CSE5539 Term Project Report – A Measure to Score Emotions in Presidential Election Speeches with Multi-Classifier System

Yiran "Lawrence" Luo, DeLiang Wang
Department of Computer Science and Engineering, The Ohio State University, Columbus
luo.81@osu.edu, dwang@cse.ohio-state.edu

Abstract

In this report we introduce a subtle approach to recognizing and measuring human emotions in speeches made by the US Presidential Election candidates. We propose a multi-classifier system which is based on Mel-frequency Cepstrum Coefficient (MFCC) feature extraction and Deep Neural Network (DNN) classification. And several methods of generating feature vectors, as well as differently configured DNN classifiers are tested in comparisons for the best training accuracy. We present the test results via a synthesized emotion scoring algorithm regarding 5 co-existing emotions, and we are able to reach a decent big picture view that corresponds well with observations made by news media.

1. Introduction

Speech Emotion Recognition (SER) has long been a promising task within Speech Recognition (SR) studies. It has seen potential usages in lie detection, education and even video gaming [1]. But unlike Facial Emotion Detection in computer vision which is born visualized and thus more straightforward to human senses, emotions in acoustic speeches are less obvious and more often treated as latent characteristics behind voice signals. To uncover such hidden information for technical purposes, due to the lack of emotion classification standards, previous researchers have favored the way to focus only on a finite set of emotions, and such methodology has been widely accepted that it leads to a vast number of speech corpora being published every year where emotions are categorized in diverse ways [3].

Nonetheless, with the range of emotions assumed, conventional approaches to classifying a discrete number of emotions are to use generalized feature models such as Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) or the hybrid of the two [4]. More recently, thanks to the growing capability of large-scale machine learning, researchers have been applying machine learning techniques, especially DNN models on SR tasks which greatly outperform early benchmarks [5]. They also have attempted combining conventional emotion classification models with Support Vector Machine (SVM) for better recognition performance [6].

In this article, we furthermore explore the potential of utilizing DNN for emotion recognition, with introducing a scoring metric to measure, and hopefully to help standardize, speech emotion categorization into a finite set. Being a primitive study, this project solely concentrates on analyzing the 2016 US Presidential Election speeches by male candidates.

The rest of the report is organized as follows. In Section 2, the fundamental components of our classification model and the emotion scoring metric are described. In Section 3, the training method comparison and several early stage experimental results are showcased. And finally, conclusions and future works are presented in Section 4.

2. Methodology

This section is to present the basic components of our model to analyze emotions in speech. The entire procedure consists of three major stages: feature extraction, classification and scoring.

2.1 Feature Extraction

Mel-frequency Cepstrum Coefficients (MFCC) are widely used features in speech recognition. MFCC algorithm is a state-of-the-art method that mimics how human cochlea perceives energies at different frequency levels from incoming audio sources. How to conduct MFCC extraction is well explained in [9]. Earlier studies in SER have already demonstrated that such spectral features are highly correlated to emotion information and, as a result, do provide good performances such as in [7] and [8].

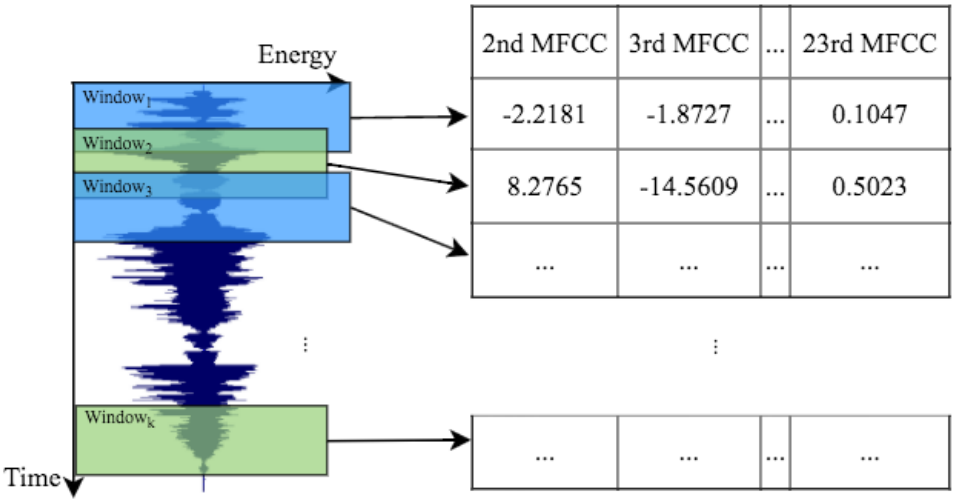

Figure 1. How an Energy-Time speech signal becomes a sequence of MFCC feature vectors. The numbers in the feature table on the right side are for demonstration purposes only.

