AUTOMATED DETECTION OF CANNABIS-INTOXICATION FROM SPEECH

Arian Shamei *[1], **Peter R. Sullivan** †[1], **Yadong Liu** ‡[1], **Muhammad Abdul-Mageed** §[1], **and Bryan Gick** ¶[1]
[1]University of British Columbia

Abstract

Machine learning can reliably distinguish a variety of mental and physical states based on acoustic alterations in the speech stream. Recent acoustic research found that cannabis intoxication results in significant differences in several acoustic correlates. Encouraged by these observations, we report models aimed at detecting cannabis intoxication from human speech. Using a small number of speakers (4 male,4 female) we exploit mel spectrograms from sober and intoxicated productions of sustained vowels, to train models under various gender-nuanced conditions (i.e., male-only, female-only, gender-agnostic) using convolutional neural networks (CNNs). In speaker-independent cross-validation, we report encouraging model performance ($avg\ macro\ F1$ - females : 68.6%, males : 67.9%).

Keywords: state detection, cannabis intoxication, automatic detection

Résumé

L'apprentissage automatique peut distinguer de manière fiable une variété d'états mentaux et physiques sur la base des altérations acoustiques du flux vocal. Des recherches acoustiques récentes ont montré que l'intoxication au cannabis entraîne des différences significatives dans plusieurs paramètres acoustiques. Encouragés par ces observations, nous présentons des modèles visant à détecter l'intoxication au cannabis à partir de la parole humaine. En utilisant un petit nombre de locuteurs (4 hommes, 4 femmes), nous exploitons des spectrogrammes de mélanges de productions sobres et intoxiquées de voyelles soutenues, pour entraîner des modèles dans différentes conditions de genre (c'est-à-dire, homme seulement, femme seulement, sans considération de genre) en utilisant des réseaux neuronaux convolutifs. Dans le cadre d'une validation croisée indépendante du locuteur, les performances du modèle sont encourageantes ($avg\ macro\ F1$ - femmes : 68,6%, hommes : 67,9%).

Mots clefs: détection d'état mental, intoxication au cannabis, détection automatique

1 Introduction

With the recent legalization of cannabis in Canada, accurate detection of cannabis intoxication is a necessity in legal and medical contexts. Aside from a blood test, which is undesirable for both ethical and logistical reasons, there is currently no reliable detection method for cannabis intoxication. The goal of this work it to investigate the feasibility of automated methods to detect cannabis intoxication from acoustic speech data. In particular, we present effective neural network models trained to detect intoxication from sustained vowels. Crucially, we obtained these results using a relatively simple and easily available network architecture and minimal data preprocessing. We believe these results highlight the robust effect of cannabis intoxication on phonation and the effectiveness of deep learning for this task.

Acoustic analysis has been shown to reliably distinguish a range of mental and physiological states, including alcohol intoxication [1] Parkinson's disease [2], heart disease [3], MDMA intoxication [4], head and neck cancer patient intelligibility [5] and emotional state [6]. Accurate classification using machine learning methods has been demonstrated for many of these states (Alcohol - see [7], Parkinson's - see [8]). A preliminary acoustic analysis on cannabis-intoxicated speech [9] found significant spectral and phonetic alterations in the speech of eight participants following cannabis intoxication. We use this very same dataset to train models based on 2-layer 2D CNNs (Convolutional Neural Network) to detect cannabis intoxication from sustained vowel articulations. Our models perform binary classification, predicting whether samples from unseen speakers were produced in sober or intoxicated conditions.

∗. arian@alumni.ubc.ca
†. prsull@gmail.com
‡. yadong@alumni.ubc.ca
§. muhammad.mageed@ubc.ca
¶. gick@mail.ubc.ca

2 Related Work

To date, we know of no previous research investigating automated detection of cannabis intoxication from speech data. However, there has been substantial work on detecting related speaker states such as alcohol intoxication from the speech stream. Much of this work was fuelled by the Interspeech 2011 Intoxication Speaker State Challenge [10]. The challenge utilized the Alcohol Language Corpus [11], which contained audio recordings of German speakers performing language tasks at a variety of intoxication levels. A gender balanced set (n=144) was randomly selected and partitioned into speaker-independent training, development and test sets. [10] reported a baseline model employing Support Vector

Machines which obtained an accuracy of 65.9%. The top performing model in the task [1] used Gaussian Mixture Model supervectors alongside normalized hierarchical features to obtain an accuracy of 70.5%. More recently, [12] used a gated recurrent neural network to achieve 75% speaker independent accuracy using FBANK features as input.

