Voice Based Intoxication Classification on Mel Filters using Deep Convolutional Networks

Prof. Emmanuel Agu, Shreesha Narasimha Murthy, Pratik Jawahar, Mitesh Agrawal, Arjun Rao

Worcester Polytechnic Institute,
100 Institute Road,
Worcester, MA, USA

emmanuel@wpi.edu,snarasimhamurthy@wpi.edu, pjawahar@wpi.edu, msagrawal@wpi.edu, arao@wpi.edu

Abstract

One of the largest risk factors that lead to accidents is driving under the influence of alcohol. In the United States, 29 people die every day involving motor-vehicle crash due to alcohol intoxication. This amounts to more that $44 billion of annual cost for alcohol-related crashes [1]. Informal intoxication tests like breathalyzers are unreliable as indicators for evaluating safe motor vehicle driving, heavy equipment operation, machine tool use, etc.. Passive detection of intoxication from voice has important applications to high-risk situations, such as driving and steering, but there is less research into this speaker state than into other speaker states. The majority of previous work uses classic machine learning techniques with Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) and learns from hand-extracted features. This paper explores a Convolutional Neural Network (CNN) based architecture for intoxication detection from voice using log Mel spectrograms as input. We experiment with the time and frequency masking architectures for feature extraction and use different combinations of dense layers and pooling techniques for classification. We also experiment with using LSTM models, auto encoder models as well as variational auto encoder model. These CNN-based architectures require little feature engineering compared to previous methods and are simple to train on different domains. On the Alcohol Language Corpus (ALC), the model achieves an unweighted average recall (UAR) of 66.28%. Our work lays important groundwork for future research into CNN-based architectures for intoxication detection. With future experimentation and adaption to the ALC dataset, the CNN architecture can be used to classify intoxicated speech with a higher UAR.

Index Terms: Deep Learning, Audio classification, Intoxication detection, Alcohol Language Corpus, CNN, VAE, AE, Disentangled VAE

1. Introduction

Alcohol has known effects on human behavior and can be dangerous when consumed in large amounts. Blood alcohol concentrations between 0.05% and 0.08% are known to impair judgment, while higher concentrations may cause nausea, slurred speech, and loss of coordination. Blood alcohol concentrations above 0.15% can leave a person unconscious and may even result in death [1]. Due to alcohol's effects on human judgment and coordination, driving while intoxicated poses a great safety risk to other drivers and passengers on the road. In fact, nearly 10.5 thousand people in the United States died in alcohol-impaired driving accidents in 2016, accounting for 28% of all driving-related deaths in the United States [2]. Thus, it is important to research and develop techniques for passive detection of intoxication to alert drivers of their state before they start driving.

In addition to affecting coordination, alcohol has known effects on a person's speech. Alcohol causes a person's speech to become slower, and the number of pauses, stutters, and speech errors also increases. The pitch and fundamental frequency of a person's voice may also increase, but this is not consistent across genders [3].

One can use voice analytics to study the changes in speech patterns of sober and intoxicated people to help detect intoxication from voice. Voice analytics is a branch of audio processing research to analyze spoken conversation and audio patterns using machine learning or deep learning models to extract and analyze information, including speech content and speaker state, by analyzing patterns in human speech. Analyzing audio data presents several challenges and these include voice samples being affected by gender, age, emotions, room acoustics and proximity to the user affect the voice sample. [4]

Previous work into intoxication detection from audio samples comes primarily from the 2011 INTERSPEECH Challenge, which included an Intoxication Detection subtask. For the subtask, participants had to identify speech in the Alcohol Language Corpus (ALC) as either sober or intoxicated. The results of the INTERSPEECH Challenge of 2011 showed that machine learning models can detect alcoholized speech from hand-engineered audio features with a highest Unweighted Average Recall (UAR) of 70.5% [5]. However, there has been little research into using deep learning methods for detecting alcohol level from voice.