Figure 1 visualizes an intuition how MFCC feature extraction is implemented in this project given a speech utterance. The utterance is scanned with a sequence of equal-length windows. Each window samples a frame that contains a sequence of energy values from the utterance. Frame steps are also of equal length and are usually smaller than the frame length so as to allow overlapping and to cover a fair extent of context information. Eventually, each frame's energy sequence is converted into MFCC vectors. The 1st MFCC is normally dropped for being an overall measure of signal loudness and therefore not containing useful spectral information. In this project, we use 22 MFCC features throughout both training and testing.

With the frame-based features extracted, we also attempt to treat an entire speech signal as one comprehensive feature vector. This is done by simply concatenating all the frame-based MFCC feature vectors in order of time into one lengthy feature vector. Since the following classification stage requires inputs be of the same dimensionality, we apply both zero-padding and trimming to unify the dimensions of the concatenated feature vectors

2.2 Multi-Classifier Classification Model

Many previous SER studies involve in using the multi-class classification as the core classifier methodology, such as in [10] [5] and [6]. Most importantly, before the final output, a *softmax* layer is deployed for a generalized regression result regarding all the classes spontaneously. This approach, however, leads to the concern regarding the psychological discovery that multiple human emotions may co-exist mutually in daily speeches. Thus human speeches are able to bear strong traits of seemingly contradicting emotions at the same time [2]. For that reason, here we present an alternative architecture that uses a multi-classifier model inspired by the text-based opinion mining [11]. We use parallel and independent classifiers to filter out, and individually learn of each emotion from the sources. By doing this, both the mutually strong emotions and the mutually weak emotions are legitimate output patterns out of the classifier bundle.

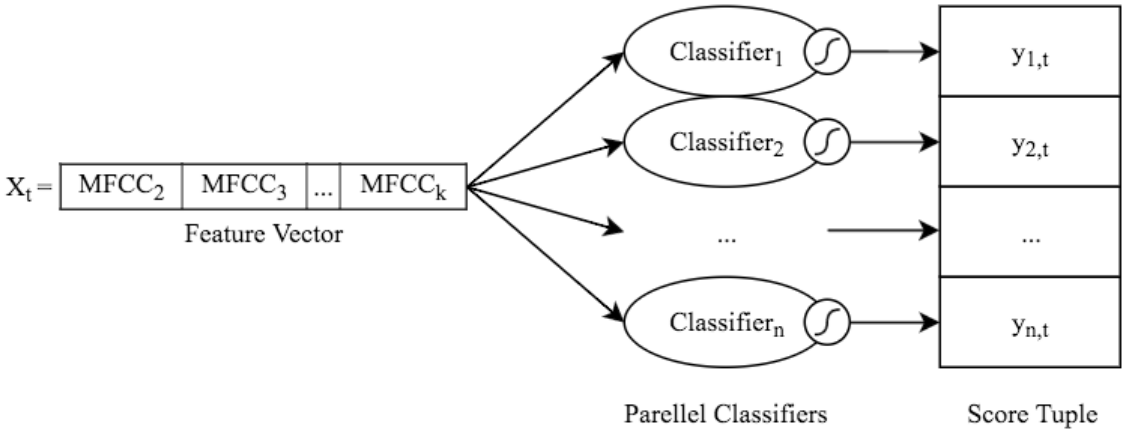

Figure 2. The Multi-Classifier Classification procedure, esp. during testing.

Figure 2 demonstrates the primary design of the classification procedure. X_t represent a feature vector extracted at frame *t*. The feature vector is then duplicated and passed to *n* identically configured classifiers.

Each classifier only concentrates on one emotion and is independent from other classifiers. For example, in a 3-classifier setting with only *happy*, *sad*, and *angry* labeled emotions, Classifier_1 is assigned to filter out only *happy* characteristics. Similarly, Classifier_2 and Classifier_3 are for *sad* and *angry* emotions respectively. The output of a classifier is through a sigmoid function and serves a scalar score that ranges from 0 to 1.