More recent work on state detection incorporates deep learning architectures and spectral imaging methods [13]. In such frameworks, audio classification tasks often make use of CNNs as their sensitivity to time-invariant data has proven to extremely valuable [14]. For example, there is a substantial body of literature reporting state-of-the-art classification of a speaker's emotion using log-mel spectrograms and 1D [15], 2D [6], and 3D CNNs [16, 17]. The use of deep learning and spectral imaging methods have recently been applied to alcohol-intoxication with promising results [7, 18]. In sum, successful state detection methodologies have been reported using traditional machine learning or more modern deep learning methods, employing either representative features of the audio or spectral imaging mediums.

3 Data

The dataset consists of eight participants (4 male, 4 female) completing a set of speech tasks in sober and intoxicated conditions using a pretest - retest format. Human research ethics approval was obtained from the University of Victoria HREB. Recordings took place in a sound-attenuated booth at the University of Victoria using a the internal microphone of a Zoom H4N Stereo recorder with a sampling rate of 44.1KHz and a bit rate of 16. Participants were asked to sit comfortably and maintain a consistent distance from the microphone. Participants arrived after abstaining from Cannabis use for at least 24 hours and completed the sober pretest tasks. Participants were then given a thirty minute break to consume their cannabis and perform the identical retest tasks. Participants provided their own cannabis and consumed varying amounts based on their medical needs. All participants consumed their cannabis by either smoking or vaporizing. Participants were asked to self-report their frequency of use and degree of intoxication on a 1-7 scale, however as this metric was highly subjective it was not utilized in the present analysis. It should be noted that accurately identifying the degree of cannabis intoxication and its interaction with a participant's individual tolerance level is difficult without real time monitoring of THC levels in the blood. Habitual users of cannabis have detectable THC levels in the bloodstream even when cannabis has not been recently consumed, which further adds to the complexity of blood-level monitoring [19]. As this study was preliminary in nature, we opted to assess the data in a manner naive to degree of intoxication and tolerance, however it is likely that these factors will influence model performance.

All participants were native speakers of English, between 20-24 years of age with no reported speech or hearing disorders. Participants included medicinal users of cannabis with a wide range of experience, frequency of use, and tolerance levels. Participants were asked to sustain seven vowels embedded within carrier words (e.g. *see*, *sue*, *saw*) for approximately five seconds each although the exact duration was not enforced. Two iterations were collected from each participant. Our decision to include only the sustained vowel task is motivated by our desire to base our models on purely *phonetic* components of speech, while avoiding the effects of *suprasegmental* or *non-phonetic* components such as speech rate, lexeme choice, or loudness. Each audio file contained either the pretest or retest recording of one participant, for a total of 16 audio files, each consisting of approximately 40 seconds of sustained vowels and their onsets following the manual removal of silences.

4 Models & Experiments

4.1 Model settings

We develop models under two main settings based on how we split our data : **(1) Gender specific** : Here we train two distinct models, each using male (n=4) or female (n=4) data exclusively. **(2) Gender agnostic** : Where we use the combined data (n=8) from both the males and females. For each setting (i.e., *male-only*, *female-only*, and *gender-agnostic*), we run *n*-fold cross validation experiments. Similar methodology has recently been applied to train models for MDMA intoxication from a small number of speakers (n=31), and presents a reasonable alternative to independent training, validation, and test sets for smaller datasets.

Our treatment of each speaker's data as an independent fold allows us to obtain both (a) speaker-independent (i.e., across all speakers) and (b) speaker-specific (on each individual speaker) results. In all cases, allowing individual scores for each participant without including their data in the training set. In our *gender-agnostic* version, this meant we trained on 7 speakers at each and tested on the eighth fold in a model which did not account for gender. In our *gender-split* versions, independent models were trained on female speakers and male speakers respectively, allowing an evaluation of whether gender-dependent speaker characteristics such as *F0* might be a factor in classification.

4.2 Model architecture

Model architecture was invariably a 2-layer 2D CNN implemented in the PyTorch [20] Python Deep learning framework. Each layer of the CNN was fed into a ReLU activation function [21] and then max pooled. Dropout [22] was applied to the second CNN layer and the subsequent dense layer. The model culminated in an output node for binary classification of speaker state (sober vs intoxicated). We use the Adadelta optimizer [23] with the default learning rate of 1.0 - note that pytorch does not recommend adjusting the adadelta learning rate. Specific model hyperparameters were selected by random search function as outlined in subsection 4.5.

4.3 Baseline

For our baseline, we train a Support Vector Machine (SVM) with a radial basis function (RBF) kernel on the data using eight-fold cross-validation.

