Speech Sentiment: Final Report

Class: CS221 / Artificial Intelligence Principles and Techniques
Students: Grigory Makarevich, Joris van Mens

Introduction
Sentences spoken with different emotion (e.g. with angry or happy tone of voice) can convey entirely different meaning. Accurately detecting this emotion in speech is useful for many different purposes. For example, products and systems based on speech recognition can use emotion data to better understand human meaning, going beyond just plain text words to improve the accuracy of the system's response.

For our project we address the problem of emotion recognition by directly comparing a number of different feature sets and models (Support Vector Machines, Neural Networks, Logistic Regression and Gaussian Mixture Models), a comparison that to our knowledge has not been done before. We use a data set with 590 short speech samples. We find that SVM and a simple Neural Net with a large number of features provide the best results on the test set, at a comparable 67-70-% accuracy rate. We believe larger data sets could improve such results further.

Literature review
A great overview of speech recognition methods is provided by Ayadi, 2011. The paper cites that highly predictive features are still missing, and the majority of methods achieve below 80% of accuracy for speaker-independent detection of multiple emotions (although higher for speaker-dependent models). It notes it is difficult to state which model serves the problem best, given almost all research has been performed on different speech databases using different measurement approaches.

Seemingly one of the best predictive models is laid out in Chavhan, 2010. Using a large set of features (capturing pitch and energy, mel-frequency spectrum coefficients, mel-energy spectrum dynamic predictive spectrum coding and predictive spectrum coding) in an SVM model, achieving > 90% accuracy.

A hidden Markov / GMM approach is described in Schuller, 2010. It takes both a temporal (using features of small frames over time) and global (using features of the full sample) approach, and finds the latter works best, achieving an 86% accuracy rate.

On Neural Networks, Hendy, 2013, provides comparative analysis of different ANN approaches to classify emotions from speech. It concludes that with proper features selection and ANN architecture it is possible to reach accuracy rate of 85% in classifying samples into 7 different emotions. It also provides strong indication that proper feature standardization may significantly affect the accuracy of the prediction. (Glüge 2011) introduced segmented-memory recurrent neural networks (SMRNN) to classify emotions in speech. The work demonstrated that SMRNN can be trained to deliver 70% accuracy on the test set. Application of Deep bidirectional LSTM to emotion in speech classification was evaluated in (Lee 2015), achieving approximately 65% accuracy. Both Glüge, 2011, and Lee, 2015, considered non-overlapping segments in the utterances. While it is not possible to directly compare the results from different NN approaches, the reported results indicate that further work on proper feature selection and varying network architecture still has opportunity to improve the accuracy.

Data
Although no public data sources (labeled speech fragments) were available for the task of speech recognition (something much needed to advance further research), we were able to get access to the "Emotional Prosody Speech and Transcripts" database through Stanford. The data consists of words being uttered under certain emotion, around 1 second per sample. The number of samples is as follows:

	Training set (# samples)	**Testing set (# samples)**
Anger	105	35
Despair	157	35
Happiness	146	35
Neutral	55	25
Sadness	127	35
Total	590	165

Baseline: logistic regression approach
The model we decided to use for the baseline approach is a logistic regression for emotion classification. We implemented this using scikit-learn (as with most methods in this paper).

The features we decided to use are loudness, intensity and fundamental frequency, as these make intuitive sense to relate to emotions. The features are all taken globally over the full sample (i.e. not splitting the audio up into frames). For every feature, we will use the first, second and first quartiles as well as the standard deviation. This gives us a total of 12 unique features per sample. We used OpenSMILE 2.3 for feature extraction over our full dataset (again often used throughout this paper).

Results: we get a 46.3% accuracy rate on the training set, and for the test set 35.8%. These values are well above the random chance mean of 20%, although still at fairly low accuracy. Results on the test set by category are as follows:

Baseline: 35.8% accuracy	**Recall**	**Precision**
Anger	60%	78%
Despair	23%	18%
Happiness	26%	31%
Neutral	8%	25%
Sadness	54%	34%

It is clear that anger is by far the best predicted emotion. This makes logical sense, as anger is more clearly distinct in intensity, loudness and frequency variation than the other emotions.

Of the features, loudness is the most strongly predictive. In fact, the fit on the test data set can be improved to 38.8% by only including 2 features: loudness standard deviation (most predictive feature), and loudness third quartile (second most predictive). This suggests that the other features contributed little, and perhaps caused overfitting on the training set.