The classifiers are trained online with binary labels. We assume every one feature vector from the training set corresponds to one most dominant emotion. For example, in the 3-classifier setting mentioned above, one feature vector from a *happy* labeled training utterance is mapped with the training label tuple (1, 0, 0). So taking this feature vector, technically, Classifier_1 is trained with the expected output 1, while Classifier_2 and Classifier_3 are both trained with 0. Besides for implementation convenience, such training labels are adapted since the ratio of the remaining less dominant emotions are not assumable. Hence it is intuitive to simply discard the ambiguous emotion values.

2.3 Emotion Scoring Algorithm

The emotion scoring algorithm is used in the testing phase. It is to provide a referential measure that approximates the dominations of different emotions in a speech. Two interfaces of generating emotion score tuples are implemented in this project.

2.3.1 Frame-wise Score

The first interface is the frame-wise score F. This score tuple comes from the direct prediction outputs of the classifier group shown on the right hand side of Figure 2.

$$score_{i,t} = \frac{y_{i,t}}{\sum_{j=1}^{N} y_{j,t}} \quad i = 1, 2, \dots, N \qquad \dots (1)$$

$$\mathrm{F_t} = (score_{1,t}, score_{2,t}, \dots, score_{N,t}) \qquad \dots (2)$$

t stands for the current frame index. *N* stands for the total number of classifiers. $y_{i,t}$ stands for the output of Classifier_i at frame *t*. Each frame-wise score tuple $\mathrm{F_t}$ is normalized for frame *t* and it mimics the speaker's emotion proportion during an instantaneous period of time.

2.3.2 Utterance-wise Score

The second interface is the utterance-wise score U. This score tuple consists of the synthesized scores from each classifier.

$$syn\_score_i = \sum_{t=1}^{T} y_{i,t}, i = 1, 2, \dots, N \quad \dots (3)$$

$$\mathrm{U} = (syn\_score_1, syn\_score_2, \dots, syn\_score_N) \quad \dots (4)$$

T stands for the total number of frames the utterance has. *N* stands for the total number of classifiers. $y_{i,t}$ stands for the output of Classifier$_i$ at frame *t*.The synthesized score syn_score_i regarding Classifier$_i$ is simply the sum of all its prediction outputs over time. The utterance-wise score U is the collection of such synthesized scores of all the classifiers and its component values are not normalized unlike the frame-wise score. This score tuple is to offer an overall view of emotion dominance with regard to an entire speech utterance.

3. Experiments

This section includes experimental preprocessing details, comparisons for the best training accuracy and showcases of several early stage results.

3.1 Setup

The training dataset uses the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [12]. All audio files are sampled at 48kHz. This project uses files of 5 emotions/classes {*happy, sad, angry, disgusted,* and *surprised*} performed by 12 male North American English-speaking actors. Each audio sample addresses a certain sentence statement of either “Kids are talking by the door” or “Dogs are sitting by the door”. And each sample is in either strong or normal intensity level. Also, each identical statement made by the same actor at the same intensity level is recorded twice. Last but not least, all training files are stripped off silences longer than 0.1 seconds. Overall, each emotion class contains 192 preprocessed audio files, and in total 480 audio samples ranging from 0.5 seconds to 3 seconds are used as the training set.

The MFCC feature extraction uses 48kHz sample rate, 25 ms window length and 10 ms window step. 22 coefficients (2nd to 23rd) are used to form a frame-based feature vector. For the training data in all, we obtain 16265 vectors of *happy*, 16270 vectors of *sad*, 18543 vectors of *angry*, 18839 vectors of *disgusted*, and 14536 vectors of *surprised*.

The two-layer feed-forward DNN classifier is chosen as the single emotion classifier. Several configurations are tested for the best training performance. In the format of (neuron number of

first layer-neuron number of second layer), structures of 100-100, 200-200, 500-200, and 1000-200 are tested for both training accuracy and training time cost. Other shared training configurations of the DNN classifiers include 0.2 Dropout rate to prevent overfitting, Mean Squared Error for calculating losses, RELU activation function for all hidden layers, sufficient epochs for training a classifier, and a mini-batch size of 250 with Stochastic Gradient Descent to speed up training. The classifiers are built with Keras ver. 0.3.3, an open-source deep learning framework in Python [13].

The test data are collected from CNN Debate Podcasts. We capture several multi-second long speeches by Donald Trump pivoting his key opinions during the South Caroline Republican Town Hall on Feb 19, 2016 [14]. And then we compare our emotion analytics to the findings made by mainstream news media such as CNN Politics and BBC News [15] [16] [17].

