SPEECH EMOTION AND DRUNKENNESS DETECTION USING A CONVOLUTIONAL NEURAL NETWORK

JOSHUA MILLER, JILLIAN DONAHUE, BEN SCHMITZ
University of Rochester, Department of Audio and Music Engineering

ABSTRACT

One problem with Artificial Intelligence (AI) is that it lacks emotional or situational knowledge about the human with which it interacts. This project 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.

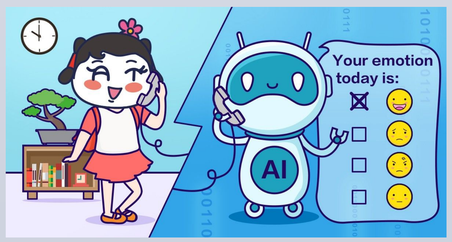
OBJECTIVES

- Create a model able to classify the following states using input voice data: happy, sad, angry, and intoxicated
- Source enough good speech data to thoroughly train a model for the desired 80% accuracy figure

BACKGROUND

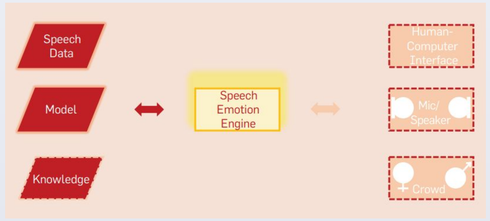
Speech Emotion Detection Problem

- Speech emotion detection has been a challenging and complex problem to tackle
- An accurate representation of the classified emotions needs to be created
- Complex networks and large amounts of data are typically required for an accurate model



Neural Network Methods

- Methods for speech emotion detection based on neural network technology are in vogue currently
- Neural nets have the ability to learn complex patterns like those that indicate emotion in speech without much guidance from programmers



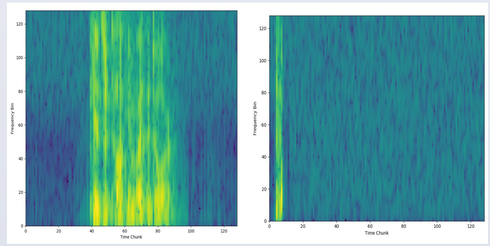
IMPLEMENTATION

Dataset

- Happy, sad angry, and neutral data was taken from The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [10]
- Contained three second audio recordings of 24 voice actors, 12 were male and 12 were female
- Recordings contained two emotional intensities (normal and strong) and two statements, each with two repetitions for a total size of 672 files
- Drunk speech data is considerably harder to come by, requiring us to self source data by recording it ourselves and fetching it from Youtube
- For our self recorded samples the subjects had at least five to six drinks, the typical amount when speech begins to slur
- A database of drunk speech exists called the Alcohol Language Corpus (ALC). However, it's cost and the fact that it contained only German speech inhibited our use of it

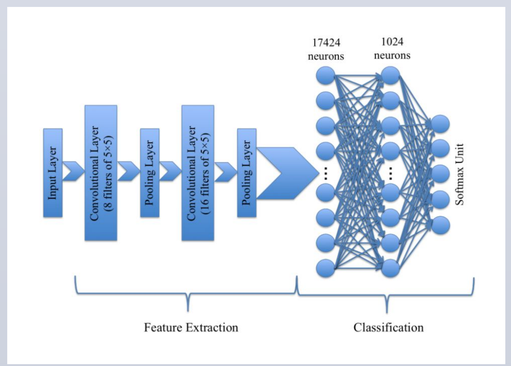
Preprocessing

- Audio files were downsampled from 44.1 kHz to 16 kHz to limit spectral data to the frequencies most relevant for speech features
- Augmented each audio file with several times its original length of white noise at 15 dB SNR to avoid overfitting [2]
- Took wide band spectrograms with an 80 sample window, 70 sample overlap, and a DFT size of 512, removed all frequency bands below 0 Hz and above 4 kHz, then rescaled using bicubic image resizing to optimize the data for quick training



CNN

- Used a combination of convolutional, pooling, and fully connected layers
- Softmax unit last to give the predicted distribution of classes
- Hyperparameters to adjust: learning rate, epochs, dropout probability, batch size
- Monitor progress to avoid overfitting



RESULTS

Confusion Matrix:
Batch size = 16, 100 Epochs, No augmentation, Dropout probability = 0.8, Learning rate = 1e-4)

	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%

- Data augmentation with noise did not produce greater accuracy
- Drunk data likely overfitting
- Unable to truly assess without more varied test data

CONCLUSIONS

- Network parameters not yet fully optimized
- Shortage of drunk data might be contributing to low classification rate
- Adding more convolutional layers may also give better results

FUTURE WORK

- Allow user input for classification in real time
- Discover the influence of language by using multiple databases
- Implementing curriculum learning for higher accuracy

SOURCES

[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