4.4 Data extraction

For each of these settings (agnostic, specific, baseline) we extract a total number of 500 mel spectrograms (250 from each class) from each participant, for a total of 4,000 spectrograms. The duration (in milliseconds) of each sample was determined by random search function, with possible values including 250, 500, and 1000ms. We use Librosa [24], the Python audio analysis package, to produce mel spectrograms from each sample by randomly sampling the specified duration 250 times from each of the 16 files. Specific hyperparameters for the generation of mel spectrograms were selected by random search function as outlined in subsection 4.5.

4.5 Random search optimization

To better understand the impact of both model structure and data preparation on the task, we randomly search 128 different combinations of model tuning hyperparameters and data settings for the creation of the mel spectrograms. Approximately one fifth of the combinations failed due to excessive memory requirements or training times and are thus omitted (male n=27, female n=27, agnostic n=25). Hyperparameters included in the random search (and their possible values) are listed in Table 1 alongside the hyperparameters selected by the top performing model (selected by highest Macro F_1) for each type (e.g, agnostic). All other hyperparamters were left to the Librosa or PyTorch defaults.

Table 1: Hyperparameters included in the random search.
Possible values for each hyperparameter are summarized in the second column. Hyperparameters selected by the top performing models are provided in their respective column. n_fft refers to FFT window size. n_fft and mel banks are sampled using a uniform distribution in the given range, while the max frequency is sampled from a log uniform distribution to better probe lower thresholds of frequency. Window length defaulted to match n_fft, and hop length indicated as fraction of n_fft.

Data Hyp.	**Range**	**Agnostic**	**Male**	**Female**	**SVM**
Duration (sec)	.25 / .5 / 1	250	250	500	250
n_fft	32 - 2048	2048	64	2048	1024
# Mels	50 - 256	244	51	78	59
Hop Len.	1/8, 1/4, 1/2	1024	8	1024	256
Max Freq. (Hz)	148 - 8100	7368	3659	2330	752
Model Hyp.					
L2 Reg.	0 - .01	.001	.001	.003	-
Dropout CNN	0.1 - 0.9	.40	.45	.19	-
Dropout Dense	0.1 - 0.9	.74	.48	.23	-
Batch Size	4 - 128	88	90	55	-
Epochs	200	6	81	2	-
CNN1 Filters	2 - 64	63	56	56	-
CNN2 Filters	2 - 64	7	38	45	-
Kernel Size	3 , 5	3	3	3	-
Dense Units	64-256	181	82	191	-

5 Results

Results for the top performing models (baseline, gender-agnostic, gender-dependent) are shown in Table 2. Results for each model are summarized in the second column in terms of $macroF_1$, the metric used to select the top performing models. Subsequent columns provide *positive* (intoxicated) and *negative* (sober) F_1 scores. Group averages are provided followed by scores for each participant. As the number of items in each class was balanced no weighting of scores was necessary (i.e. chance is equal to 50%). Mean accuracy as calculated from the mean of negative and positive recall (the metric of choice for the Interspeech Intoxication Challenge [10]) did not noticeably differ from Macro F_1 scores (SVM = 57.0%, gender-agnostic = 60.5%, female = 67.2%, male = 68.2%)

All three models (i.e., gender agnostic, male, and female) outperformed the SVM baseline metrics (F_1, positive/negative precision, recall, and F_1). Despite training on much smaller datasets, the F_1 of gendered combinations (male = 67.9, female = 68.6) are substantially higher than the gender agnostic model (59.0). A multivariate regression evaluating the effect of model type and speaker on F_1 values found that gender-dependent models significantly outperformed the SVM (p=.01) and the gender-agnostic model (p=.02). A type II multivariate analysis of variance (MANOVA) suggests that the effect of speaker (p=.004) is stronger than that of model (p=.025). These observations are reflected in the tendency for certain speakers to perform extremely well across all experiments.

For the gender-agnostic model, performance is highly variable across participants, while notable improvements to the consistency of precision and recall across participants can be observed for the gendered models. Between the two gendered models, the male-only model demonstrates better consistency in performance across participants (St. Dev of F_1 - Males : 8%, Females : 18.8%). The rate of false positives is highest for the female-only model, but lowest for the male model (positive precision - Female : 63.0%, agnostic : 64.5%, male : 69.3%). The rate of false negatives is highest for the gender-agnostic model, but lowest for the female model (negative precision - agnostic : 61.2%, male 68.8%, female 75.2%).

