SPEECH EMOTION AND DRUNKENNESS DETECTION USING A CONVOLUTIONAL NEURAL NETWORK

Joshua Miller jmill97@u.rochester.edu

Jillian Donahue jdonahu2@u.rochester.edu

Benjamin Schmitz bschmitz@u.rochester.edu

University of Rochester

ABSTRACT

One problem with Artificial Intelligence (AI) is that it lacks emotional or situational knowledge about the human with which it interacts. This paper attempts to propose a solution to this problem by detecting emotion or drunkenness through speech input. Using convolutional neural networks, models for four states were created: happy, sad, angry, and intoxicated. Our network aims to classify these four states with accuracy above 80% by building upon previous research in emotion detection. The ultimate goal of the project is to compare these models with new speech input from a user in real-time that will yield an estimation of what emotion the user is feeling or if they are intoxicated. This method could be implemented in personal assistant systems such as Alexa to give better and more appropriate state-based interactions between AI and humans.

1. INTRODUCTION

Speech emotion detection is anything but a simple problem. The individualistic nature of human emotional expression means that in order for a model to be accurate over a large set of people, it needs to lie in a sweet spot between being precise enough to classify accurately and broad enough to avoid speaker influenced errors.

Despite the difficulty of this problem, the potential usefulness of a good speech emotion model makes the problem worth tackling. This is especially true when we include the speech state for drunkenness into our model.

The primary uses we envision for our model are in AI assistants like Alexa and Siri. If these assistants were able to detect a user's emotional state, they could adapt their behavior to improve the user experience. For example, if an AI assistant detected anger in a user's voice, they could prompt the user to check if the assistant had done something wrong or misinterpreted an input. This would enhance the assistant's ability to learn from its mistakes.

Further, if an assistant detected a drunk user, it could dissuade the user from sending texts or making large online purchases in an impaired state. If the model was very accurate, it could be implemented in voice activated car audio systems, dissuading or preventing the user from driving if drunkenness was detected.

2. OBJECTIVES

The goal of the project was to create a model able to classify the following states using input voice data: happy, sad, angry, and intoxicated. There is a great deal of prior research on emotion detection from speech but detecting drunkenness is a much less explored topic, partially due to the limited availability of drunken speech databases [1]. The first necessary step for creating a model for these states was to find training data for these states. We used a combination of databases for emotion detection and self-sourced intoxicated speech from personal recordings and video clips of intoxicated people. After accumulating this data the next step was to design the network. To the best of our knowledge, there were no other research papers which used neural networks for drunken speech detection, so we wanted to apply the techniques used in speech emotion detection for this additional classification challenge. We set the initial goal of hoping to achieve 80% accuracy for each state when testing new data with our trained model, inspired by a paper using a similar network and spectrogram processing approach for emotion detection which achieved comparable if not greater accuracy [2].

3. BACKGROUND

In this section we hope to elucidate the motivations for our specific technical approach of our project. The first of these is the reasoning for choosing a neural network to achieve our goals. The biggest draw of neural nets for audio analysis is their ability to identify features on their own without explicit guidance from programmers. This is extremely useful for problems like speech emotion, where a variety of features including loudness, envelope, pitch, and formants all play some part in what make speech sound like one emotion or another. Feeding spectrograms to a neural net and letting it figure out what to look for is far easier than manually identifying what features to analyze.

This is all especially true with drunk speech detection, where research is considerably more sparse than with emotion detection. This lack of information makes the use of neural nets the most efficient way of getting good results with drunk speech.

One of the few studies that we found for drunken speech detection utilized RMS rhythmicity and formant features to attempt to classify drunk and sober speech [11]. They did not publish their rate of success for classification, but it was clear that even with their best metrics results were lackluster. This drove us towards raw spectrogram data as our input instead of specific features to avoid spending excessive time searching for metrics that would aid most in classification.

In our pre-processing, discussed more in detail in section 4.2, we chose to use a wide-band spectrogram over a narrow-band spectrogram. Wide-band spectrograms have a lower time resolution, which allows individual glottal pulses to be shown rather than the specific harmonics of the voice [2]. The glottis is responsible for the pitch in speech, and occurs when air passes through the vocal folds. When a short burst of air emerges, it is called a glottal pulse. The pitch of speech changes continuously and is set by the frequency of these glottal pulses [6].

