SPEECH MARKERS FOR CLINICAL ASSESSMENT OF COCAINE USERS

Carla Agurto[1], *Raquel Norel*[1], *Mary Pietrowicz*[1], *Muhammad Parvaz*[2], *Sivan Kinreich*[3], *Keren Bachi*[2], *Guillermo Cecchi*[1], *Rita Z. Goldstein*[2]

[1]Computational Biology Center, T.J. Watson IBM Research Laboratory, New York, [2]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York
[3]SUNY Downstate Medical Center, New York

ABSTRACT

One of the main foci of addiction research is the delineation of markers that track the propensity of relapse. Speech analysis can provide an unbiased assessment that can be deployed outside the lab, enabling objective measurements and relapse susceptibility tracking. This work is the first attempt to study unscripted speech markers in cocaine users. We analyzed 23 subjects performing two tasks: describing the positive consequences (PC) of abstinence and the negative consequences (NC) of using cocaine. We perform two main experiments: first, we analyzed whether acoustic and semantic features can infer clinical variables such as the Cocaine Selective Severity Assessment; then, we analyzed the main problem of interest: to see if these features are powerful enough to infer if the subjects remains abstinent. Our results show that speech features have potential to be used as a proxy to monitor cocaine users under treatment to recover from their addiction.

Index Terms— drug addiction, cocaine, acoustic, semantic, abstinence.

1. INTRODUCTION

Drug addiction is a complex disease process of the brain that involves a recurring cycle of intoxication, bingeing, withdrawal and craving, resulting in an excessive drug use despite the devastating consequences [1]. Cocaine is considered as one of the most addictive psychoactive drugs. Given its short half-life and mechanism of action, the US government classifies this stimulant as having a high risk of dependency and high potential for abuse.

Evaluation of cocaine users is usually done using interviews and questionnaires, the primary objective being to explore cues that are associated with the propensity to relapse among those users who seek treatment. Although these assessments are useful for characterizing cocaine addiction, cognitive and biological markers are also desirable because they offer objective and reliable measurements of conditions of interest such as cravings in cocaine addicted individuals. The advantages of using speech over other biological markers such as brain function, blood tests, etc., include 1) low cost, 2) ease of data sample collection, and 3) the ability to infer cognitive information via semantic analysis.

The chronic use of cocaine, as many of other drugs, may cause injuries that affect speech such as inflammation of the vocal cords, resulting in hoarseness. Indeed, cocaine consumption can result in vocal fatigue, loss of vocal range, and laryngitis [2,3]. However, prior work studying speech markers in cocaine users is sparse. To the best of our knowledge, the only other relevant prior work used semantic fluency [4] to measure the ability to name as many words from a specified semantic category (e.g., animals, fruits or vegetables) within a discrete time (e.g., 1 min) in cocaine addicted individuals positive (current users) or negative (abstinent) for cocaine in urine. However, this work did not examine either acoustic or content properties.

In this work, we analyzed cocaine addicted individuals' free speech through two different tasks: describing positive consequences of being abstinent and negative consequences of drug consumption. We computed acoustic and semantic features using well-known acoustic toolboxes (Praat [5] and OpenSmile [6]) and a word vector representation package (GloVe [7]). The main goal of this study is to find objective speech markers than can be used in the evaluation of cocaine users in initially treatment-seeking. In addition, we take a first step in applying our speech marker selections by using them to predict abstinence in cocaine users (a component of relapse prediction). This paper is organized as follows. In section 2, we describe all the features (acoustic and semantic) used to characterize speech, section 3, we present the results and their discussion using statistical and machine learning methods. Finally, conclusions are stated in Section 4.

2. DATA ACQUISITION

2.1 Participants

The study was approved by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai. Participants were recruited from local addiction treatment facilities as well as using newspaper ads and word of mouth. A total of 23 cocaine users were recruited. Six of the participants were females with an average age of 47 (SD=11), and 17 were males with an average age of 52 (SD=8). Participants were asked to perform the following two free speech tasks: describing the positive consequences (PC) of abstinence, and describing the negative consequences (NC) of using cocaine. All participants were American and spoke in English. Each speech task produced a 5-minute recording. Five of the subjects performed the tasks again after a period of 3 months.