Oracle: OpenEAR auto classification approach

As “oracle” method we have used the approach described by Eyben et al, 2009 without altering anything. OpenEAR provides the mechanism to extract 988 unique features from the audio files, including Mel-Spectrum cepstral coefficients, Mel-frequency cepstral coefficients, line spectral frequencies, intensity, loudness, zero-crossings, pitch, and different statistical transformations of these. Once features are extracted, a pre-configured SVM is used to train the model on the training data, which is then used to classify the test samples. The SVM is implemented by publicly available libsvm package (Chang, et al.). This approach has been previously applied to classify data from Berlin Speech Emotion Database (Burkhardt et al, 2005). Extracted features were scaled into -1, + 1 range, and trained using C-SVC svm-type and radial-basis kernel function.

Results: the accuracy on the training set was 65.9% (389/590), and on the test set 58.2% (95/165). Results on the test set by category:

Oracle: 58.2% accuracy	**Recall**	**Precision**
Anger	77%	87%
Despair	45%	41%
Happiness	45%	47%
Neutral	88%	73%
Sadness	42%	48%

Support Vector Machine approach

For our own experimentation with an SVM approach, we will start by loosely copying the approach laid out in Chavhan, 2010. The features this paper has found work best, confirmed on two different data sets, are energy (intensity), mel-frequency spectrum coefficients (MFCC) and mel-energy spectrum dynamic coefficients (MEDC).

We start by using the mean and standard deviation for these features, and find that these features already improves the logistic regression test accuracy 47.9%. For the SVM we pick a non-linear kernel and scale the data (to zero mean and unit variance). This gives us a 45.5% accuracy (below logistic regression).

Subsequently, we add many additional relevant features: the first and third quartiles, the first derivative of all features, and the first 12 orders for MFCC (72 features total). This brings us to a 70.3% test accuracy, a remarkably good result and surpassing the SVM approach from the “oracle” (shedding doubt on its classification as “oracle”). Results by category are as follows:

SVM: 70.3% accuracy	**Recall**	**Precision**
Anger	66%	100%
Despair	49%	47%
Happiness	97%	68%
Neutral	88%	92%
Sadness	57%	63%

Note we are using a non-linear kernel for this. The same test with a linear kernel gives significantly inferior results at 50.6% test set accuracy (c.f. 70.3% with a non-linear kernel). This hints at significant non-linearities.

Gaussian Mixture Model approach

For the Gaussian Mixture Model approach we will rely on Schuller, 2003, who uses global features in a Gaussian Mixture Model with great results. We train a separate Gaussian Mixture Model for every emotion using the training data, and then have every model provide scores for every sample in the test set data. The emotion with the highest score is chosen as the prediction.

We will continue to use the same features we used for the Support Vector Machine model, given these have already shown to provide high explanatory power. However, initial tests show that using all 72 features used previously leads to overfitting with near 100% training set accuracy and only 44.2% test set accuracy (inferior to the SVM results).

To counter-act the overfitting, we strive to find the most valuable features by looking at the coefficients (weights) that the Logistic Regression and SVM models apply to each feature. We take the absolute coefficients for every category and sum these, giving us the “sum of absolute coefficients”. We then rank the features using this metric (highest sum is most important feature). We find that both Logistic Regression and SVM show a preference for the same features. The top 10 features are shown in below table.

	Sum of abs. coefs.		Feature rank		
Feature	**LogReg**	**SVM**	**LogReg**	**SVM**	**Combined**
mfcc_sma[0]_amean	62.99	4.41	2	1	1
pcm_Mag_melspec_sma_de_de[0]_quartile1	69.16	3.78	1	2	1
pcm_Mag_melspec_sma[0]_amean	55.69	3.68	3	3	3
mfcc_sma[1]_amean	42.08	2.96	4	4	4
mfcc_sma[3]_amean	29.49	2.70	6	8	5
pcm_LOGenergy_sma_stddev	31.20	2.33	5	10	6
mfcc_sma[2]_quartile1	26.56	2.77	10	5	6
pcm_LOGenergy_sma_amean	28.98	2.55	7	9	8

mfcc_sma[0]_quartile1	27.91	2.15	8	11	9
mfcc_sma[2]_amean	25.33	2.71	12	7	9

Using these top 10 features in our GMM model gives 65.7% training set accuracy and 45.5% test set accuracy, indicating less overfitting and a slight improvement in accuracy. Subsequently, we tested with various different numbers of features (5, 10, 20, 30, 40) and found the optimal test set accuracy to reside around 30 features, leading to 58.8% test set accuracy (95.6% training set accuracy). See table below for the full results.

