MP5: Hidden Markov Model

Group: Akhil Alapaty & Ojus Deshmukh

I. Introduction.

In the case of this MP, we are directly given the audio and visual features. The audio features were found using the Cepstrum of each of the raw audio vectors. This is done by converting the raw pixel feature vector signal into a vector of frames using the sig2frames function. Next, you take the Inverse Fourier Transform of the logarithm of the frames vector's magnitude spectrum. The Cepstrum feature vector is more effective than the Raw Pixel feature vector because it catches features such as pitches and frequencies in a speech signal, since they are part of the logarithmic representation. The cepstrum is generally characterized as containing information that depicts the rate of change in a signal's separate spectrum bands.

As for the visual data, it is computed using three distances measured on each subject's face: the lip width, the distance between the upper/lower lips, and the line connecting the mouth corners. Clearly, these visual features are more effective than the Raw Pixel feature vector since each of these measurements are almost unique to each person, which would prove to be a stronger match than the Raw Pixel feature vector.

The learning algorithm used in this MP is the Hidden Markov Model learning metric. The basic goal of this learning algorithm is to use its train and test sequences and find the best set of alpha (emission) and beta (transition) probabilities. The "hidden' refers to how each of the individual states isn't visible; only the output, which is dependent on the all of the states, can be seen. However, each state has its own probability, which gives information as to the particular sequence of the states. We specifically implemented the Baum-Welch forward-backward algorithm. This algorithm serves to compute the maximum likelihood estimate of the given parameters of a HMM along with the associated audio/video/etc feature vectors.

One of the important transforms was calculating the forward transition probability. To put it simply, we assumed that we had already seen a certain sequence of events occur, and then used the HMM to guess what event would occur next. Our other critical transform was calculating the backward transition probability. To do this, we assumed that we had already seen a certain sequence of events occur, and then used the HMM to guess what event had occurred just before that sequence.

II. Methods

Calculating α:

We defined α(i) as “the probability of the partial observation sequence O1 , O2 , … Ot (until time t), and state Si at time t, given the model λ”. We calculated α(i) entirely in ghmm_fwd.m, using the following equations:

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N.$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad \begin{array}{l} 1 \le t \le T-1 \\ 1 \le j \le N. \end{array}$$

Equation graphics were taken from Lawrence R. Rabiner, “A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”.

Calculating β:

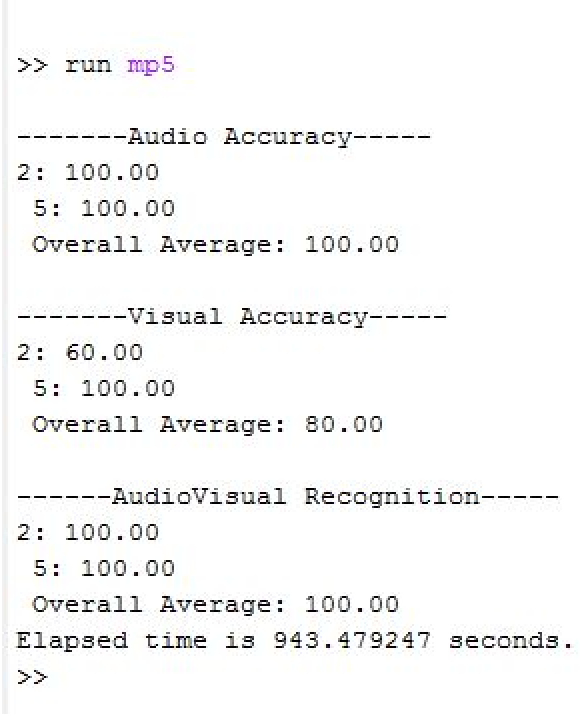
We defined β (i) as “the probability of the partial observation sequence from t+1 to the end, given state Si at time t, and model λ”. We calculated β(i) entirely in ghmm_bwd.m, using the following equations:

$$\beta_T(i) = 1, \qquad 1 \le i \le N.$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j),$$
$$t = T-1, T-2, \cdots, 1, \; 1 \le i \le N.$$

Equation graphics were taken from Lawrence R. Rabiner, “A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”.

III. Results



```
>> run mp5

-------Audio Accuracy-----
2: 100.00
 5: 100.00
 Overall Average: 100.00

-------Visual Accuracy-----
2: 60.00
 5: 100.00
 Overall Average: 80.00

------AudioVisual Recognition-----
2: 100.00
 5: 100.00
 Overall Average: 100.00
Elapsed time is 943.479247 seconds.
>>
```

IV. Discussion

The main result that stands out is the discrepancy between the 2 and 5 digits for Visual accuracy. Our best guess would be that the mouth makes a more defined and specific movement when uttering the digit ‘5’, as opposed to the digit ‘2’. For example, when saying “two”, the mouth comes forward in an “o” shape, a single movement. However, when saying “five”, the mouth closes slightly to make the “f” sound. The mouth then opens and closes to make the”v”. The many distinct steps allow the HMM to more accurately predict behavior. This is probably why ‘5’ registers as a 100% accuracy and ‘2’ registers as a lower accuracy.