Emotion Detection from Speech

1. Introduction

Although emotion detection from speech is a relatively new field of research, it has many potential applications. In human-computer or human-human interaction systems, emotion recognition systems could provide users with improved services by being adaptive to their emotions. In virtual worlds, emotion recognition could help simulate more realistic avatar interaction.

The body of work on detecting emotion in speech is quite limited. Currently, researchers are still debating what features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together.

In this project, we attempt to address these issues. We use K-Means and Support Vector Machines (SVMs) to classify opposing emotions. We separate the speech by speaker gender to investigate the relationship between gender and emotional content of speech.

There are a variety of temporal and spectral features that can be extracted from human speech. We use statistics relating to the pitch, Mel Frequency Cepstral Coefficients (MFCCs) and Formants of speech as inputs to classification algorithms. The emotion recognition accuracy of these experiments allow us to explain which features carry the most emotional information and why.

It also allows us to develop criteria to class emotions together. Using these techniques we are able to achieve high emotion recognition accuracy.

2. Corpus of Emotional Speech Data

The data used for this project comes from the Linguistic Data Consortium's study on Emotional Prosody and Speech Transcripts [1]. The audio recordings and corresponding transcripts were collected over an eight month period in 2000-2001 and are designed to support research in emotional prosody.

The recordings consist of professional actors reading a series of semantically neutral utterances (dates and numbers) spanning fourteen distinct emotional categories, selected after Banse & Scherer's study of vocal emotional expression in German [2]. There were 5 female speakers and 3 male speakers, all in their mid-20s. The number of utterances that belong to each emotion category is shown in Table 1. The recordings were recorded with a sampling rate of 22050Hz and encoded in two-channel interleaved 16-bit PCM, high-byte-first ("big-endian") format. They were then converted to single channel recordings by taking the average of both channels and removing the DC-offset.

Neutral	Disgust	Panic	Anxiety
82	171	141	170
Hot Anger	Cold Anger	Despair	Sadness
138	155	171	149
Elation	Happy	Interest	Boredom
159	177	176	154
Shame	Pride	Contempt	
148	150	181	

Table 1: Number of utterances belonging to each Emotion Category

3. Feature Extraction

Pitch and related features

Bäzinger et al. argued that statistics related to pitch conveys considerable information about emotional status [3]. Yu et al. have shown that some statistics of the pitch carries information about emotion in Mandarin speech [4].

For this project, pitch is extracted from the speech waveform using a modified version of the RAPT algorithm for pitch tracking [5] implemented in the VOICEBOX toolbox [6]. Using a frame length of 50ms, the pitch for each frame was calculated and placed in a vector to correspond to that frame. If the speech is unvoiced the corresponding marker in the pitch vector was set to zero.

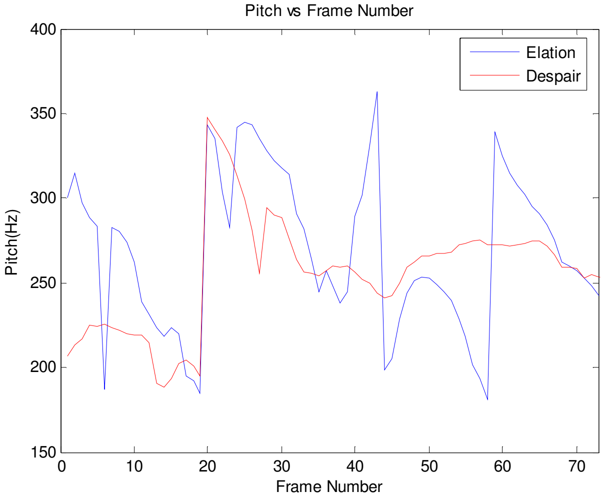

Figure 1: Variation in Pitch for 2 emotional states

Figure 1 shows the variation in pitch for a female speaker uttering “Seventy one” in emotional states of despair and elation. It is evident from this figure that the mean and variance of the pitch is higher when “Seventy one” is uttered in elation rather than despair. In order to capture these and other characteristics, the following statistics are calculated from the pitch and used in the pitch feature vector:

- Mean, Median, Variance, Maximum, Minimum (for the pitch vector and its derivative)
- Average energies of voiced and unvoiced speech
- Speaking rate (inverse of the average length of the voiced part of the utterance)

Hence, the pitch feature vector is 13-dimensional.

MFCC and related features

MFCCs are the most widely used spectral representation of speech in many applications, including speech and speaker recognition. Kim et al. argued that statistics relating to MFCCs also carry emotional information [7].