We also experimented with more than 1 Gaussian component per category, but found this worsened the problem of overfitting, and did not provide superior results (52.1% as best test set accuracy vs. 58.8% using only 1 component):

# Features	5	10	20	30	40	10	20	30
# Components	1	1	1	1	1	2	2	2
Train acc.	49.8%	65.7%	87.3%	95.6%	98.3%	80.7%	95.1%	99.0%
Test acc.	37.0%	45.5%	56.4%	58.8%	56.9%	44.2%	52.1%	46.7%

Given the overfitting issues and limitations this imposes on the number of features and Gaussian components we can use, we believe that the GMM method's results could be significantly improved by obtaining more training data. We report the recall & precision for the highest achieved test set accuracy (using 30 features and 1 component) below.

GMM: 58.8% accuracy	**Recall**	**Precision**
Anger	86.2%	71.4%
Despair	42.4%	40.0%
Happiness	54.9%	80.0%
Neutral	88.9%	32.0%
Sadness	51.2%	62.9%

Neural Net approach

We have experimented with two different types of neural networks to classify the audio recordings into the emotion categories - multilayer perceptron and recurrent neural network.

Multilayer perceptron network

For the input data for multilayer perceptron network we have used the same 72 feature set we have previously used to train the SVM model. Our initial network consisted of two identical hidden layers, consisting of 512 hidden neurons. The output from each layer was processed through rectified linear unit (ReLU). The network was implemented using the TensorFlow framework.

The network could easily be trained to produce 100% accuracy on the training set. The following results were produced on the test set:

MLP: 58.78% accuracy	**Recall**	**Precision**
Anger	40.0%	77.7%
Despair	45.7%	51.6%
Happiness	80.0%	51.8%
Neutral	76.0%	65.5%
Sadness	57.1%	60.6%

We have also experimented with varying the number of hidden layers (2, 3, 4) and changing the number of the neurons in the layers (16, 32, 64, 128, 256, 512). None of the experiments produced better results, with most results being very similar to the results above.

Based on the observed results we have concluded that our network is significantly overfitted. We have tried the following approaches to fight against the overfit:

1. Decreasing the number of iterations during the training:

Fixing the other parameters of the network, we have tried to stop training when the accuracy of the model on the training set varied between 60% and 90%. In all the experiments the accuracy of the trained model on the test data did not improve compared to the original results.

2. Applying l2-regularization:

We have applied l2-regularization to weights on both hidden layers and output layer and on the corresponding biases. We have observed that the accuracy of the trained model on the training set maxed out at 70.3%. The corresponding results on the test set:

MLP with L2-regularization: 43.0% accuracy	**Recall**	**Precision**
Anger	25.7%	100.0%
Despair	74.2%	30.1%
Happiness	77.1%	46.5%
Neutral	32.0%	61.5%
Sadness	2.3%	100.0%

3. Applying dropout:

Dropout is another popular approach to avoid overfitting of the neural network. We have experimented dropout to both hidden layers with keep probability varying between 0.75 and 0.95. We have observed

that decreasing the keep probability significantly decreased accuracy (below 0.5) on both training and test datasets.

Finally, we have decided to see how the preprocessing of the data could affect the performance of the network. We have processed the input data using standard MinMax scaling. We have also used single hidden layer with the 512 hidden neurons. We have observed 91.7% accuracy of the model on the training set. The results of the model on the test set were the best we have observed so far - 67.3%

MLP with single layer, MinMax scaling: 67.3% accuracy	**Recall**	**Precision**
Anger	74.2%	96.3%
Despair	48.5%	44.7%
Happiness	77.1%	65.9%
Neutral	84.0%	80.8%
Sadness	57.1%	60.6%

Increasing the number of hidden layers or decreasing the number of the neurons in the layers did not improve the performance of the network.

Recurrent neural network

As speech sentiment has time duration, we have decided to experiment with RNN to classify the recordings. Using the TensorFlow framework we have implemented simple LSTM network consisting of single LSTM cell of 512 hidden neurons. For this experiment we have used different method to extract features from the inputs. We have used Bregman Audio Visual information toolbox to extract 95 MFCC, 12 Chroma and 1 Power features. The extracted features are collected per overlapping window size. As all the recordings have slightly different number of frames, we have limited the number of collected frames by the minimum length among the complete data set - 64. Hence, we have constructed an RNN consisting of 64 steps.

We have trained the model on the complete feature set and observed 100% accuracy for the training set. The corresponding result for the test data set was 45.5%:

RNN: 45.5% accuracy	**Recall**	**Precision**
Anger	42.6%	55.5%
Despair	20.0%	46.6%
Happiness	57.1%	33.3%
Neutral	64.0%	88.8%
Sadness	48.5%	37.3%

To find out how the selection of features affects the performance of the network, we have trained the same model using only Chroma and Power feature selection, expecting that power and tone variation are the mostly significant indicators of the emotions. Even on the training set the accuracy of the trained model was maxed out at 34.7%. The accuracy on the test data set was at 25.5%.