2.2 Clinical Variables

The following variables were collected during the data acquisition: 1) Cocaine Selective Severity Assessment (CSSA), which assesses withdrawal signs and symptoms within the past 24 hours; 2) Beck Depression Inventory (BDI), which assesses depression symptoms within the past 2 weeks; 3) days since last drug use/days of abstinence (DoA); 4) Craving Questionnaire (CQ) score, which measures craving; and 5) Cocaine Consumption days in the last 30 days (CC30d). To evaluate if the speech features were robust enough to infer abstinence, the last variable was binarized (0 days vs. at least 1 day).

3. METHODS

Here we describe the acoustic and semantic features extracted to analyze the data.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

3.1. Acoustic Features (AF)

To characterize the cocaine users' speech, we computed 315 features that captured 6 different types of voice characterization metrics (See Table 1), which are explained below:

3.1.1 Pitch variations

Pitch values are obtained in frames of 40ms using the autocorrelation method with a Hanning window of 5ms in Praat [8]. Since a distribution of values is obtained, we computed 6 statistical descriptors: median, interquartile range from 75th to 25th percentiles (IQR), 5th (pct5) and 95th (pct95) percentiles and 3rd and 4th moments (skewness and kurtosis). In addition to analyzing statistics from the pitch distribution, we also calculate the mean and standard deviation of the glottal pulse period (pitch period length).

3.1.2 Voice Quality

For voice quality, we extracted features that captured voice stability and noise. The features were calculated using Praat and were corrected for gender. The voice stability features included jitter (local absolute and ppq5) and shimmer (local absolute and apq5) [9]. In addition, we computed the fraction of locally unvoiced frames and the total duration of the breaks between the voiced parts of the signal, divided by the total duration of the analyzed part of the signal. To measure the noise, we compute the harmonics to noise ratio, noise to harmonics ratio (also called Harmonicity), and mean autocorrelation.

3.1.3 Vowel space assessment

Changes in formant distribution, specifically in the vowel space area composed by formants #1 and #2, have been found to be informative for predicting degeneration in Parkinson's disease and amyotrophic lateral sclerosis [10-12] as well as in predicting emotional states (e.g. anger) [13,14]. In addition, previous research [15] shows that formant # 2 (F2) can help discriminate subjects who are under the influence of drugs from those who are not. Therefore, we consider the information of the formants in the characterization of cocaine addiction. To extract formant information, we extracted the vowels from each recording using a Praat plug-in (details can be found in [16]). Then, we applied a pre-emphasis filter and set the maximum number of formats to 5 and the highest frequency to 5000 Hz (males) or 5500Hz (females). From the 5 formants, only the first 3 (F1, F2, F3) were analyzed along with their respective bandwidths. Since the values are obtained using 25-ms frames, we calculate the median, IQR, pct5, pct95, skewness and kurtosis over the obtained formant and their bandwidth distributions.

To build the vowel space, we considered F1 and F2 and used the method of Sandoval et al. [17] to identify clusters containing the vowels that characterize the American English language. In this way, we can also measure variation with respect to the American English standard triangle vowel (a-i-u) or polygon vowel area (based on the overall area) as well as calculate centroid position and orientation (angle).

3.1.4 Temporal features

This category includes features obtained by measuring time between syllables and pause durations. The pause duration distribution, which is characterized by a silence threshold of -25 dB and a minimum duration of 100 ms, was estimated in the recording. In addition, we also estimated the number of syllables using a method proposed by Jong et al. [18] and compute the duration between syllables to generate its distribution. From the pause and syllable duration distributions, we compute the following statistical descriptors: median, IQR, pct5, pct95, skewness and kurtosis. In addition, based on the duration between syllables, we computed two more features: speech rate (number of syllables over the total duration of the recording) and articulation rate (number of syllables over the total recording time after pauses were removed).

3.1.5 Spectral Characterization

Changes in the frequency spectrum of the voice were measured by computing the long-term average spectra (LTAS). From the LTAS, we calculate its slope, the maximum energy and the frequency were the maximum value is obtained, as well as the median and the energy IQR.