For each 25ms frame of speech, thirteen standard MFCC parameters are calculated by taking the absolute value of the STFT, warping it to a Mel frequency scale, taking the DCT of the log-Mel-spectrum and returning the first 13 components [8]. Figure 2 shows the variation in three MFCCs for a female speaker uttering “Seventy one” in emotional states of despair and elation. It is evident from this figure that the mean of the first coefficient is higher when “Seventy one” is uttered in elation rather than despair, but is lower for the second and third coefficients. In order to capture these and other characteristics, we extracted statistics based on the MFCCs. For each coefficient and its derivative we calculated the mean, variance, maximum and minimum across all frames. We also calculate the mean, variance, maximum and minimum of the mean of each coefficient and its derivative. Each MFCC feature vector is 112-dimensional.

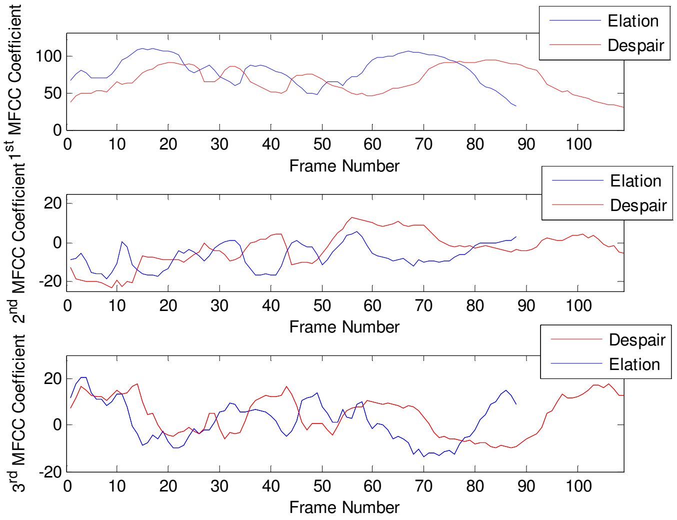

Figure 2: Variation in MFCCs for 2 emotional states

Formants and related features

Tracking formants over time is used to model the change in the vocal tract shape. The use of Linear Predictive Coding (LPC) to model formants is widely used in speech synthesis [9]. Prior work done by Petrushin suggests that formants carry information about emotional content [10]. The first three formants and their bandwidths were estimated using LPC on 15ms frames of speech. For each of the three formants, their derivatives and bandwidths, we calculated the mean, variance, maximum and minimum across all frames. We also calculate the mean, variance, maximum and minimum of the mean of each formant frequency, its derivative and bandwidth. The formant feature vector is 48-dimensional.

4. Classification

We tried to differentiate between "opposing" emotional states. Six different "opposing" emotion pairs were chosen: despair and elation, happy and sadness, interest and boredom, shame and pride, hot anger and elation, and cold anger and sadness.

For each emotion pair, we formed data sets comprising of emotional speech from all speakers, only male speakers, and only female speakers because the features are affected by the gender of the speaker. For example, the pitch of males ranges from 80Hz to 200Hz while the pitch of females ranges from 150Hz to 350Hz [11]. This corresponds to a total of eighteen unique data sets.

For each data set, we formed inputs to our classification algorithm comprising of feature vectors from: Pitch only, MFCCs only, Formants only, Pitch & MFCCs, Pitch & Formants, MFCCs & Formants, and Pitch, MFCCs & Formants. Hence, for each emotion pair, the classification algorithm was run on twenty one different sets of inputs.

K-Means Clustering

For each emotion pair, all input sets were clustered using K-Means clustering (k = 2) for all twelve combinations of the parameters listed below:
Distance Measure Minimized: Squared Euclidean, L1 norm, Correlation, and Cosine (the Correlation and Cosine distance measures used here are as defined in the MATLAB 'kmeans' function).
Initial Cluster Centroids: Random Centroids and User Defined Centroids (UDC). A UDC is the centroid that minimizes the distance measure for the input features of one emotion in the emotion pair.
Maximum Number of Iterations: 1 (only when the initial cluster centroid is a UDC) and 100 (for both Random and UDC centroids).

The error used to obtain the recognition accuracy is the average of the training errors obtained by 10-fold cross validation and is an estimate of the generalization error. The variance of the recognition accuracy is the variance of these training errors. For each experiment, the highest recognition accuracy achieved, its variance, the inputs features and clustering parameters used, is listed in Table 2.

