Emotion Recognition and Synthesis in Speech

Dan Burrows
Electrical And Computer Engineering
Carnegie Mellon University
dburrows@andrew.cmu.edu

Ajay Ghadiyaram
Electrical And Computer Engineering
Carnegie Mellon University
aghadiya@andrew.cmu.edu

Maxwell Jordan
Electrical and Computer Engineering
Carnegie Mellon University
maxwelljordan@cmu.edu

Amandianeze Nwana
Electrical and Computer Engineering
Carnegie Mellon University
aon@andrew.cmu.edu

Amber Xu
Electrical and Computer Engineering
Carnegie Mellon University
axu@andrew.cmu.edu

Abstract

In this paper we describe an emotion recognition system that uses Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs). We also describe how to synthesize speech that has emotion using state-of-the-art speech synthesis tools. Our feature set includes f_0, mel cepstral coefficients (MCEPs) and power. Four-fold cross validation is used to test the accuracy of our recognition system. We synthesize and recognize four fundamental emotions: happiness, hot anger, neutrality, and sadness. Our system classifies speech into one of the four emotion categories and responds with speech synthesized in that same emotion.

1 Introduction

Emotion synthesis and recognition has many applications. It can be used in automated call centers, lie detector systems and by psychologists. To recognize emotions we used a decision tree to decide between emotions. We built GMMs around three features and used the probabilities they output to feed into a SVM which makes decisions throughout the decision tree. After determining the spoken emotion we synthesis one of four emotions; happy, hot anger, neutral or sadness in response. The voices were created using the Festival Speech Synthesis System and the Festvox project. We explored whether power conversion and durational models can be used to improve synthesis emotional speech. We will discuss the contents of our database in Section 2, describe the emotion classifier in Section 3, and elaborate on converting a voice into different emotions in Section 4.

2 Database

The database that was used for the project was the Emotional Prosody Speech and Transcripts database provided by the Linguistic Data Consortium (LDC2002S28). The database is composed of recorded speech from three male and four female professional actors. The database consists of utterances containing only numbers and dates where each utterance is approximately 2 seconds long. The utterances are expressed in 15 unique categories: neutral, disgust, panic, anxiety, hot anger,

cold anger, despair, sadness, elation, happy, interest, boredom, shame, pride, and contempt. The sampling rate is 22.05 kHz and the speech is stored using dual-channel interleaved 16-bit PCM. The database has a text transcript for each speaker that documents the words that were spoken during each utterance. Each utterance is also labeled with a single emotion category in this transcript.

For the purpose of this paper we will focus on classification between four emotions on Male subjects only; 'Happy', 'Hot Anger', 'Sadness' and 'Neutral'. This decision was motivated by the low accuracy in classification between all 15 emotions and both sexes in [4]. We normalized the power in each utterance and downsampled the audio to 16 kHz so that our data was compatible with Festival.

3 Recognition

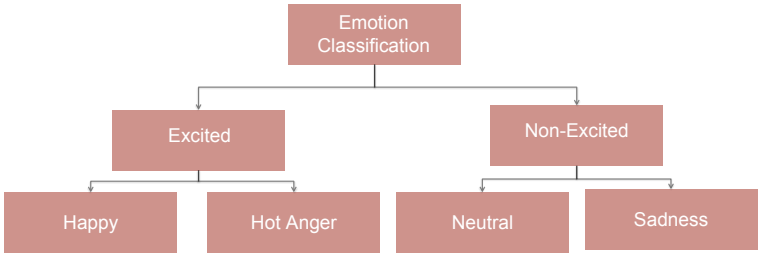


Figure 1: Hierarchal classification tree.

Our speaker dependent emotion classifier is a decision tree. Each level contains a binary decision between two emotional categories. At the first level the decision is between the 'Excited' and 'Non-Excited' classes. The 'Excited' class is a combination of the 'Happy' and 'Hot Anger' classes and the 'Non-Excited' classes is a combination of the 'Neutral' and 'Sadness' classes.

After determining which top level class the utterance comes from, we enter the second level of the tree where a binary division occurs between 'Happy' and 'Hot Anger' or 'Neutral' and 'Sadness' depending on which first level class was picked.

Three GMMs were constructed for each class and trained on the Mel-Cepstrum coefficients, statistics of the f_0, and statistics of the power. This results in a total of 18 GMMs whose outputs were fed into the SVMs that make up the decision tree. In the following discussion we will describe in detail the features we chose, how we constructed the GMMs, how we determined the SVM kernel to use and our recognition accuracies.