In similar experiment we have extended Chroma and Power feature with the MFCC[1:13] features, most commonly used in speech recognition applications. The observed results were very similar to the results previously observed for the complete feature set.

Combined our experiments with Neural Network approaches to classify sentiments into emotion categories did not produce overly promising results. While it is relatively easy to achieve high accuracy on the training set, the accuracy on the test data set is worse than results produced by the simpler SVM method. Common methods to avoid overfitting did not improve the classification results. The experiments demonstrate that proper feature selection and data preprocessing may be most important to obtain better results. We also believe that at least partially, poor performance of the network approach in our study is caused by the limited size of the training data set.

Conclusion

We implemented various methods for classifying speech utterances into one of five emotion categories and compared the results. We observed that accuracy of the classification varies from as low as 35% using naive logistic classifier with a few features to around 70% using SVM or a simple neural network. 70% accuracy reached by our classification methods matches the accuracy of the methods previously reported in the literature.

In our GMM and Neural Network approaches we have observed that the models may be trained to reach near 100% accuracy on the training set, producing much lower accuracy on the test set, indicating overfitting. We have tried several approaches to counteract the overfitting, and observed that careful feature selection and preprocessing of the features have the most profound effect on the accuracy of the models. This observation correlates with previously published comments that feature selection and data preprocessing is the mostly critical component of any speech to emotion classifier.

Our study similarly to most reported works suffered from limited training data. Difficulty in the assembly of such data sets is partially caused by artificial nature of the collection procedure - the actors are asked to produce the “emotional” reading of some raw texts. Collecting data from natural speech recordings using manual annotations could be also complicated as emotion recognition by humans is ambiguous. Probably because of the lack of large data sets consisting enough data, ANNs were not the most used method to classify emotions from speech. In this study we have only experimented with naive DNN and RNN architectures. The reached accuracy strongly suggests that to counter overfitting, feature selection and feature preprocessing should be the focus of future work to get better results. Finally, experimenting with more sophisticated RNNs capable of picking up dependencies in the time series of non-global (local, time based) features extracted from the utterances could be another way to improve the accuracy of classification.

Appendix

The code implementing methods described in this work is available at https://github.com/jorisvanmens/speech_emotion_recognition.

Works Cited

Ayadi, Moataz El, Mohamed S. Kamel, and Fakhri Karray. "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases." Pattern Recognition 44.3 (2011): 572-87. Print.

“Bregman Audio-Visual Information Toolbox¶." Bregman Audio-Visual Information Toolbox — Bregman V1.0 Documentation - http://digitalmusics.dartmouth.edu/~mcasey/bregman/.

Burkhardt, Paeschke, Rolfes, Sendlmeier, and Weiss. A database of german emotional speech. In Proceedings Interspeech 2005, Lissabon, Portugal, pages 1517–1520, 2005.

Chang, Chih-Chung, and Chih-Jen Lin. "Libsvm." *ACM Transactions on Intelligent Systems and Technology* 2.3 (2011): 1-27. Print.

Chavhan, Yashpalsing, M. L. Dhore, and Pallavi Yesaware. "Speech Emotion Recognition Using Support Vector Machine." *International Journal of Computer Applications* 1.20 (2010): 8-11. Print.

Eyben, Florian, Martin Wollmer, and Bjorn Schuller. "OpenEAR — Introducing the Munich Open-source Emotion and Affect Recognition Toolkit." *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (2009). Print.

Glüge, S., Böck, R., and Wendemuth A. 2011. “Segmented-memory recurrent neural networks versus hidden Markov models in emotion recognition from speech”. Conference: Proceedings of the International Conference on Neural Computation Theory and Applications (NCTA 2011)

Hendy, N. and Farag, H. “Emotion Recognition Using Neural Network: A Comparative Study”. 2013 International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:7, No:3.

Koolagudi, Shashidhar G., and K. Sreenivasa Rao. "Emotion Recognition from Speech: A Review." International Journal of Speech Technology 15.2 (2012): 99-117. Print.

Lee, J., Tashev, I. 2015 “High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition” International Speech Communication Association (Interspeech) 2015.

Schuller, B., G. Rigoll, and M. Lang. "Hidden Markov Model-based Speech Emotion Recognition." 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). Print.

Shaw, A., Kumar, R., and Saxena, S. “Emotion Recognition and Classification in Speech using Artificial Neural Networks”. 2016 International Journal of Computer Applications Volume 145 – No.8

Ververidis, D., and C. Kotropoulos. "Emotional Speech Classification Using Gaussian Mixture Models." *2005 IEEE International Symposium on Circuits and Systems*. Print.