All Speakers						
Experiment	Features	Distance Measure	Centroid	Iterations	Recognition Accuracy	Variance
despair-elation	MFCC	L1 norm	UDC	100	75.76%	1.74%
happy-sadness	MFCC	L1 norm	UDC	1	77.91%	14.34%
interest-boredom	Pitch	L1 norm	UDC	100	71.21%	2.48%
shame-pride	MFCC	L1 norm	UDC	1	73.15%	3.23%
hot anger-elation	MFCC	L1 norm	UDC	1	69.70%	10.75%
cold anger-sadness	MFCC	L1 norm	UDC	1	75.66%	3.35%
Male Speakers						
Experiment	Features	Distance Measure	Centroid	Iterations	Recognition Accuracy	Variance
despair-elation	MFCC & Pitch	Correlation	UDC	1	87.80%	0.14%
happy-sadness	MFCC	L1 norm	UDC	1	88.80%	3.66%
interest-boredom	MFCC & Pitch	Cosine	Random	100	81.20%	6.36%
shame-pride	MFCC & Pitch	Correlation	UDC	1	74.24%	15.53%
hot anger-elation	MFCC	L1 norm	UDC	1	65.89%	14.95%
cold anger-sadness	MFCC	L1 norm	UDC	1	88.43%	9.78%
Female Speakers						
Experiment	Features	Distance Measure	Centroid	Iterations	Recognition Accuracy	Variance
despair-elation	MFCC	L1 norm	UDC	1	80.42%	9.66%
happy-sadness	MFCC	L1 norm	UDC	1	72.80%	15.24%
interest-boredom	MFCC	L1 norm	UDC	1	70.62%	18.06%
shame-pride	MFCC	L1 norm	UDC	1	81.18%	19.79%
hot anger-elation	MFCC	L1 norm	UDC	1	77.16%	4.37%
cold anger-sadness	MFCC	Correlation	UDC	1	72.04%	15.00%

Table 2: Highest Recognition Accuracies using K-means Clustering

Support Vector Machines (SVMs)

A modified version of the 2-class SVM classifier in Schwaighofer's SVM Toolbox [12] was used to classify all input sets of each emotion pair. The two kernels used and their parameters are:

1. **Linear Kernel** (with parameter C, corresponding to the upper bound for the coefficients α_i's, ranges from 0.1-100, with multiplicative step 10).
2. **Radial Basis Function** (RBF) Kernel (parameter C, corresponding to the upper bound for the coefficients α_i's, ranges from 0.1-10, with multiplicative step $\sqrt{2}$)

The recognition accuracy and variance was calculated using the same technique as for K-means. For each experiment, the highest recognition accuracy achieved, its variance, the inputs features and clustering parameters used is listed in Table 3.

All Speakers					
Experiment	Features	Kernel	C	Recognition Accuracy	Variance
despair-elation	MFCC	RBF Kernel	6.4	83.44%	5.52%
happy-sadness	MFCC	Linear Kernel	1	72.50%	19.65%
interest-boredom	MFCC + Pitch	Linear Kernel	10	65.15%	17.68%
shame-pride	MFCC	RBF Kernel	1.6	76.55%	7.93%
hot anger-elation	MFCC + Pitch	Linear Kernel	1	60.00%	19.79%
cold anger-sadness	MFCC	RBF Kernel	6.4	86.00%	4.10%
Male Speakers					
Experiment	Features	Kernel	C	Recognition Accuracy	Variance
despair-elation	MFCC + Pitch	Linear Kernel	1	96.67%	6.57%
happy-sadness	MFCC	Linear Kernel	1	99.17%	24.55%
interest-boredom	MFCC + Pitch	Linear Kernel	10	96.15%	22.78%
shame-pride	MFCC + Pitch	Linear Kernel	100	91.54%	24.59%
hot anger-elation	MFCC	Linear Kernel	10	90.00%	16.22%
cold anger-sadness	MFCC	Linear Kernel	100	96.67%	21.25%
Female Speakers					
Experiment	Features	Kernel	C	Recognition Accuracy	Variance
despair-elation	MFCC + Pitch	Linear Kernel	1	79.50%	14.03%
happy-sadness	Pitch	Linear Kernel	1	62.00%	11.11%
interest-boredom	Pitch	Linear Kernel	100	80.53%	11.85%
shame-pride	Pitch	Linear Kernel	1	76.88%	23.61%
hot anger-elation	MFCC + Pitch	Linear Kernel	10	88.75%	13.52%
cold anger-sadness	MFCC	Linear Kernel	10	96.11%	7.15%

Table 3: Highest Recognition Accuracies using 2-Class SVMs

6. Discussion

The results obtained by the experiments performed allow us to make the following observations.

Using the formant feature vector as an input to our classification algorithms, always results in sub-optimal recognition accuracy. We can infer that formant features do not carry much emotional information. Since formants are used to model the resonance frequencies (and shape) of the vocal tract, we can postulate that different emotions do not significantly affect the vocal tract shape.

