Detecting Emotion in Speech with a Tiered Machine Learning Model

Elizabeth Ling

Abstract

With the rise of new Artificial Intelligence (AI) technologies in all walks of life, the problem of speech emotion recognition is vital to improving human- computer interactions. Recent research has proposed a myriad of feature extraction methods and classifiers ranging from human and animal based to new structures. Without consistent accuracies across methods, this paper aims to propose a new model structure for improved accuracy and consistency. The proposed tiered method combines the benefits of generalizing emotion classes and improved accuracy on binary classifiers with the detailed learning power of neural nets. Using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Toronto Emotional Speech Set (TESS), it achieves accuracies of .8626 and .9760. These findings are a step towards a better, more consistent speech emotion recognition model.

Keywords

Speech Emotion Recognition — Machine Learning — Clustering

Contents

1. Introduction

With the rise of new Artificial Intelligence (AI) technologies in all walks of life - health care, economics, automobiles, our society is moving toward a future integrated with machines completing every day tasks. Machine learning algorithms power companies including Google, Coca-cola, IBM, Netflix and Disney, with over 85% of online advertising coming from AI algorithm based suggestions [1]. Subsequently, the need for enhanced human-computer interaction is becoming more prevalent. Speech emotion recognition (SER) is a growing research field because of its crucial role in enhancing human-computer interactions [2]. Important emotional cues are contained in vocal cues in speech, which allow listeners to identify additional meaning within dialogue [3, 4]. Speech emotion recognition's practical applications range from medical stress and pain detection, social coaching, interactions with robots, call centers, and more [5].

Speech emotion recognition consists of two major parts: feature extraction and the predictive model. As early as the 1980s, simple statistical analysis of acoustic features of speech was used as a means to analyze emotion. Later, a hidden Markov model was used to classify emotions in speech, finding a 7% increase in accuracies when suprasegmental information was added [6]. However, debate continues on which features are the most influential in recognizing emotion from speech signals, with Mel Frequency Cepstral Coefficients (MFCCs) and formants being two of the most prominent [7, 8]. Various feature packages, such as openSMILE and Praat, can be useful for generating a large number of low-level descriptors, which I utilize in this paper [9, 10]. Researchers have also proposed a feature selection methods to mathematically reduce the number of features used to only the most statistically influential [11]. Furthermore, He and Cao introduce a mix of hand-crafted and deep-learned features, while some other findings report that hand-crafted features are not as effective [12]. In recent years, SER has been explored with k-means clustering, SVMs, Artificial Neural Nets, and hidden Markov models [13]. Additionally, deep learning methods are increasingly visible, including deep belief networks, deep LSTMs and deep convolutional neural nets [14]. Recent research utilizes a brain-based model to move the data through various frameworks that feed into each other, or a whale-based learning model to find optimize weights for a neural net. [15, 16]

Without consistent accuracies across methods, this paper aims to propose a new model structure for improved accuracy and consistency. I propose a novel network structure for the classifier to improve speech emotion recognition performance.

Based on previous research, generalizing and grouping emotions into binary classifiers improves performance. [9] Mordkovich et al. used binary classfiers to classify one or two emotions against all the remaining emotions, which yielded a 15 to 20% error compared to 40% for larger clusters. Similarly, He and Xia propose a Joint Binary Neural Network to predict multiple binary classifications to create a multi-label output [17]. Thus, their model is able to identify co-existing emotions in a sentence with by answering an obviously simpler question: determining whether the sentiment is love or not love, joy or not joy, etc. Our model uses this idea to combine the effectiveness of a binary classifier with the detailed deep learning of a neural net.

2. Methods

2.1 Data Sets

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a publicly available validated multimodal database. [18] Only the speech audio data is used in this paper, which consists of 1440 files which each contain one of two spoken sentences. The emotional speech data is recorded by 24 professional actors in a North American accent in eight emotions: calm, happy, sad, angry, fearful, disgusted, surprised, and neutral. Each actor produces emotions at two levels of intensity, normal or strong. The data has been validated by a group of 247 untrained research participants with reports of high emotional validity and intrarater reliability.

The Toronto Emotional Speech Set (TESS) is a data set of 200 target words spoken by two professional female actresses aged 26 and 64. The seven target emotions consist of happiness, sadness, anger, fear, disgust, pleasant surprise, and neutral, for a total of 2800 data points. Note that the TESS data is one word per data point while the RAVDESS data is about 6 words.

Table 1. Distribution of speech recordings used in the paper

Data Set	HP	SD	AG	FR	DG	SP	N	C
RAVDESS	192	192	192	192	192	192	96	192
TESS	400	400	400	400	400	400	400	0

HP = happy, SD = sad, AG = anger, FR = fear, DG = disgust, SP = surprise, N = neutral, C = calm

2.2 Feature Extraction Method

Feature extraction was done with the emobase feature set in the openSMILE package. The feature set contains 989 acoustic features including average pitch, formants, and MFCC information. Individual analysis was performed on each audio clip with a frame time of .125 seconds. The data were then scaled with a StandardScaler, which transforms the data by removing the mean and scaling to unit variance, or a range of -1 to 1. All 989 features were used as numerical input into classifiers.

