**INTRODUCTION**

Emotions are one of the most essential components of communication. They affect education, decision-making, personal relationships and human perception in general. Detecting emotion of a person is the next step after perfecting speech-to-text algorithms. Understanding the text and using state of the art Natural Language Processing algorithms is not enough for making further strides in human-computer interaction. Support Vector Machines, artificial neural networks, linear discriminant analysis, k-nearest neighbors have been used in the past to tackle this problem. Due to the temporal nature of speech signals, we decided to use Hidden Markov Models (HMM) to capture the temporal complexity of speech. We were mainly interested in classifying the following emotions:

1. Anger
2. Boredom
3. Disgust
4. Happy
5. Neutral
6. Sadness

**OBJECTIVE**

Hidden Markov Models are used for time series data where the true states are “hidden” from us and what we see are surrogate observations.

- \( a(t-1) \)
- \( a(t) \)
- \( a(t+1) \)

![Figure 1: Representation of a Hidden Markov Model (HMM)](Image)

Speech-related features are calculated (pitch, intensity, MFCCs, formants...). These are our observations, and in the HMM, they come from hidden variables: the states. To build an emotion recognition engine, we had to train one HMM per emotion (6 models) and at the time of prediction, the maximum likelihood model determines the emotion.

We also used GBM as a first baseline, and to get some insights into the relative importance of the features (for features selection purposes). We chose GBM because it is a relatively powerful classifier and gives a good measure of the importance of each feature.

**FEATURES OF THE DATA**

The ‘Emotional Prosody Speech and Transcripts’ database, courtesy of the Linguistics Department at Stanford was used because of good articulation of words and proof of previous success for other authors.

This dataset initially contained 14 emotion labels on which we tried k-means clustering to check for a natural clustering of emotions. Clustering was hard to perform for this problem, emotions are not clearly separable and that’s the difficulty of the problem. We decided to manually group similar emotions into 6 categories.

**RESULTS**

Gradient boosting machines using decision trees led to a good classification accuracy of 83.4% on our 6 emotion classes as can be seen in Table I.

As an initial attempt at applying HMM’s to this application, we assumed the speech data to be distributed as a mixture of four Gaussian distributions. Each of the six emotions was modeled as a separate HMM with 4 hidden states. This resulted in an accuracy of about 45%. We plan on increasing this accuracy by using more effective locally found features like pitch, energy and their differential and acceleration coefficients.

<table>
<thead>
<tr>
<th>Table I</th>
<th>CONFUSION MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction / Truth</td>
<td>anger</td>
</tr>
<tr>
<td>anger</td>
<td>38</td>
</tr>
<tr>
<td>boredom</td>
<td>0</td>
</tr>
<tr>
<td>disgust</td>
<td>0</td>
</tr>
<tr>
<td>happy</td>
<td>3</td>
</tr>
<tr>
<td>neutral</td>
<td>3</td>
</tr>
<tr>
<td>sadness</td>
<td>2</td>
</tr>
</tbody>
</table>

**CONCLUSION**

- Features like pitch, intensity and MFCC coefficients show a clear trend with different emotions (see Figure 3 and 4). This provided us direction to effectively choose features that would maximize classification accuracy.
- The Gradient Boosting Machines (GBM) algorithm gave a very good accuracy of 83.4% with the global features listed on the left.
- An initial attempt at implementing HMM’s led to a rough accuracy of about 45% which can be improved by a more effective feature selection and modifying the number of states for each HMM.
- MFCC coefficients did not prove to be useful features for use with our HMM model.

**FUTURE RESEARCH**

In order to improve the accuracy of the HMM’s model, we plan to include more features like pitch, energy, their differential and acceleration coefficients and also features like MFCC’s and formants, which are generally used in speech applications.

We also intend to use a more sophisticated HMM model while training over the dataset. To prove that our model is language agnostic, we plan to merge our current dataset with other datasets in different languages.

**REFERENCES**