3.1 Features

We used a set of three distinct features to train our models and then classify new data. We calculated the Mel-Cepstrum Coefficients by using the Speech Signal Processing Toolkit that is part of Festival. The Mel Cepstrum is described as:

$$\mathscr{F}^{-1}(log(\mid Mel\ Scale\mid^2)) \tag{1}$$

$$Mel\ Scale\ =\ 2595 * log_{10}(\frac{\mathscr{F}(x(t))}{700} + 1) \tag{2}$$

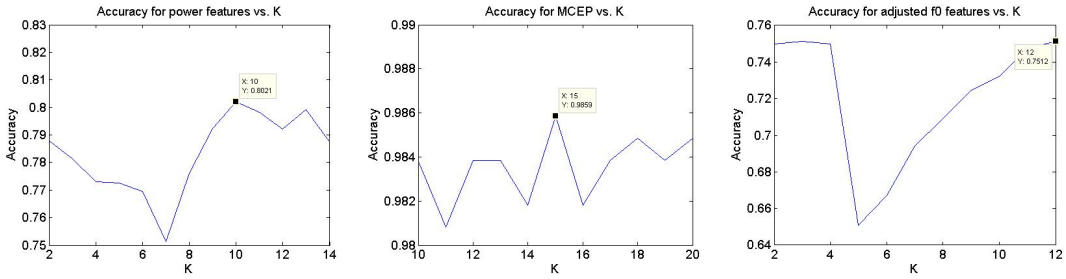
We calculated 24 coefficients for each 10ms frame of speech. This creates $\frac{Length\ of\ Utterance(s)}{.01(s)}$ points in 24 dimensional space.

For the fundamental frequency we used a similar method as by Medan et al. [1]. Through the implementation provided in Festival which autocorrelates adjacent windows to determine the τ value, frequency, that maximizes the correlation. We then threshold the overall pitches based on human speech to remove unvoiced segments and outliers due to cracks in the speaker's voice. This builds a vector of voiced pitches per utterance. We calculated the mean, variance, minimum and maximum of the original f_0 and its first and second derivatives to use as features for training the GMMs.

Power is also calculated within Festival over 10ms segments as for MCEPs. Similarly to f_0, the mean, variance, minimum and maximum of the power values, its first derivative, and its second derivative make up the features for training the GMMs for power.

3.2 Building the Gaussian Mixture Models

We constructed one GMM per class for each of the three features using an Expectation Maximization algorithm as described by Reynolds et al. [2]. We determined experimentally that a full covariance GMM worked best for MCEP and diagonal covariance GMM worked best for f_0 and power. To construct the models we then determined how many component Gaussian densities and nodal variances to include in each model. We did this by using MATLAB's gmdistribution.fit command for different model orders, K, and tracked the resulting accuracies.



(a) (b) (c)

Figure 2: 4-Fold cross validation for the first level classification for model order K.

To determine the optimal K values for each model we compared 4-fold cross validation accuracies (Figure 2). We also considered convergence and computation time of the model. For f_0 and power, the range of K was limited by the amount of test data available which prevented convergence at high values of K; whereas for MCEPs, computation time became unrealistic after $K = 20$. From our tests we determined that $K = 10$ for power, $K = 12$ for f_0, and $K = 15$ for MCEPs created the most accurate mixture models.

3.3 Building the Support Vector Machine

The probabilities from the GMMs for three features over the six classes are fed into an SVM which is used to make decisions through our decision tree. To train the SVMs, we use MATLAB's built in svmtrain to develop an SVM for each of the three decision we make. We also determine which kernel best supports our model. After attempting various kernels as shown in Figure 3, we found a linear kernel produced the highest accuracy between individual emotion with an accuracy of 85.4%. In the graph it is evident that all kernels have a very high accuracy for the first decision (in red) between the 'Excited' and 'Non-Excited' classes but there is less reliability after the second level deciding between specific emotions (in blue).

3.4 Testing

After we have created our GMMs and SVMs from the training set we are ready to start testing the model. We calculate the features in the same manner for each test utterance as we did for the training set. At the first level we calculate the probability that the utterance came from either the 'Excited' or 'Non-Excited' class for f_0 and power features, and the normalized log likelihood for MCEPs from the GMMs. We feed these probabilities into the top level linear kernel SVM we trained earlier to determine which class the result is most likely from. After a decision is made, we move down that branch of our tree and classify between either 'Happy' and 'Hot Anger' or 'Neutral' and 'Sadness'.