After pre-processing, several Low Level Descriptors (LLD) are calculated: intensity, loudness, twelve Mel Frequency Cepstral Coefficients (MFCC), pitch, probability of voicing, F_0 envelope, eight Line Spectral Frequencies (LSF), and the zero-crossing rate.

Formants, distinct frequencies that define vowels, can also be used to detect screams because they shift upwards in screams. Each vowel sound has three formants known as F1, F2, and F3 which show up as dark bands in spectrograms of vowels. Screaming tends to increase F1 values because the tongue is pushed back and the jaw is positioned lower.

The Mel-Frequency Cepstrum Coefficient (MFCC) is also commonly used to measure loudness. The calculation splits the signal into short frames, which represent an envelope of speech, so that differences/similarities can be detected between frames. Essentially, these short sound bytes can be used to recognize words. The coefficient is calculated by taking the FFT of each frame, mapping the powers onto the mel scale (which is essentially a measure of energy higher than Hz), and performing a cosine transform on the logs of the energies. “Mapping powers onto the Mel scale” is essentially splitting the signal into N energy bins on the Mel scale, with each bin containing the sum of the FFT energies in that range. This tells us how much of the signal is in each energy range. Vocal effort classification with MFCCs has been found to be around 80%, which may play a role in classifying emotion.

2.3 Classifier

All machine learning classifiers in this paper are implemented using Keras via Tensorflow, written in Python 3.

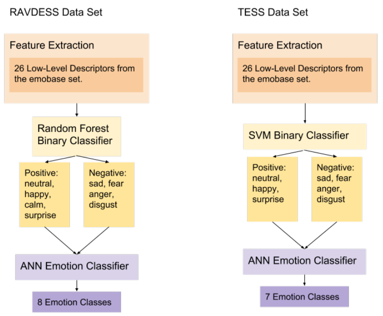

RAVDESS Data Set
Feature Extraction
26 Low-Level Descriptors from the emobase set.
Random Forest Binary Classifier
Positive: neutral, happy, calm, surprise
Negative: sad, fear anger, disgust
ANN Emotion Classifier
8 Emotion Classes
TESS Data Set
Feature Extraction
26 Low-Level Descriptors from the emobase set.
SVM Binary Classifier
Positive: neutral, happy, surprise
Negative: sad, fear anger, disgust
ANN Emotion Classifier
7 Emotion Classes

Figure 1. Diagram of the overall tiered model structure

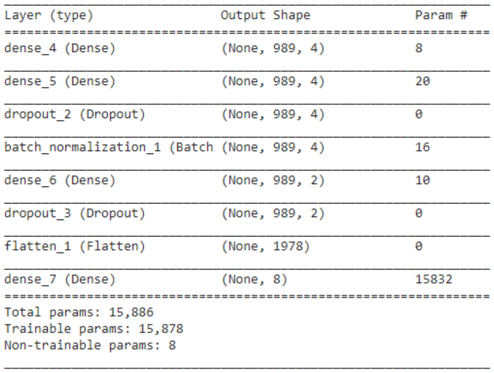
2.3.1 Binary Classifiers

To begin, the 989 features are taken as input into a binary model that identifies whether the sound clip reflects a positive or negative emotion. Happiness, neutral, calm, surprise, are all considered ”positive” emotions, while sadness, anger, fear, and disgust are considered ”negative” emotions.

Since Support Vector Machines (SVMs) are generally better suited to binary classification methods compared to Random Forests, so I tested it for this task and verified its efficacy. A GridSearch determined the most optimal parameters: a penalty parameter C of 1, a radial basis function kernel with a kernel coefficient of .001. A Random Forest was also used for binary classification after testing the accuracies produced by both methods for each dataset. Ultimately, I used an SVM for the TESS data while a Random Forest was more effective for the RAVDESS data. A GridSearch was again utilized to determine the most optimal parameters for the random forest. From the GridSearch, the maximum number of features used was capped at 50, the minimum number of samples in each leaf was 3, and the minimum number of samples to split was 6.

