2) Training
3) Testing
4) Analysis

4.3.1- Data Preparation

The speech signals were parameterize using the MFCCs. So the first step in the processing stage is to parameterize the speech signal using the HTK command called HCopy.

4.3.2 – Training

Once the data is prepared, the prototype HMMs are initialized and then re-estimated using the HTK commands called HInit and HRest respectively. The prototype HMMs are defined by the user, where number of stages, number of mixtures, means, variances, and the transition matrix have to specified. For our experiment, we used 5 stage model and the 8 Gaussian mixtures in the prototype HMM for each vowel model.

4.3.3 – Testing

Once the models are trained, they are tested using the HTK command called HVite. User has to create a dictionary and the word network before executing HVite. HVite uses the Virterbi algorithm to test the models.

4.3.4 – Analysis

The output of HVite is compared against the ground truth to analyze the recognition accuracy. There is a command called HResults which is used for this comparison.

HResults provides a confusion matrix which shows the recognition accuracy of the each individual model and the over all recognition accuracy. A confusion matrix is shown below as an example:

```
====================== HTK Results Analysis =======================
  Date: Fri Jul 21 11:07:34 2006
  Ref : ref_nv
  Rec : hvite_nv
------------------------ Overall Results --------------------------
SENT: %Correct=58.04 [H=783, S=566, N=1349]
WORD: %Corr=58.04, Acc=58.04 [H=783, D=0, S=566, I=0, N=1349]
------------------------ Confusion Matrix -------------------------
       a   a   a   a   a   a   a   a   e   e   e   i   i   i   o   o   u   u
       a   e   h   o   w   x   x   y   h   r   y   h   x   y   w   y   w   x
                             r                                          Del [ %c / %e]
  aa  47   0  12  10   2   0   1   5   1   1   0   0   0   0   1   0   0   0    0 [58.8/2.4]
  ae   3  53   2   0   5   1   0   4  11   0   1   1   0   0   0   0   0   0    0 [65.4/2.1]
  ah   4   5  33   0   1   6   0  11   6   1   0   0   2   0   7   3   0   0    0 [41.8/3.4]
  ao   8   1   3  58   2   0   0   0   0   1   0   0   0   0   6  11   0   0    0 [64.4/2.4]
  aw   2   2   0   0   9   0   0   0   0   0   0   0   0   0   0   1   0   0    0 [64.3/0.4]
  ax   3   0  14   1   1  13   0   3   0   2   0   1   6   0   5   1   3   0    0 [24.5/3.0]
 axr   2   0   2   4   1   0  19   0   6   4   0   0   1   0   1   0   0   0    0 [47.5/1.6]
  ay   9   1   8   1   0   0   1  60   0   0   3   1   0   0   0   2   0   0    0 [69.8/1.9]
  eh   2  19   4   1   1   2   5   3  17   1   2  12   9   0   2   0   0   0    0 [21.2/4.7]
  er   1   0   0   0   0   0  18   2   3  16   0   1   1   0   0   0   0   0    0 [38.1/1.9]
  ey   0   0   0   0   0   0   0   5   1   0  59   8   5   8   0   0   0   0    0 [68.6/2.0]
  ih   0   4   0   0   0   2   0   2   5   1   9  32  16   8   7   0   0   5    0 [35.2/4.4]
  ix   0   4   2   0   0   7   3   8  10   0   9  20  29   8   0   0   0   0    0 [29.0/5.3]
  iy   0   0   0   0   0   0   0   0   0   1  15  14   8 221   0   0   0   0    0 [85.3/2.8]
  ow   2   4   7   7   1   1   0   0   0   0   0   0   1   0  53   2   1   0    0 [67.1/1.9]
  oy   3   0   0   0   0   0   0   1   0   0   0   0   0   0   4  56   0   0    0 [87.5/0.6]
  uw   0   0   0   0   0   0   0   0   0   0   0   1   0   0   1   0   0   1    0 [ 0.0/0.2]
  ux   0   0   0   0   0   0   0   0   0   0   1   4   0   7   0   0   2   8    0 [36.4/1.0]
Ins    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
===================================================================
```

Figure 4- Confusion Matrix for the NV category showing the percentage accuracy of each vowel and the overall recognition accuracy

So in this case the over all recognition accuracy was 58.04%.

5 – **RESULTS**

We conducted two experiments:

5.1 – Experiment-1:

In this experiment we made 20 different models for each vowel. Training was done using all the vowels. However, for testing we broke down the vowels into four categories.

1) All vowels
2) Oral Vowels (OV)
3) Nasal Before Vowels (NV)
4) Vowel Before Nasal (VN)

These categories were made to find out how much error does each category contributes to the over all recognition. The results obtained are summarized in the following table:

Category	Recognition Accuracy
ALL	52.92%
OV	55%
NV	58.25%
VN	39.75%

Table 1- Percentage Recognition Accuracy for Different Categories

It is obvious from the table that the VN category, where nasal consonant occur after the vowel, has the least recognition accuracy. Also the OV and NV recognition accuracies are almost the same. These two results are consistent with our hypothesis.

5.2 – Experiment-2

In this experiment, our objective was to figure out that if we create separate models of vowels for each category (OV, NV, VN) and train and test each category separately, does that increase the recognition accuracy. Therefore the vowel models were made, initialized, and then re-estimated for each of the above mentioned category separately.

So, the training was not done using all the vowels in each category as in Experiment 1; rather each category was individually trained and then tested. The results are summarized in the table below:

Category	Recognition Accuracy
ALL	52.92%
OV	56%
NV	58.04%
VN	42.86%

Table 2 – Percentage Recognition Accuracy for Different Categories

So we see from the table that the recognition accuracy for the VN case did improve when different vowel models were created for each category.

6- CONCLUSION

The results show that when the system is trained on vowels in general, the recognition of nasalized vowels is 17% below that of oral vowels. The recognition accuracy for the VN category is improved by 8% when different vowel models are created for each category. This result suggests that automatic recognition of vowels can be improved by first detecting nasalization and then using different models.

REFERENCES

[1] Bond, Z. S, “Identification of vowels excerpted from neutral and nasal contexts,” JASA, Vol. 59, May 1976.

[2] Bell-Berti, F., “Segmental Effects,” in “Phonetics and Phonology,” Eds: Huffman, M.K. and Krakov, R.A., Academic Press Inc, 1993.

[3] TIMIT, “TIMIT acoustic-phonetic continuous speech corpus”, NIST, October 1990.