Another choice we made for pre-processing was to augment the speech data and add white noise at 15 dB signal to noise ratio (SNR). The reason for this is to avoid overfitting with a small dataset. Overfitting occurs when the neural network models the training speech data too well, which negatively affects performance of the model when testing new data. Since neural networks are very sensitive to small changes in the input, adding noise increases the “robustness” of the model. It allows for the output to be unaffected by any disruptions in the input. However, in practice this approach did not increase classification accuracy so we chose to omit this augmentation.

Lastly, we choose to use speech input data because we assume in a human-computer interaction scenario, speech is the best indicator of drunkenness and the easiest way for a computer to detect drunkenness. There have been multiple studies that analyze driver fatigue, but not specifically drunkenness. Some studies look into physical features of drunkenness, such as eye movement and head positions [7]. Another study uses thermal infrared imagery to detect when a person is intoxicated based on how “flushed” they are [8]. One study uses a mobile phone to detect typical drunk driving patterns [9]. There are very limited studies that analyze drunk speech. This was a main motivation we had for this project, as we believe this can be a comparable technique in detecting intoxication levels and used for multiple applications.

4. IMPLEMENTATION

4.1. Data Sourcing

The data used for training emotional states (happy, sad angry, neutral) was taken from The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [10]. This database contained mono audio recordings (about three to four seconds long) of 24 voice actors - 12 were male and 12 were female. The recordings contained two kinds of emotional intensities (normal and strong), two statements (“kids are talking by the door” and “dogs are sitting by the door”) and with two repetitions. This gave us a total of 672 audio files sourced from this database.

Given the relatively small and inaccessible corpus of intoxicated speech, we decided to self-source drunken speech samples. For these samples the subjects had at least five to six drinks, the typical amount when speech begins to slur. To expand our collection, we also pulled audio drunken speech recordings from interviews or television shows. We had initially come across a database of drunk speech, called the Alcohol Language Corpus (ALC). It is the first “publicly available audio library of drunk (and sober) speech” [5]. However, for the price of $1,200 plus shipping, it was too costly for us to use for this project.

Additionally, this corpus contains exclusively German speech samples while our network will be entirely trained with English. Most studies separate their models by language if they are using multiple databases [2]. In a paper by Rajoo and Aun called “Influences of Languages in Speech Emotion Recognition: A Comparative Study Using Malay, English and Mandarin languages” comparisons are made between different languages using native speakers and those with the respective language as a second language. Their results show that there are “language specific differences in emotion recognition in which English shows higher recognition rate,” and that native speakers have higher accuracy rates [3]. Since English had the highest accuracy and their results showed slight differences across languages, we decided to stick to just the English database. After sourcing as much drunk data as possible, we used 96 samples per emotion to match the size of the smallest dataset for one class - the neutral emotion recordings.

4.2. Audio Processing

A number of steps of pre-processing were performed on our dataset in order to optimize it for the training of our network. Firstly, the audio files were downsampled from 44.1 kHz to 16 kHz in an effort to reduce training time by limiting spectral data to the frequencies most relevant for speech features. We then initially augmented each audio file with 20 times its original length of white noise at 15 dB SNR to aid in avoiding overfitting with a smaller dataset [2]. After experimenting with training our network, we found that no noise augmentation was actually better for our accuracy results. This is elaborated on in section 5.

Next, we took wide band spectrograms with a five millisecond (80 samples at 16 kHz) window, 70 samples of overlap, and a DFT size of 512. We then removed all frequency bands below 0 Hz and above 4 kHz to further optimize the data for quick training, leaving only information essential for speech. This left us with a spectrogram for each file, which we resized to 129 by 129 using bicubic image resizing. This allowed us to train with uniform data. This processing was all done after reading each .wav file into Python using SciPy.

4.3. Network Architecture

A combination of convolution, pooling, and fully-connected layers were used in our convolutional neural network to create models for each of the four states we hoped to identify. The CNN used in this study contained two convolutional layers (each followed by a max-pooling layer) and one fully connected layer with 1024 hidden neurons. The kernel size of the first convolution layer is 10 by 10 having 8 kernels and the second convolutional layer is size 5 by 5 having 16 kernels. The pooling layer is set to a kernel size of 2 by 2. Figure 1 is a depiction of this neural network.