3.1.6 Mel Frequency Cepstral Coefficients (MFCCs)

MFCC coefficients have been previously used for the characterization of emotion. To obtain these features, we used the open source toolbox called OpenSmile. MFCC calculations were smoothed using a moving average window of length 3 [6]. A total of 14 coefficients were calculated and for each of them, we compute the following 17 functionals (using as a basis the ones proposed for the Emotional challenge in 2009) as listed here: stddev, skewness, kurtosis, range, percentile 99, percentile1, peakMeanAbs, meanPeakDist, peakDistStddev, minPos (absolute position of the minimum value), maxPos (absolute position of the maximum value), linregc1 (the slope of a linear approximation of the contour), IQR50th-25th, IQR75th-50th, IQR75th-25th, flatness, amean (arithmetic mean of the contour).

Table 1: Summary of all features used in this work.

Number of features	Description
6	Pitch variation
9	Voice Quality
43	Vowel space features
14	Temporal Features
5	Spectral features
238	MFCCs (OpenSmile)
126	Semantic Features

3.2. Semantic Features (SF)

To analyze the semantic information of the free speech of cocaine users, we automatically transcribed the recordings using IBM® Speech to Text. Once we obtained the transcripts for all the recordings, we processed them using WordNetLemmatizer from NLTK [19]. In this work, we extracted only nouns, verbs, adjectives and adverbs. After that, we analyzed the words using a method called Global Word Vector representation (GloVe) [7]. This unsupervised method trained with Wikipedia 2014 and Gigaword 5, represents 6 billion words (tokens) in 300 dimensions. One of the main advantages of this method is that it preserves linear substructures in the word space (e.g. comparative-superlative, women-men).

Next, we selected words of interest to learn whether specific content could help us characterize cocaine users. We selected two lists of words based on 1) the most common topics expressed during the two speech tasks such as: 'steal', 'lie', 'happy', 'sad', 'danger', 'healthy', 'trust', 'job', 'sexual', 'money', 'family', and 2) prior knowledge on drug addiction research, including words such as: 'drug', 'addiction', 'intoxication', 'bingeing', 'withdrawal', 'craving', 'relapse', 'pleasure', 'cocaine', 'abstinence'. Finally, the cosine distance was calculated between each word obtained after applying the lemmatizer and the words of interest. Since a distribution of distances was obtained, 6 features were extracted, including median, IQR, skewness and kurtosis, pct90 and pct10.

3.3 Experimental design

Both of the proposed experiments evaluate the association between speech features and clinical variables independently for each speech task.

3.3.1 Experiment 1: Can we use AF and SF as a proxy for clinical variables?

To discover whether the obtained speech features could be used as objective measurements to evaluate cocaine users, we estimated the R^2 coefficient of the correlation between each feature and clinical variables. We also calculated its significance by computing its p-value. Since multiple comparisons were performed, we used false discovery rate correction at $q<0.05$ to detect statistical significant features.

3.2.2 Experiment 2: Can AF and SF be used to infer if the subjects remain abstinent for 30 days prior to performing the speech task?

The ideal situation would be to be able to monitor cocaine users at home to determine if they have been able to withdraw their addiction or not without requiring them to be evaluated by a clinician. For that purpose, we evaluate whether speech features could distinguish between two groups: subjects that were abstinent during the last 30 days of the trial, or current cocaine users. First, we performed a two-sample t-test to find the most relevant features. Then we evaluated their potential to discriminate both groups using multivariate predictive analysis.

Specifically, we standardized the features (mean = 0 and standard deviation = 1), and use linear support vector machines (SVM) classifier with a nested leave-one-subject-out cross validation approach. Performance was measured in terms of accuracy, and the confidence interval was calculated using bootstrap resampling.

In addition, we were also interested to see how the association between concepts (using semantic analysis) changed between abstinent and current cocaine users. For that reason, we computed the partial correlations for the most significant features.