In this paper, we use a CNN with pooling to classify speech segments in the Alcohol Language Corpus (ALC) using log Mel spectrograms as input. Convolution Neural Networks (CNNs) have not been yet explored for feature extraction in the audio classification pipeline.It is interesting to note that all steps needed to compute filter banks were motivated by the nature of the speech signal and the human perception of such signals[6]. The paper discusses various approaches using CNN and mel scale filter banks to classify voice notes.The proposed CNN-based architecture requires minimal feature engineering compared to previous methods in intoxication detection. Our experiments show detailed analysis of data-set, complexity of problem statement and a competitive result in classification.

The remainder of this paper is organized as follows: We present background information in Section II, and our proposed intoxication detection architecture is explained in Section III. In Section IV, we outline the experiments we ran and present results in Section V. Discussion is presented in Section VI fol-

lowed by Conclusion and Future work in Section VII.

2. BACKGROUND

2.1. INTERSPEECH 2011 Challenge: Intoxication Detection Sub-challenge

Previous work into intoxication detection comes primarily from the 2011 INTERSPEECH Challenge. The Challenge focused on intoxication detection and sleepiness detection, two speaker states that were less researched in the past. Sleepiness and intoxication detection have applications to the security and medical domains, in situations such as driving, steering, and controlling [5]. The Intoxication Detection sub-task of the Challenge was a supervised binary classification task using the Alcohol Language Corpus (ALC) Dataset, which is described in Section IV. Speech recordings were labeled according to the Blood Alcohol Content (BAC) of the speaker, and they had to be classified as either alcoholized for BAC exceeding 0.5 per mill or sober for BAC equal to or below 0.5 per mill. The official audio feature set consisted of 4368 Low-Level Descriptors (LLDs) extracted using openSMILE and known to be useful for intoxication detection. Contestants could also extract additional low-level and hierarchical features for audio classification. Participants were given training and test sets and reported model performance using Unweighted Average Recall (UAR) [5]. Our work uses the ALC dataset used in the Intoxication Detection sub-challenge for supervised binary classification of speech segments as either sober or intoxicated. We use UAR to evaluate our models to maintain uniformity in comparing results obtained from the original challenge.

2.2. Deep Learning in Audio Classification

Deep learning has shown success in a variety of other speech processing tasks as well, including emotion detection, speaker recognition, and audio event detection [7] [8] [9]. These deep models learn high-level features on top of the Low- Level Descriptors (LLDs) that are typically used in machine learning methods, and they outperform standard machinelearning approaches. Berninger et. al [10] lay the foundation of using a deep neural network for the speaker intoxication detection task on the ALC dataset. They use a bi-directional Recurrent Neural Network (Bi-RNN) with 2 Gated Recurrent Unit (GRU) layers and Gaussian dropout for the binary intoxication detection task. The Bi-RNN model has a forward GRU layer and a backward GRU layer to capture dependencies in the speech signal in both the forward and backwards directions while avoiding the vanishing gradient problem. [10] use the CMU Sphinx speech recognition toolkit 40-dimensional filter bank (FBANK) features from speech segments in the ALC dataset. The spectrogram representations of the audio signals are input to the network. The model achieves an accuracy of 71.30%and unweighted average recall of 71.03%, outperforming the winning submission of the 2011 Challenge with minimal feature engineering.

Other deep learning techniques for speech processing use Convolutional Neural Networks (CNNs) in the classification pipeline. Deep CNN models are robust to different audio environments and speaker styles [11]. They require little feature engineering and learn high-level feature representations as they train [7], [12]. Hershey et al. [7] show that CNN architectures such as AlexNet, VGG, Inception, and ResNet, which are typically used for image classification, are also effective for large-scale Audio Event Detection (AED) on the Youtube- 100M dataset.CNNs have also been used for health-related speech processing tasks. Wu et al. [13] use a Convolutional Neural Network (CNN) to detect pathological voice disorders on the Saarbruecken dataset. They compute spectrogram representations of normal and pathological speech in the dataset and input these representations into a CNN pretrained with a Convolutional Deep Belief Network (CDBN). The CNN comprises of 10 convolutional and maxpooling layers followed by a Dense layer for classification. The CNN-based model performs with 71% accuracy on the test set and achieves an F1-score of 72%. Moreover, CNNs can be used to successfully classify audio end-to-end from raw data, requiring no additional feature extraction [12].