6 Discussion

In this study, we present further evidence that cannabis intoxication results in salient alterations to the speech stream. We have also demonstrated that a deep learning system can successfully exploit these alterations to predict cannabis intoxication from the vowel phonations of a previously unseen speaker at a rate substantially higher than chance. Considering the simplicity of the model architecture and the limited dataset, these results are encouraging. We believe this reflects the substantial degree to which cannabis intoxication affects voice quality. We suspect more training data from a larger number of speakers will help further improve the system and achieve state-of-the-art performance comparable to the detection of alcohol [7] and MDMA intoxication [4] from speech.

Our results indicate that biological speaker characteristics, such as sex, may influence the necessary data and model hyperparameters for accurate discrimination. This is evidenced by our gender-specific models outperforming the gender-agnostic model, despite the latter having more than double the pool of training data when testing on each fold. This suggests that acoustic correlates used by the model may differ in their manifestation in male and female speech. We note that the most notable acoustic difference between males and females is that of fundamental frequency (F0). Our previous work on

Table 2: Model performance for 8-fold gender-agnostic, and 4-fold male & female models.

Model	Macro F_1	Pos/Neg	Avg	F1	F2	F3	F4	M1	M2	M3	M4
SVM	55.0	+ F_1	56.6	69.5	37.3	75.8	82.3	59.6	53.9	39.2	35.0
		- F_1	53.8	42.6	39.0	76.1	85.4	20.2	61.3	65.5	40.3
Agnostic	59.0	+ F_1	56.0	38.8	35.2	74.3	88.8	43.1	68.7	44.1	54.9
		- F_1	62.0	72.0	52.1	69.7	89.5	51.0	52.7	64.2	44.8
Female	68.6	+ F_1	71.7	45.0	62.6	74.8	86.8				
		- F_1	60.9	41.6	78.4	76.4	89.3				
Male	67.9	+ F_1	67.6					74.2	75.7	54.7	66.0
		- F_1	68.1					67.8	74.2	55.2	75.3

this dataset [25] found notable alterations to several features of F0, namely range, trajectory, and acoustic shimmer. We acknowledge that the small dataset and aggressive hyperparameter search likely resulted in arbitrary selection of some hyperparameters. In spite of this, we find it noteworthy that the length of the fourier transform (n_fft) differed for top performing male and female models. Figure 1 provides the n_fft for the top ten performing models of each CNN variant (female, male, combined).

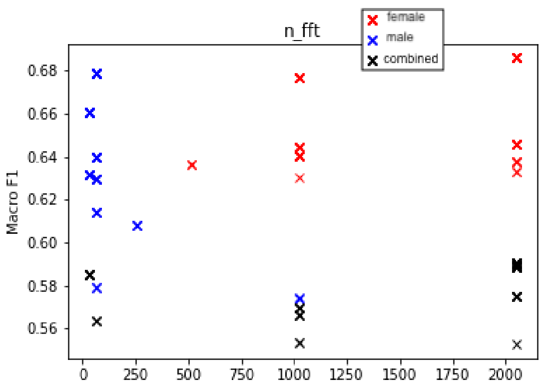

Figure 1: Performance ($macro\ F_1$) of top 10 gender-dependent models sorted by n_fft.

This may suggest that increased temporal resolution was necessary for male classification, and frequency resolution for female classification. It is possible that temporal aspects of F0 such as jitter and shimmer were more predictive for the male model, whereas F0-dependent features such as trajectories, range, and deltas were more predictive for both the female and gender-agnostic models. Another hyperparameter which differed consistently between top performing model variants was the maximum frequency (fmax) of the mel spectrogram. Figure 2 provides the maximum frequency of the spectrogram for the top ten performing models of each CNN variant (female, male, combined) and the respective F1 score of each configuration. Top performing female models had a tendency to select a lower fmax than the male model or combined models. In sum, higher temporal resolution and maximum frequency range produced better classification for the male dataset. Whereas higher frequency resolution and lower range produced higher performance on the female datset.

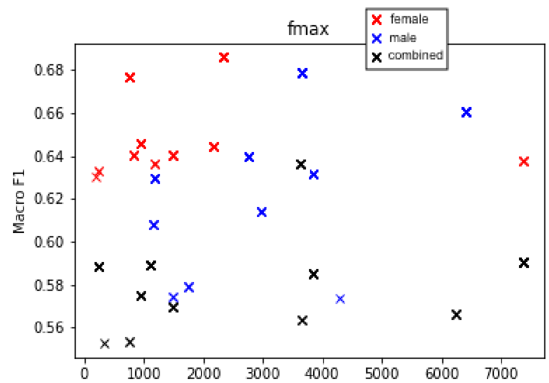

Figure 2: Performance ($macro\ F_1$) of top 10 gender-dependent models sorted by fmax

The results of our MANOVA suggest that individual speaker characteristics substantially influence model performance. Considering the small number of speakers, the risk for individual speaker differences to influence hyperparameter selection is high. Future work including a larger number of speakers may allow more confident insight into the utility of correlates and more nuanced evaluations of individual differences such as degree of intoxication, tolerance, and frequency of use.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Compute Canada.