4. RESULTS & DISCUSSION

4.1. Experiment 1: Infer clinical variables

Table 2 shows the top correlation obtained between the clinical variables and the two types of analyzed features. We observe a statistically significant correlation for almost all clinical variables and the acoustic features (marked by * in Table 2) but these correlations vary according to the speech task. For example, we observe that first MFCC coefficient (MFCC#1) flatness is significant regardless the task performed, but for DoA - F3 was found to be significant only for the PC task. This indicates that certain features are task specific. For example, the depression index (BDI) appears to correlate more with acoustic features when the task is PC; while the correlation is better with semantic features when the task is NC.

4.2. Experiment 2: Abstinence detection

4.2.1 Univariate Analysis

In table 3, we show the results of the two-sample t-test to discover for significant features which discriminate abstinence from current cocaine use. We observe that none of the features pass the statistical significance after FDR correction. However, we observe that AF are slightly better when the task is PC. On the other hand, SF are slightly better when the task is NC. Even though univariate analysis does not help detect significant differences between both classes, it may be the case that interactions among them may be more informative.

Table 2: *Most significant correlation using R^2 of extracted features and clinical variables*

Task	Type	Clinical Variable	R^2	p-value	Feature name
PC	AF	**CSSA**	0.34	1E-3	MFCC#10 linregc1
		CQ	0.25	7E-3	MFCC#7 maxPos
		BDI	0.43*	1E-4	MFCC#10linregc1
		DoA	0.51*	3E-5	F3 kurtosis
		CC30d	0.49*	5E-5	MFCC#1flatness
	SF	**CSSA**	0.19	2E-2	Healthy IQR
		CQ	0.22	1E-2	Happy IQR
		BDI	0.32	2E-3	Healthy kurtosis
		DoA	0.55*	1E-5	Craving pct90
		CC30d	0.19	2E-2	Healthy pct90
NC	AF	**CSSA**	0.48*	7E-5	MFCC#3 meanPeakDist
		CQ	0.39	6E-4	MFCC#8 amean
		BDI	0.26	5E-3	MFCC#12 maxPos
		DoA	0.28	5E-3	MFCC#14 amean
		CC30d	0.47*	9E-5	MFCC#1 flatness
	SF	**CSSA**	0.34	1E-3	Danger pct90
		CQ	0.21	2E-2	Steal median
		BDI	0.43*	2E-4	Danger pct90
		DoA	0.18	3E-2	Happy pct10
		CC30d	0.15	5E-2	Money pct10

4.2.2 Multivariate Analysis

Figure 1 shows the best performance results obtained with linear SVM. We observe that acoustic features better characterize abstinent from current cocaine users than semantic features in both tasks (81% NC and 89% PC). We also observe that both types of features present higher results for PC than NC which may indicate that the PC task is better to characterize abstinence. Bootstrap confidence intervals show that only acoustic features result in significantly different performance from chance accuracy.

4.2.3 Partial correlations

Fig.2 shows the results of partial correlations for the NC speech task which is the one that presents the lower p-value in Table 3. We observe that there are different associations for the two classes. Abstinent cocaine users show positive correlation between 'danger' and 'intoxication', and 'happy and trust'. They also present a negative correlation between 'withdrawal' and 'intoxication' which is expected in healthy people. However, the correlations found in current cocaine users show positive correlation between 'danger' and 'happy', and 'abstinence' and 'intoxication', which are unexpected for a healthy subject. This provides evidence that the perception of a cocaine-addicted individual changes when he or she is using the drug.

Table 3: Statistical significance for Abstinence classification (based on CC30d)

Task	Type	p-value	t-stat	Feature name
PC	**AF**	2.6E-4	4.24	MFCC#9 (linregc1)
	SF	1.3E-1	1.58	Relapse (IQR)
NC	**AF**	1.0E-3	3.71	Shimmer (local)
	SF	2.5E-2	2.38	Withdrawal (pct90)

5. CONCLUSIONS

We demonstrated that the acoustic and semantic features could potentially be used as a proxy for clinical evaluation of cocaine addicted individuals that seek treatment through objective assessments. In addition, our classification results showed high accuracy (81% for NC and 89% for PC) in detecting abstinence (in a period of 30 days) in cocaine users. To the best of our knowledge, this is the first attempt to study speech markers in cocaine users. This is an ongoing longitudinal study, where 3- and 6-month follow-up data is still being collected, which might help in the discovery of markers which predict relapse.