Use of Auto Encoders(AE) is known for improving performance of classifiers for detecting anomelies in a better way [14]. Autoencoders have been widely used for obtaining useful latent variables from high-dimensional datasets. Variational Auto Encoders (VAE) has shown improvement in complex classification problems using latent space representation. VAE can also be used as an tool to generate spectrum of raw audio to learn specific parameters that are more relevant to classification. J Hennig used Classifier in a VAE to improve performance of music regeneration [15]. Corollary research work shows more accuracy for emotion classification of emotions as compaired to RNN models [16][17]

Proper Data augmentation methods has shown great improvement in the classification of images.Learned augmentation techniques have explored different sequences of augmentation transformations that have achieved state-of-the-art performance in the image domain [18]. SpecAugment, an augmentation method that operates on the log mel spectrogram of the input audio, rather than the raw audio itself, consists of three kinds of deformations of the log mel spectrogram.The first is time warping, a deformation of the time-series in the time direction. The other two augmentations, inspired by "Cutout", proposed in computer vision [19], are time and frequency masking, where we mask a block of consecutive time steps or mel frequency channels. SpecAugment time and frequency masking has shown great improvement in results of Automatic Speech Recognition network[20]

It is interesting to note that all steps needed to compute filter banks were motivated by the nature of the speech signal and the human perception of such signals[6]. Using these filter banks as Low-Level audio Descriptors (LLDs) to learn and require careful model adaptation and tuning to account for speaker variability. CNNs have performed well in audio classification tasks and are robust to different speaker styles and acoustic environments [21]. Moreover, they are easy to train and perform well on weakly labeled datasets [22]. Given the imbalanced classes in the Alcohol Language Corpus (ALC) and the variability in speaker gender and style in the corpus, we focused our research on CNN-based architectures with time and frequency masking for intoxication detection along with Filter banks used as LLDs.

3. Experiments

In general data is pre-processed by splitting audio files using sliding window techniques with minimum overlap. Audio files are read using LibROSA library functions and applied Mel filter bank.LibROSA is a python package for music and audio analysis[23]. There are four main experiments tried to classify data and they are discussed below.

3.1. Convolutional Neural Network (convnet)

CNN models are robust to different audio environments and speaker styles [11]. The Network uses four layers of CNN.After post-processing the extracted features into equallength, the features are input into 3 Dense layers followed by a pooling mechanism and then 2 dense layers followed by pooling layers to classify the input audio segment as either Sober or Intoxicated. The 3 fully connected-layers compute high-level embeddings from the features output from Mel filters. Pooling reduces the feature map of embeddings while retaining information of an activation of features [24]. Maximum pooling functions like the global max pooling layer in a Convolutional Neural Network (CNN) consists of a single Dense layer with RELU activation. The maximum prediction is used for classification. The model is experemented with and without time and frequency masking techniques where we mask a block of consecutive time steps or mel frequency channels. Loss function used is binary cross entropy.

3.2. LSTM and Conv LSTM Networks

LSTM have shown great results in classifying ACL dataset [5]. The LSTM network comprises of bi-directional Recurrent Neural Net-work (Bi-RNN) with 2 Gated Recurrent Unit (GRU) layers and Gaussian dropout for the binary intoxication detection task. The Bi-RNN model has a forward GRU layer and a backward GRU layer to capture dependencies in the speech signal in both the forward and backwards directions while avoiding the vanishing gradient problem. The Conv LSTM network comprises of combination of previously discussed convnet followed by LSTM layer. The network comprises of two dense CNN layers followed by pooling and a LSTM layer. Binary cross entropy is used as loss function in both the models.

3.3. Convolutional Auto Encoder and Variational Auto encoders Networks

Auto Encoders (AE) is known for improving performance of classifiers for detecting anomelies in a better way[14]. AEs have been widely used for obtaining useful la-tent variables from high-dimensional datasets. In general AE comprise of a Encoder, Latent space and Decoder. In the model the latent space generated after multiple layers of CNN in encoder is passed through Decoder CNN network for reconstruction of the input Mel filtered data as well as a classifier network of CNN to identify class. The MSE loss function is used to calculate reconstruction loss and binary cross entropy for classification loss. For Variational Auto encoder classifying model is trained with evidence lower bound (ELBO) loss and binary cross entropy.