References

[1] Daniel Bone, Matthew P Black, Ming Li, Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. Intoxicated speech detection by fusion of speaker normalized hierarchical features and gmm supervectors. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[2] Alex Frid, Edmond J Safra, Hananel Hazan, Lorrey L Lokey, Dan Hilu, Larry Manevitz, Lorraine O Ramig, and Shimon Sapir. Computational diagnosis of parkinson's disease directly from natural speech using machine learning techniques. In *2014 IEEE International Conference on Software Science, Technology and Engineering*, pages 50–53. IEEE, 2014.

[3] Olivia Murton, Stefanie Shattuck-Hufnagel, Jeung-Yoon Choi, and Daryush D Mehta. Identifying a creak probability threshold for an irregular pitch period detection algorithm. *The Journal of the Acoustical Society of America*, 145(5) :EL379–EL385, 2019.

[4] Carla Agurto, Guillermo A Cecchi, Raquel Norel, Rachel Ostrand, Matthew Kirkpatrick, Matthew J Baggott, Margaret C Wardle, Harriet de Wit, and Gillinder Bedi. Detection of acute 3, 4-methylenedioxymethamphetamine (mdma) effects across protocols using automated natural language processing. *Neuropsychopharmacology*, 45(5) :823–832, 2020.

[5] Li Bin, Matthew C Kelley, Daniel Aalto, and Benjamin V Tucker. Automatic speech intelligibility scoring of head and neck cancer patients with deep neural networks. In *International Congress of Phonetic Sciences (ICPHs' 19), Melbourne, Australia*, 2019.

[6] Che-Wei Huang and Shrikanth Shri Narayanan. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 583–588. IEEE, 2017.

[7] Weiqing Wang, Haiwei Wu, and Ming Li. Deep neural networks with batch speaker normalization for intoxicated speech detection. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1323–1327. IEEE, 2019.

[8] Tobias Bocklet, Stefan Steidl, Elmar Nöth, and Sabine Skodda. Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues. In *Interspeech*, pages 1149–1153, 2013.

[9] Arian Shamei and Sonya Bird. An acoustic analysis of cannabis-intoxicated speech. *Canadian Acoustics*, 47(3) :108–109, 2019.

[10] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski. The interspeech 2011 speaker state challenge. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[11] Florian Schiel, Christian Heinrich, and Sabine Barfüsser. Alcohol language corpus : the first public corpus of alcoholized german speech. *Language resources and evaluation*, 46(3) :503–521, 2012.

[12] Kim Berninger, Jannis Hoppe, and Benjamin Milde. Classification of speaker intoxication using a bidirectional recurrent neural network. In *International Conference on Text, Speech, and Dialogue*, pages 435–442. Springer, 2016.

[13] Nicholas Cummins, Alice Baird, and Björn W Schuller. Speech analysis for health : current state-of-the-art and the increasing impact of deep learning. *Methods*, 151 :41–54, 2018.

[14] Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3) :328–339, 1989.

[15] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47 :312–323, 2019.

[16] Noushin Hajarolasvadi and Hasan Demirel. 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*, 21(5) :479, 2019.

[17] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10) :1440–1444, 2018.

[18] Zixing Zhang, Felix Weninger, Martin Wöllmer, Jing Han, and Björn Schuller. Towards intoxicated speech recognition. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1555–1559. IEEE, 2017.

[19] Morris S Odell, Matthew Y Frei, Dimitri Gerostamoulos, Mark Chu, and Dan I Lubman. Residual cannabis levels in blood, urine and oral fluid following heavy cannabis use. *Forensic science international*, 249 :173–180, 2015.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch : An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[21] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2018. cite arxiv :1803.08375Comment : 7 pages, 11 figures, 9 tables.

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1) :1929–1958, 2014.

[23] Matthew D Zeiler. Adadelta : an adaptive learning rate method. *arXiv preprint arXiv :1212.5701*, 2012.

[24] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa : Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[25] Arian Shamei and Sonya Bird. An acoustic analysis of cannabis intoxicated speech, 2019.