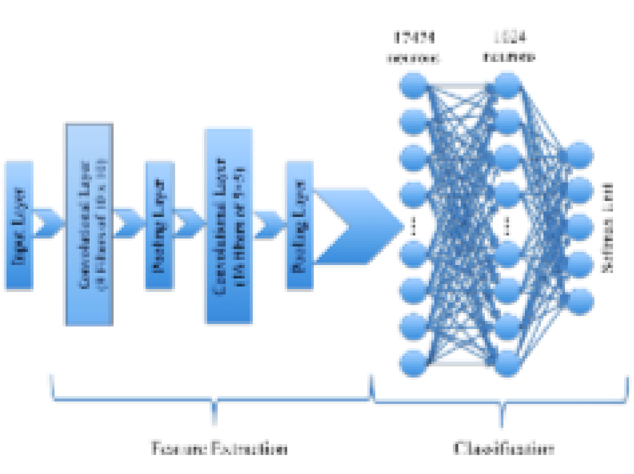


Figure 1: Baseline architecture of convolutional neural network [2]

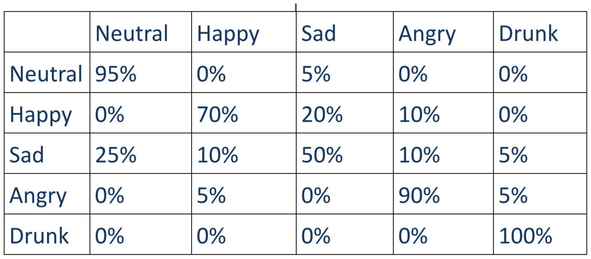
With four classes of emotions and a neutral class, a five-way softmax unit was used to estimate the probability distribution of the classes. To introduce non-linearity to the model, rectified linear units were used as activation functions in the fully connected and convolutional layers. To measure the loss, cross-entropy was used, and to minimize the loss function over the mini batches (between 16 and 256) of the training data the Adam optimizer algorithm was used. The number of epochs was varied as our experiment progressed, but the final value of training iterations used was 100.

Lastly, the dropout algorithm was utilized to reduce the possibility of overfitting by randomly omitting neurons in the hidden layers [2]. With these specifications we expected to obtain the best results and highest accuracy of our model. More in-depth details of each network component can be found in [2].

5. RESULTS

Initially we began training our network with the same parameters as Shahsavarani [2] and listed above, but found that his methods did not translate well for our network. For example, a batch size of 512 was much too large for our database so this had to be fine tuned. Other parameters we adjusted were the learning rate of the Adam optimizer, the dropout probability, the number of epochs the network was run for, and the testing/training split. Additionally, augmenting our dataset with noise did not in any case improve accuracy. Figure 2 shows our best results, a test/train split of about 20%/80%, a class-averaged accuracy of 81%, achieved with a batch size of 16, learning rate of

1e-4, dropout probability of 0.8, no data augmentation, and 100 epochs.



	Neutral	Happy	Sad	Angry	Drunk
Neutral	95%	0%	5%	0%	0%
Happy	0%	70%	20%	10%	0%
Sad	25%	10%	50%	10%	5%
Angry	0%	5%	0%	90%	5%
Drunk	0%	0%	0%	0%	100%

Figure 2: Confusion matrix of CNN with highest accuracy run for 100 epochs.

Using a larger database with emotions happy, sad, angry and drunk, we expect similar results. It is possible that the model is overfitting for drunk data, but this is difficult to test without new test data and cross validation. Another possible explanation for the drunk data being overfit is the varied speech content of the drunk samples. Whereas all other classes have controlled sentences for training, e.g. “the dog is sitting by the door,” the drunk data contains random sentences. Since we acquired drunk data from multiple sources, recording techniques varied greatly and our network might simply be just recognizing those differences. The model could possibly be classifying based on the voice actors and actresses (speaker-dependence). To fix this problem, we would need to test on separate data and introduce enough dissimilarity for the proper classification to further validate our results.

It is also possible our network is not fully optimized and that classification accuracy for non-drunk classes could be increased via more fine tuning of hyperparameters such as batch size and learning rate. Lastly, due to time constraints we were not able to see the effect of training for over 100 epochs. It is possible the results we obtained could be bettered by training for more epochs.

6. CONCLUSIONS

6.1. Achievements

Given a fairly limited database and a restricted time frame for training, we have a network which still achieves high accuracy in detecting emotions and drunkenness. Using intelligent pre-processing and network architecture we successfully captured the features of these states. After experimenting with parameters of our network, we were able to make some improvements on the methods we used to increase our accuracy over time.

6.2. Future Work

For future presentation of our research, we hope to have users input speech in real-time in a demonstration of how our model works. This input would be fed through our pre-processing steps and tested in our network to detect the specific emotion of the user.