2.3.2 Feed Forward Neural Net

An Artificial Neural Network (ANN) is used to classify the specific emotion, using the above predicted positive/negative label as an input for the model. The neural net was trained for 60 epochs, and it utilizes stochastic gradient descent at a learning rate of .001 and categorical cross entropy loss. The softmax activation function is used for the final multi-class prediction layer. This ANN architecture follows the general rules of using no more than three hidden layers to avoid overfitting, as well as dropout and batch normalization layers.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 989, 4)            8
_________________________________________________________________
dense_5 (Dense)              (None, 989, 4)            20
_________________________________________________________________
dropout_2 (Dropout)          (None, 989, 4)            0
_________________________________________________________________
batch_normalization_1 (Batch (None, 989, 4)            16
_________________________________________________________________
dense_6 (Dense)              (None, 989, 2)            10
_________________________________________________________________
dropout_3 (Dropout)          (None, 989, 2)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 1978)              0
_________________________________________________________________
dense_7 (Dense)              (None, 8)                 15832
=================================================================
Total params: 15,886
Trainable params: 15,878
Non-trainable params: 8
_________________________________________________________________
```

Figure 2. Artificial neural net network architecture for TESS dataset

3. Results and Discussion

Table 2. Binary Classifier Results

Data Set	Positive	Negative
RAVDESS	.934	.923
TESS	1.0	.966

These results are on par with or better than many other models in current literature. I attribute the success to the combination of the power of generalizing and grouping emotions in a binary classification and the detailed learning done by the neural net. A similar 2011 Stanford paper that uses a similar technique reports 13% error on its binary classification task, and a 2018 brain-based speech emotion recognition model reports 90.28% accuracy. [9, 16]

Table 3. Multiclass Classifier Results

Data Set	Neutral	Calm	Happy	Sad
RAVDESS	.65	.93	.86	.81
TESS	1.0	–	.99	1.0
	Fear	Anger	Surprise	Disgust
RAVDESS	.83	.89	.71	.82
TESS	1.0	.98	.98	.98

One of the major shortcomings of the model is that it may only be suited to one word or one sentence audio samples for the TESS and RAVDESS models, respectively. Currently, when the model is trained on one of these datasets, it performs well on testing data of the same format but not of a different length. This may be because audio features associated with emotion -pitch, MFCCs, etc- are expressed differently depending on the length of speech considered. To address this, I am currently working on a continuation of this project that applies transfer learning to train the models on the other dataset. Other literature in the field often reports findings on only one dataset or two separate datasets, so this is an ongoing research question. A successful transfer learning model will likely be better suited for real-time audio data, which may vary in length. Alternatively, real time data can always be clipped to a certain length to suit the model, so the model can still be used commercially.

Various steps were taken to reduce overfitting the data to the test datasets. First, the complexity of the model has been reduced to contain at most 4 nodes per dense layer as well as several dropout layers of 25% and 20% each.

4. Conclusions and Open Questions

In this paper, I present a novel network structure for speech emotion recognition. The model consists of a two-part classifier, one part that determines positive and negative emotions followed by a specific emotion classifier. The accuracies are .8626 and .9760 on the RAVDESS and TESS data sets, respectively, which is notable for this task. Future research can focus on creating models that work for multiple distinct data sets to account for various types of audio data, since the current model only works on one data set at a time. Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

5. Acknowledgments

Thank you to Mr. White, my Syslab director.

References

[1] Lauren Fisher. Us programmatic ad spending forecast 2018, Apr 2018.

[2] Roberto Pieraccini. *The voice in the machine: building computers that understand speech*. MIT Press, 2012.

[3] Grant Fairbanks and Wilbert Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs*, 6(1):87–104, 1939.

[4] Grant Fairbanks and LeMar W. Hoaglin. An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monographs*, 8(1):85–90, 1941.

[5] Adam Conner-Simons and Rachel Gordon. Wearable ai system can detect a conversation's tone, Feb 2017.

[6] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. *International Conference on Spoken Language Processing, ICSLP, Proceedings*, 3, 12 1996.

[7] Assel Davletcharova, Sherin Sugathan, Bibia Abraham, and Alex Pappachen James. Detection and analysis of emotion from speech signals. *Procedia Computer Science*, 58:91 – 96, 2015. Second International Symposium on Computer Vision and the Internet (VisionNet'15).

[8] Shah Hewlett. Emotion detection from speech. 2007.

[9] Alex Mordkovich, Kelly Veit, and Daniel Zilber. Detecting emotion in speech. 2011.

[10] David Griol, José Manuel Molina, and Zoraida Callejas. Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances. *Neurocomputing*, 326-327:132 – 140, 2019.

[11] Turgut Özseven. A novel feature selection method for speech emotion recognition. *Applied Acoustics*, 146:320 – 326, 2019.

[12] Lang He and Cui Cao. Automated depression analysis using convolutional neural networks from speech. *Journal of Biomedical Informatics*, 83:103 – 111, 2018.

[13] Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60 – 68, 2017. Advances in Cognitive Engineering Using Neural Networks.

[14] Yaxiong Ma, Yixue Hao, Min Chen, Jincai Chen, Ping Lu, and Andrej Košir. Audio-visual emotion fusion (avef): A deep efficient weighted approach. *Information Fusion*, 46:184 – 192, 2019.

[15] Shaoling Jing, Xia Mao, and Lijiang Chen. Prominence features: Effective emotional features for speech emotion recognition. *Digital Signal Processing*, 72:216 – 231, 2018.

[16] Zhen-Tao Liu, Qiao Xie, Min Wu, Wei-Hua Cao, Ying Mei, and Jun-Wei Mao. Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing*, 309:145 – 156, 2018.

[17] Huihui He and Rui Xia. Joint binary neural network for multi-label learning with applications to emotion classification. *CoRR*, abs/1802.00891, 2018.

[18] Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35, 05 2018.