3.2 Training Performance

Feature Space	Concat. Trimmed	Concat. Zero-padded	Frame-based Features			
Classifier	100-100 DNN			200-200 DNN	500-200 DNN	1000-200 DNN
Training Error	0.1972	0.1981	0.1130	0.0942	0.0800	0.0705
Time Cost	<10 sec	<10 sec	~1 min	~3 min	~6 min	~15 min

Table 1. The training performance comparison. The error is the average training error of the five chosen emotion classifiers. The time cost is the average time cost to complete training one classifier.

The training performance comparison results of using the comprehensive feature vectors vs the frame based features, and the performance comparison of the number of hidden neurons are all migrated into Table 1. Apparently, concatenating all the frame features into one vector for each utterance has little use for classification. Since their training error is extremely close to 0.20 which is the exact proportion of one emotion in the training set, using these features is no better than randomization. On the other hand, the classifier with the most hidden neurons has the best overall training performance, albeit the relatively lengthy time to complete training.

3.3 Test Scoring Results

This section showcases using the frame-wise score to plot emotion trends and using the utterance-wise score to provide a big picture view of emotion dominance. The presentations shown in Figure 3 and Table 2 are about the utterance “I have a lot of respect for the Pope”. The utterance, which lasts for roughly 1.5 seconds, is converted into a 100-frame MFCC feature matrix.

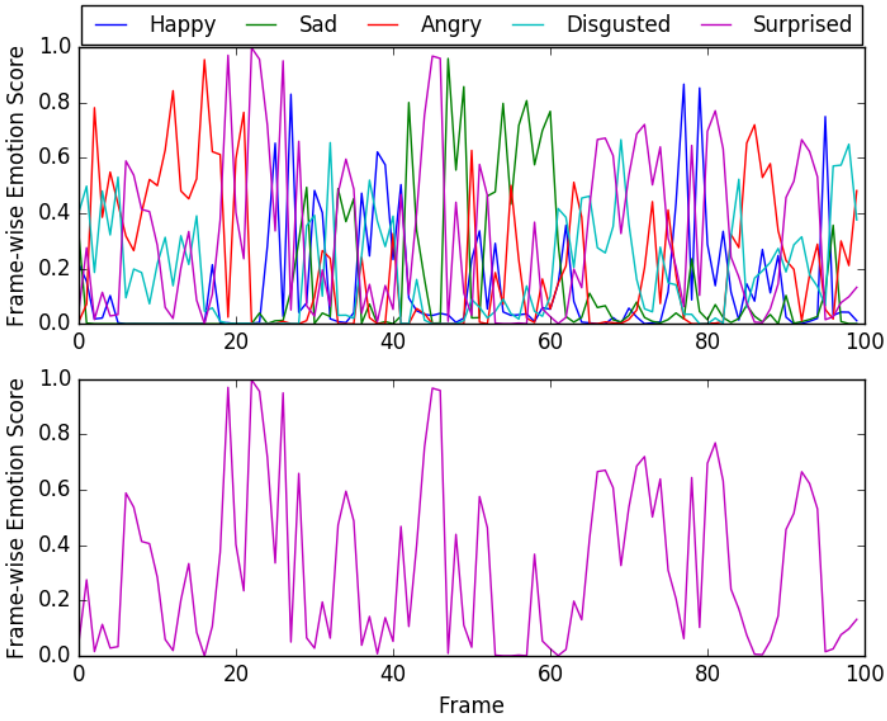

Figure 3. Tracks of the frame-wise emotion score changes through the utterance of the Pope comment. The lower subplot picks the most dominant emotion of the five which is *surprised*.

Happy	Sad	Angry	Disgusted	Surprised
7.68853533	12.179586	17.25051569	15.78155226	**27.65836445**

Table 2. The utterance-wise emotion scores regarding Donald Trump's opinion on the Pope.

We are also able to align the utterance-wise emotion scores up with observations made by news sources in Table 3. Even though being a subjective validation, the following result table does present a decently grounded consistency in the big picture with media investigations, especially about the most prominent emotion(s).

4. Conclusions and Future Works

As a summary, this project manages to reach a consistent metric through early-stage experiments. Utilizing the Multi-Classifier approach along with DNN and MFCC, this proposed measure is capable of not only tracking the most dominant emotion in order of time, but also keeping the instantaneous mutuality of multiple strong emotions. In all, this methodology, although being comparatively naïve, is able to grasp a big picture insight into a speaker's performance by serializing his/her emotions.