From the Confusion Matrix in Figure 4 we can see that 'Happy' and 'Hot Anger' or 'Neutral' and 'Sadness' are most often confused however 'Neutral' and 'Happy' or 'Neutral' and 'Hot Anger' are

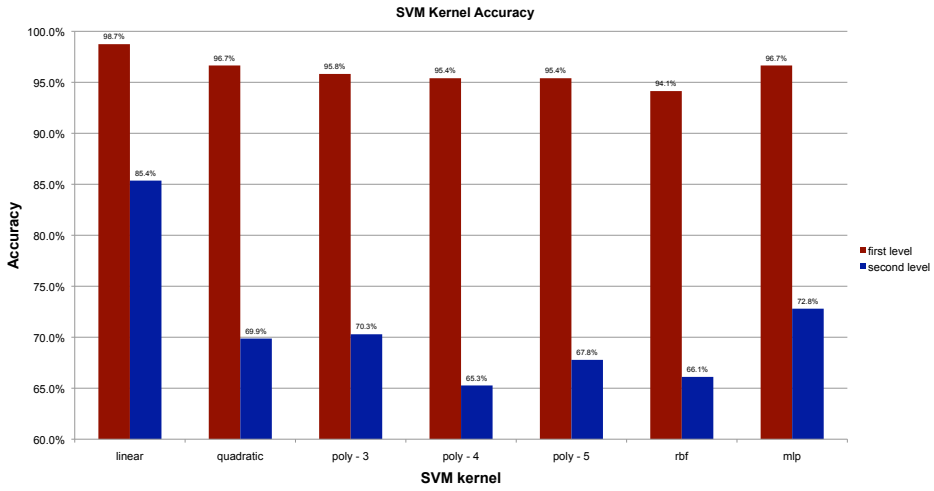


Figure 3: SVM Kernel Accuracy Comparison

not. This is due to the high accuracy of the top level of the hierarchy split and therefore limits confusion between emotions from different first level classes. The hierarchy clearly limits our error rates. We can also see that the second level of classification skews the decision towards over classifying sadness and happy over neutral and hot anger. We speculate that this is due to having a slightly higher number of utterances for Sadness and Happy. However, due to our small dataset, we could not afford to remove utterances to have the same number of utterances for each emotion.

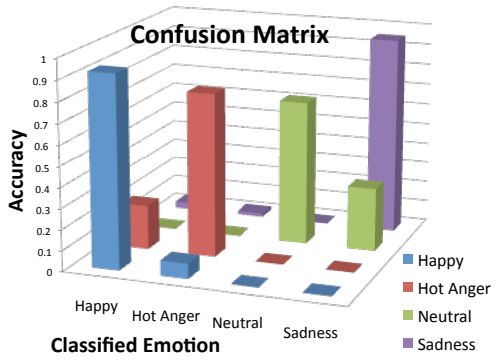


Figure 4: Confusion Matrix of Overall System

4 Voice Conversion

We use the Festival Speech Synthesis System and the Festvox project to synthesize speech. In order to build a synthetic voice from scratch using these tools, a large database of utterances is needed. However, a full Festival voice for a target speaker based on an existing Festival voice can be built using much less data than is needed to build a Festival voice from scratch. This technique is called voice conversion. We use voice conversion because existing databases of emotional speech are very small. The voice conversion technique modifies the pitch, MCEPs and power of the source speaker to generate speech that sounds like the target speaker. We also tried building a duration model for the target speaker. We synthesized emotional speech using several source voices and found that the kal_diphone voice has the best quality. We built transforms for the happy, angry, neutral, and sad emotions. There was a noticeable difference between the excited (happy and hot anger) and non-excited (neutral and sadness) speech but there was little difference between happy and angry speech.

There was also little difference between neutral and sad speech. The techniques used to convert pitch, MCEPs, and powers are discussed below.

4.1 MCEP Mapping

The technique that is used to convert the source MCEPs is described in detail by Toda et al. and summarized below [3]. We extract the MCEPs and ΔMCEPs from source and target frames (each frame is 10ms of speech). Next, we align the source and target speech in time using dynamic time warping. A GMM is built to model the joint pdf of source and target MCEPs and ΔMCEPs. Using the mean and variance of each mixture component, another GMM containing a penalty term for reduction of global variance is built to model the conditional PDF per target frame given the source frame. Finally the MLE of the target frames are computed using the conditional PDF. Power is also converted using this technique.