3.4. Disentangled Classification

The output of the LSTM Variational Autoencoder per-formed better than expected since the reconstruction of signals is hard and we achieved good accuracy on the model. Rather than use the output of the VAE directly for classification, we decided to use the representation for classification. We replace the decoder with a classifier since we want to perform classification. According to the paper, we can separate time-variant from time-invariant features and we decided to implement this as EEG data contains a lot of additional noise which are time-variant like muscle movement, eye blinks, etc.. [25] shows that beats perform well when it comes to classification.

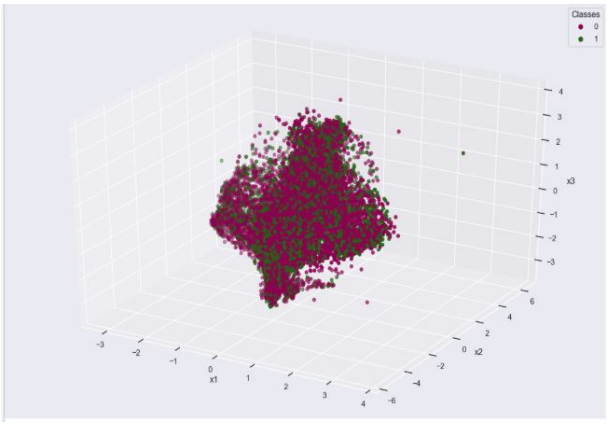

Figure 1: *Tnse train data 3D visualization*

4. Discussion

The results of the experiments implemented are shown below. In general CNN architectures with Mel-Filter input performed better than other architecture. The data imbalance is one of the main issues that affect the network. To solve this we initially use standard undersampling techniques to reduce the majority class samples and make the event rate of the train set to an equal class distribution. There is a marked improvement in network performance. However, this is significantly outperformed by augmenting the minority class using the SpecAugment techniques presented by Google.

We observed that time and frequency masking augmentation presented in SpecAugment help network learn the parameters in best way for the CNN network. The time-warping technique from SpecAugment again, is a very task-oriented and parameter sensitive technique. This means picking the right parameters is of utmost importance, but the lack of a well defined method to pick these parameters makes it difficult to work with for our task, and thereby it produced poor results

It was noted that LSTM networks had poor performance, with one of the main reasons being large sequence sizes that the network couldn't possibly handle. As a result, performance with Bi-LSTMs was equally poor.

Subsequent experiments involved the best CNN model being conditioned with demographic information such as speakers' gender, age, BMI, mood before and after the interview, drinking and smoking habits and the weather conditions and car details in which the recordings were made. Despite a marked improvement expected, none of these techniques could outperform the best CNN model.

Table 1 shows compiled results of various experiments with ALC dataset. The complexity of dataset is very high and along with high-dimensional spectrum too it is difficult to classify intoxicated and sober sample to high accuracy because of no clear distinction. This can be observed in figure 1 and 2 which represent TNSE 3D visualization of log Mel data.

5. Conclusions

Our work explores how CNN architectures can be used in the audio classification pipeline for the intoxication detection task. Our work establishes a solid baseline for further experimentation into CNN techniques for intoxication detection. Although many of the experiments we ran did not train, adding differ-