# of Frames	Utterance Transcripts	Utterance-wise Emotion Scores	Media Comments
100	"…I have a lot of respect for the Pope…" [14]	Happy: 7.68853533 Sad: 12.179586 Angry: 17.25051569 Disgusted: 15.78155226 **Surprised: 27.65836445**	Trump even praised Pope Francis, … [15]
	"(The protesters are professionals) … and making noise. And I think it's a disgrace." [14]	Happy: 5.71182216 **Sad: 28.71805771** Angry: 9.19883203 Disgusted: 13.4112369 **Surprised: 28.26004205**	Mr. Trump was repeatedly interrupted by the protesters, whom he called a "disgrace". [16]
200	"(Ted Cruz) did a voter violation notice, looks like right under the IRS…"[14]	Happy: 36.21011501 Sad: 8.82838024 Angry: 25.66392993 **Disgusted:45.36478681** Surprised: 26.58861101	He quipped that other tactics the Cruz campaign has used in this election were "disgusting." [17]

Table 3. The utterance-wise emotion scores of Donald Trump compared to media comments. The highest scored emotions are marked in bold.

Future directions of this project should focus on solidifying the foundations of this approach. On one hand, other than using the spectral features like MFCC, temporal features shall be also taken into consideration and be tested for training performance. Also, since this project only involves in analyzing male speakers, female speakers will be a task subject in future, and we would like to compare the subtle differences between the two genders on topics such as the emotion trends and the score values. Last but not least, a crowd-sourcing accuracy test on the emotion scores should be conveyed in order to furthermore validate how legit our proposed emotion scoring algorithm is.

5.Acknowledgements

We would like to thank Professor DeLiang Wang for his support throughout the project. Less would have been achieved without his tremendous advices.

References

1. Cowie, R. *et al*., "Emotion recognition in human-computer interaction," (2001). in *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80, Jan 2001.
2. Williams, P., and Aaker, J. L. “Can Mixed Emotions Peacefully Coexist?”. (2002). *Journal of Consumer Research*, pp. 636–649, Apr 28, 2002.
3. Schuller, B., Rigoll, G. and Lang, M. "Hidden Markov model-based speech emotion recognition." (2003). *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, pp. II-1-4 vol.2.
4. Jiang, D. N. and Cai, L. H. "Speech emotion classification with the combination of statistic features and temporal features." (2004). *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, 2004, vol. 3, pp. 1967-1970.
5. Hinton, G., *et al*. “Deep Neural Networks for Acoustic Modeling in Speech Recognition”. (2012). *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, Nov 2012.
6. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier G. and Schuller B., "Deep neural networks for acoustic emotion recognition: Raising the benchmarks." (2011). *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011, pp. 5688-5691.
7. Slaney, M., McRoberts, G. (2003) Baby ears: a recognition system for affective vocalizations. *Speech Commun*, vol 39, no. 3-4, pp. 367-384.
8. Krishna Kishore, K.V. and Satish, P. K. "Emotion recognition in speech using MFCC and wavelet features," (2013). *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, Ghaziabad, 2013, pp. 842-847.
9. Davis, S. Mermelstein, P. (1980). “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28 no. 4, pp. 357-366.
10. Ververidis, D. and Kotropoulos, C., (2006). “Emotional speech recognition: Resources, features, and methods.” *Speech communication*, 48(9), pp.1162-1181.
11. Batista, L. B. and Ratte, S. (2012). A Multi-Classifier System for Sentiment Analysis and Opinion Mining. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (ASONAM '12). IEEE Computer Society, Washington, DC, USA, pp. 96-100.
12. Livingstone, S. R., Peck, K., & Russo, F. A. (2012). “RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song.” *The 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS), Kingston, ON.*
13. Chollet and François. (2015). Keras. *GitHub Repository*. https://github.com/fchollet/keras .
14. “Donald Trump - Republican Town Hall, South Carolina”. *CNN Debates by CNN on ITunes.* ITunes.

15. Bradner, E. "5 takeaways from the CNN Republican town hall." *CNN*. Cable News Network. Feb 19, 2016. http://www.cnn.com/2016/02/19/politics/republican-town-hall-recap/ .
16. "Trump calls off Chicago rally following violent clashes." *BBC NEWS*. BBC. Mar 12, 2016. http://www.bbc.com/news/election-us-2016-35791008
17. Lee, MJ. " Donald Trump challenged over 9/11, Iraq War comments." *CNN*. Cable News Network. Feb 19, 2016. http://www.cnn.com/2016/02/18/politics/republican-town-hall-highlights/ .