4.2 Pitch Mapping

The fundamental frequency is converted using the following formula:

$$\hat{y_t} = \frac{\sigma^y}{\sigma^x}(x_t - \mu^x) + \mu^y \tag{3}$$

Where $\hat{y_t}$ and x_t are the log scaled f_0 of the source speaker and the target speaker at frame t. μ^x and μ^y are the mean log-scaled f_0 from the source and target speaker. σ^x and σ^y are the standard deviation of the log-scaled f_0 from the source and target speaker. This formula is derived by calculating the conditional expected value of the target pitch given the source pitch assuming both source and target pitch are from a gaussian distribution.

4.3 Duration

The phone durations for the target speech were modeled using a Classification and Regression Tree (CART). The CART tree was built using the Edinburgh Speech Tools Library which is distributed with Festival. The features used include previous and future phones, whether the phone is stressed or unstressed, word position and phrase position. We varied the size of the leaf nodes in the tree from one to fifty. The results are shown in Figure 5. The correlation is always in the neighborhood of 0.1 so we decided not to use durational information in our synthesizer.

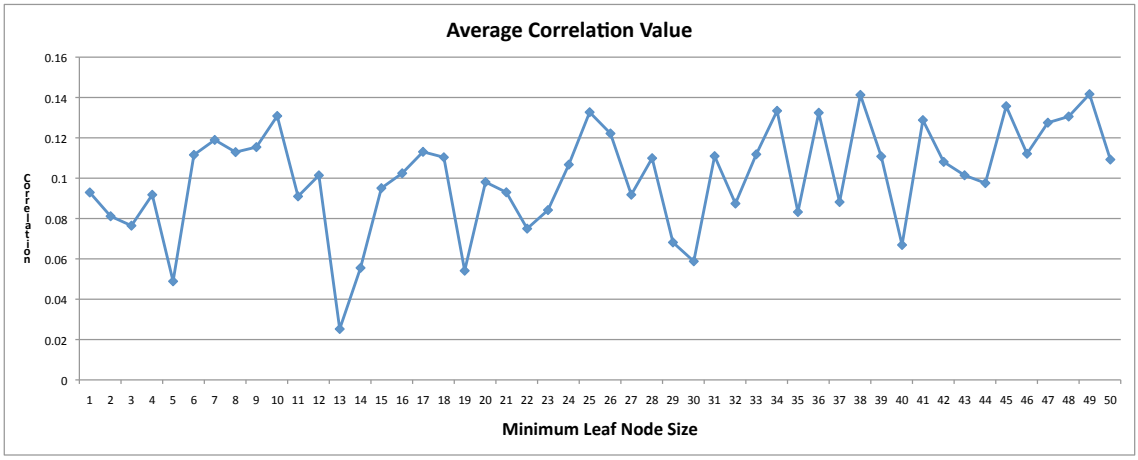


Figure 5: Plot of the average correlation values versus minimum leaf node size.

5 Conclusion

In the end we created a system that could identify emotion in speech and respond with an emotional response. The primary shortcoming through our work was the lack of data. The utterances were very short and made training and building voice models most difficult. Attempting to combine emotions such as happy and elation or sadness and despair produced mixed results and skewed classification towards emotions with larger training sets. This also affected synthesis with weaker models to build new voices with. With these shortcomings we still produced accurate results. The recognition portion had a classification accuracy of 85.4% after four-fold cross validation. Subsequently the synthesis aspect could create four distinct emotions based on our classification results.

We determined empirically that mixing emotions to create new 'in between' emotions had little effect. This is due to the fact that it is easier for the human ear to recognize emotional extremes. For this reason we built four transformations to uniquely synthesize each of the four emotions we used in training. Once an emotion has been classified we can synthesize a response in the classified emotion.

6 Acknowledgments

We would like to thank Dr. Alan Black, Dr. Bhiksha Ramakrishnan, and the 18-797 Machine Learning course staff for their guidance and support during this project.

References

[1] Yoav Medan, Eyal Yair, and Dan Chazan. Super resolution pitch determination of speech signals. *IEEE Transactions On Signal Processing*, 39(1):40–48, 1991.

[2] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[3] Tomoki Toda, Alan Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transations of Audio, Speech and Language Processing*, 15(8):2222–2236, 2007.

[4] Sherif Yacoub, Steve Simske, Xiaofan Lin, and John Burns. Recognition of emotions in interactive voice response systems. Technical report, Hewlett-Packard Company, 2003.