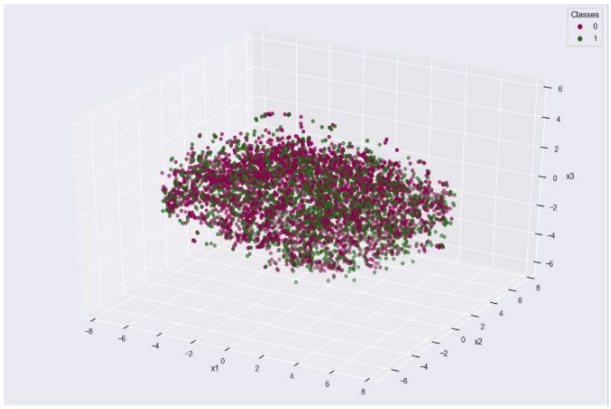

Figure 2: *Tnse test data 3D visualization*

Table 1: *Test UAR for trained networks on ACL dataset*

Network	Demographics	Augmentation method	UAR
Convnet	None	None	61
Convnet	None	Time and frequency masking	66.28
Convnet	None	Time wrapping	62
LSTM	None	None	58.12
Conv- LSTM	None	None	62.27
Autoencoder	None	None	65.53
Variational Autoencoder	None	None	62.78
Convnet	All	Time and frequency masking	65.2
Convnet	Only gender	Time and frequency masking	64.3
Convnet	Only drinking habits	Time and frequency masking	63.1
Convnet	Only BMI	Time and frequency masking	62.6
Convnet	Only smoking habits	Time and frequency masking	60.4
Disentangled VAE	None	None	60.2

ent augmentation like time and frequency masking techniques to the standalone CNN architecture helped with model training and improved UAR by 7%, compared to the worst-performing model in the table. The best result we achieved was using the CNN model, using Time and frequency masking. The model achieved a UAR of 66.28%. This result is 9% higher than the results we were getting from standalone LSTM model with UAR of 58.12%.

.

Other experiments we can explore for future work are outlined below:

- Extract audio features:
  - Spectral Centroid
  - Spectral Rolloff
  - Spectral Bandwidth
  - Zero Crossing Rate
  - MFCC (mean over time)
  - Chrome Features
- With the above extracted features, feed it along with Mel filters (our existing state of art model)
- Focus on low level features
- Work on fine-grained classification

6. References

[1] SAHealth.gov, "Blood alcohol concentration (bac) and the effects of alcohol," pp. https://www.sahealth.sa.gov.au/wps/wcm/connect/public+content/sa+health+internet/health+topics/health+conditions+prevention+and+treatment/alcohol/blood+alcohol+concentration+bac+general+information.

[2] CDC.gov, "Impaired driving: Get the facts," pp. https://www.cdc.gov/motorvehiclesafety/impaired driving/impaired–drv factsheet.html.

[3] E. Tisljár-Szabó, R. Rossu, V. Varga, and C. Pléh, "The effect of alcohol on speech production," *Journal of psycholinguistic research*, vol. 43, 11 2013.

[4] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, pp. 1733–1746, 10 2015.

[5] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states—a review on intoxication, sleepiness and the first challenge," *Computer Speech Language*, vol. 28, p. 346–374, 03 2014.

[6] H. Fayek, "Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between," 04 2016.

[7] S. Hershey, S. Chaudhuri, D. Ellis, J. Gemmeke, A. Jansen, R. Moore, M. Plakal, D. Platt, R. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," 03 2017, pp. 131–135.

[8] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," 09 2015.

[9] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 03 2017.

[10] K. Berninger, J. Hoppe, and B. Milde, "Classification of speaker intoxication using a bidirectional recurrent neural network," vol. 9924, 09 2016, pp. 435–442.

[11] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, J. Chen, J. Chen, Z. Chen, M. Chrzanowski, A. Coates, G. Diamos, K. Ding, N. Du, E. Elsen, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," 12 2015.

[12] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," 03 2016.

[13] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," 09 2018, pp. 446–450.

[14] K. Babaei, Z. Chen, and T. Maul, "Data augmentation by autoencoders for unsupervised anomaly detection," 12 2019.

[15] J. Hennig, A. Umakantha, and R. Williamson, "A classifying variational autoencoder with application to polyphonic music generation," 11 2017.

[16] J. Hennig and R. Williamson, "Sequence generation and classification with vaes and rnns," 11 2017.

[17] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion," 12 2017.

[18] E. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. Le, "Autoaugment: Learning augmentation policies from data," 05 2018.

[19] T. DeVries and G. Taylor, "Improved regularization of convolutional neural networks with cutout," 08 2017.

[20] D. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," 09 2019, pp. 2613–2617.

[21] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng, "Deepspeech: Scaling up end-to-end speech recognition," 12 2014.

[22] Q. Kong, Y. Xu, W. Wang, and M. Plumbley, "Audio set classification with attention model: A probabilistic perspective," 04 2018.

[23] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

[24] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," 06 2016.

[25] Y. Li and S. Mandt, "Disentangled sequential autoencoder," 2018.