Using Squared Euclidean as a distance measure for K-means always results in sub-optimal recognition accuracy. Using this distance metric effectively places a lot of weight on the magnitude of an element in the feature vector. Hence, an input feature that might vary a lot between the two opposing emotions may be discounted by this distance measure, if the mean of this feature is smaller than that of other features.

Tables 2 and 3 indicate that the recognition accuracy is higher when the emotion pairs of male and female speakers were classified separately. We postulate two reasons for this behavior. First, using a larger number of speakers (as in the all speaker case) increases the variability associated with the features, thereby hindering correct classification. Second, since MFCCs are used for speaker recognition, we hypothesize that the features also carry information relating to the identity of the speaker. In addition to emotional content MFCCs and Pitch also carry information about the gender of the speaker. This additional information is unrelated to emotion and increases misclassification.

Tables 2 and 3 also suggest that the recognition rate for female speakers is lower than male speakers when classifying emotional states of elation, happiness and interest. The higher number of female speakers than male speakers in our data set may contribute to this lower recognition accuracy. Further investigation suggested that in excitable emotional states such as interest, elation and happiness, the variance of the Pitch and MFCCs increases significantly. However, variance of the Pitch and MFCCs is higher for female voices than male voices. Hence, this increase in variance is masked by the natural variance in female voices, which could make the features less effective at correctly classifying agitated emotions in female speakers.

Of all the methods implemented, SVMs with a linear kernel give us the best results for single-gender classification, especially in male speakers. This indicates that this feature space is almost linearly separable. The best results using K-Means classification are usually obtained when the cluster centroids are UDCs which we think indicates that unsupervised learning algorithms such as K-Means cannot pick up on all the information contained in the feature sets, unless we add some bias to the features.

7. Conclusion & Future Work

Although it is impossible to accurately compare recognition accuracies from this study to other studies because of the different data sets used, the methods implemented here are extremely promising. The recognition accuracies obtained using SVMs with linear kernels for male speakers are higher than any other study. Previous studies have neglected to separate out male and female speakers. This project shows that there is significant benefit in doing so. Our methods are reasonably accurate at recognizing emotions in female and all speakers. Our project shows that features derived from agitated emotions such as happiness, elation and interest have similar properties, as do those from more subdued emotions such as despair and sadness. Hence, ‘agitated’ and ‘subdued’ emotion class can encompass these narrower emotions. This is especially useful for animating gestures of avatars in virtual worlds.

This project focused on 2-way classification. The performance of these methods should be evaluated for multi-class classification (using multi-class SVMs and K-Means). In addition the features could be fit to Gaussians and classified using Gaussian Mixture Models. The speakers used here uttered numbers and dates in various emotions – the words themselves carried to emotional information. In reality, word choice can indicate emotion. MFCCs are widely used in speech recognition systems and also carry emotional information. Existing speech recognition systems could be modified to detect emotions as well. To help improve emotion recognition we could combine methods in this project and methods similar to the Naïve Bayes in order to take advantage of the emotional content of the words.

8. References

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28

[2] R.Banse, K.R.Scherer, “Acoustic profiles in vocal emotion expression”, *Journal of Personality and Social Psychology*, Vol.70, 614-636, 1996

[3] T.Bänziger, K.R.Scherer, “The role of intonation in emotional expression”, *Speech Communication*, Vol.46, 252-267, 2005

[4] F.Yu, E.Chang, Y.Xu, H.Shum, “Emotion detection from speech to enrich multimedia content”, *Lecture Notes In Computer Science*, Vol.2195, 550-557, 2001

[5] D.Talkin, “A Robust Algorithm for Pitch Tracking (RAPT)”, *Speech Coding & Synthesis*, 1995

[6] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[7] S.Kim, P.Georgiou, S.Lee, S.Narayanan. “Real-time emotion detection system using speech: Multi-modal fusion of different timescale features”, *Proceedings of IEEE Multimedia Signal Processing Workshop*, Chania, Greece, 2007

[8] http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[9] L.R.Rabiner and B.H.Juang. “Fundamentals of Speech Recognition”, Upper Saddle River; NJ: Prentice-Hall, 1993

[10] V.A Petrushin, “Emotional Recognition in Speech Signal: Experimental Study, Development, and Application”, *ICSLP-2000*, Vol.2, 222-225, 2000

[11] L.R.Rabiner and R.W.Schafer. “Digital processing of speech signals”, Englewood Cliffs; London: Prentice-Hall, 1978

[12] http://ida.first.fraunhofer.de/~anton/software.html