In our initial research of speech emotion, we came across multiple databases of speech audio. Some of these databases contained different languages. An initial question we had was if we could use these databases in combination with the English speech database to train and test our model. Would we get similar accuracy results? This is a question we aim to answer with future work. Along the same lines, it would be interesting to compare more similar languages such as Spanish and Portuguese, or English and German and to discover the influence of language using our neural network.

One promising method we hope to implement in our future work is curriculum learning. Curriculum learning is similar to how humans learn. In the neural network, the model would be trained with audio that has clear emotional content and then gradually trained with speech samples that have ambiguous emotional content, increasing the difficulty. This could be beneficial because curriculum learning generally gives better accuracy results. In a paper by Lotfian and Busso titled “Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels,” a curriculum was designed for emotion recognition based on multiple evaluators, using disagreements between individuals as a measure of difficulty [4]. Using these human judgments led to significant improvements in emotion recognition over systems trained without curriculum learning, assuming that recognizing these difficult emotions for humans is just as difficult for computers. This is a method that we could put into action in the future to see if we could achieve better accuracy results with our neural network.

Another improvement that could be made is to add more convolutional layers to our network. Adding more layers will allow the dimension of the fully-connected layer (depicted in Figure 1) to be reduced. Initially, this was not something we considered due to the increase in training time it would add as well as the success of others studies with the same network architecture. Lastly, k-fold cross validation could be used to make sure the model is not overfitting. K-fold cross validation is a technique which randomly separates data into k groups of the same size and uses k-1 of these groups for training while the remaining group is used for testing. This procedure is repeated k times until each group has been used as test set. Although we have achieved high classification accuracy for one test set it is possible given a different test set the network would not perform quite as well.

In conclusion, we hope to keep updating our findings with our neural network to expand the applications of machine learning.

7. REFERENCES

[1] Bone, Daniel et al. "Intoxicated Speech Detection: A Fusion Framework with Speaker-Normalized Hierarchical Functionals and GMM Supervectors" *Computer Speech & Language* vol. 28, 2 (2012): 10.1016/j.csl.2012.09.004.

[2] Shahsavarani, Somayeh. "Speech Emotion Recognition using Convolutional Neural Networks" (2018). Computer Science and Engineering: *Theses, Dissertations, and Student Research*. 150. https://digitalcommons.unl.edu/computerscidiss/150.

[3] R. Rajoo & C.C. Aun. "Influences of languages in speech emotion recognition: A comparative study using Malay English and Mandarin languages" *Computer Applications & Industrial Electronics (ISCAIE) 2016 IEEE Symposium on. IEEE*, pp. 35-39, 2016.

[4] Buso, Carlos & Lotfian Rezo. "Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels" (25 May 2018). arXiv:1805.10339

[5] Braga, Matthew. "Inside The First Audio Library of Alcohol-Addled Speech (Which Just Might Help Stop Drunk Driving)." *Fast Company*, 25 November 2014, https://www.fastcompany.com/3038889/inside-the-first-audio-library-of-alcohol-addled-speech-which-just-m, accessed 10 November 2018.

[6] Puckette, Miller. "Acoustics for Musicians and Artists: The Voice." Music 170, 24 November 2014, University of California San Diego. Online Course Notes. http://msp.ucsd.edu/syllabi/170.13f/course-notes/node5.html

[7] Rezaei M. & Klette R. (2011) "3D Cascade of Classifiers for Open and Closed Eye Detection in Driver Distraction Monitoring." In: Real P., Diaz-Pernil D., Molina-Abril H., Berciano A., Kropatsch W. (eds) Computer Analysis of Images and Patterns. CAIP 2011. Lecture Notes in Computer Science, vol 6855. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23678-5_19

[8] Koukiou, Georgia & Anastassopoulos, Vassilis. "Neural networks for identifying drunk persons using thermal infrared imagery" *Forensic Science International*, vol 252, pp 69-76, 2015. https://doi.org/10.1016/j.forsciint.2015.04.022.

[9] J. Dai, J. Teng, X. Bai, Z. Shen and D. Xuan, "Mobile phone based drunk driving detection" *4th International Conference on Pervasive Computing Technologies for Healthcare*, Munich, 2010, pp. 1-8. doi: 10.4108/ICST.PERVASIVEHEALTH2010.8901

[10] Livingstone, Steven R., & Russo, Frank A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Version 1.0.0) [Data set]. PLoS ONE. Zenodo. http://doi.org/10.5281/zenodo.1188976

[11] Florian Schiel, Christian Heinrich, Veronika Neumeyer (2010). Rhythm and Formant Features for Automatic Alcohol Detection . Bavarian Archive for Speech Signals.