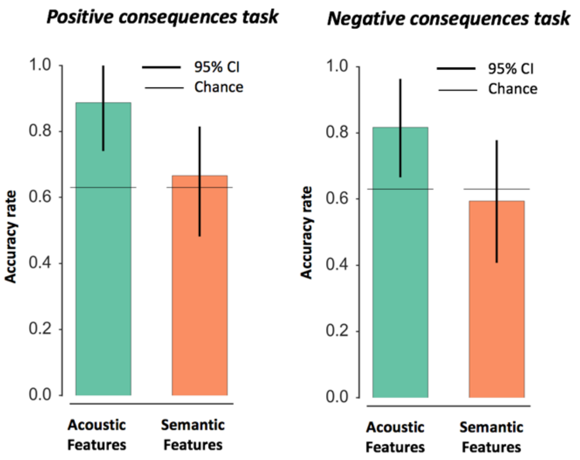

Figure 1: Best performance results for identifying abstinence using the two speech tasks.

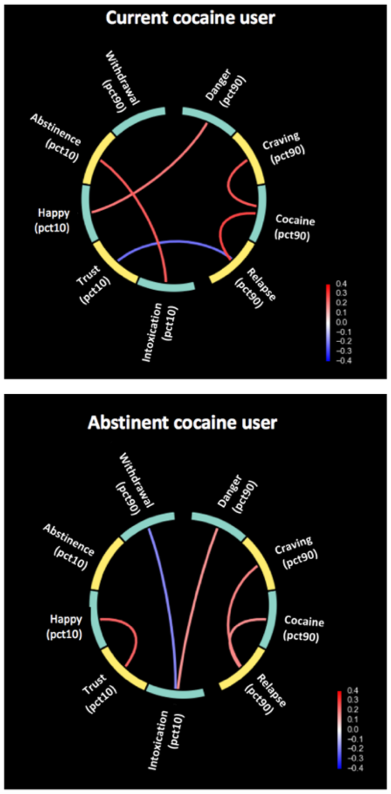

Figure 2: Partial correlations of semantic similarity obtained for speech task: NC.

6. ACKNOWLEDGEMENTS

We would like to thank to the participants of this study. The collection of data in this study was funded by NIH grant R01DA041528 (PI: Goldstein)

6. REFERENCES

[1] Underlying Neurobiological Basis: Neuroimaging Evidence for the Involvement of the Frontal Cortex," *American Journal of Psychiatry*, vol. 159, no. 10, pp. 1642–1652, 2002.

[2] T. Moreira, C. Gandez, L.R. Figueiró, D.M. Capobianco, K. Cunha, M. Ferigolo, H.M.T.Barros, M.Cassol, "Substance use, voice changes and quality of life in licit and illicit drug users," *Revista CEFAC*, vol. 17, no. 2, pp. 374-384, 2015.

[3] A.C.N. Filho, S.G. Bettega, S. Lunedo, J.E. Maestri, F. Gortz, "Repercussões otorrinolaringológicas do abuso de cocaína e/ou crack em dependentes de drogas," *Rev. Assoc. Med. Bras*. Vol 45, no. 3, pp. 237-241, 1999.

[4] R. Goldstein, P. Woicik, T. Lukasik, T. Maloney, and N. Volkow, "Drug fluency: A potential marker for cocaine use disorders," *Drug and Alcohol Dependence*, vol. 89, no. 1, pp. 97–101, 2007.

[5] A. R. Bradlow, "A cross-language comparison of vowel production and perception: language- specific and universal aspects*," *Working Papers of the Cornell Phonetics Laboratory*, vol. 8, pp. 29–85, 1993.

[6] Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller: "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", *In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM*, pp. 835-838, October 2013.

[7] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[8] P. Boersma and V. van Heuven, "Speak and unSpeak with Praat," *Glot Int.*, vol. 5, no. 9–10, pp. 341–347, 2001.

[9] Teixeira, J. P.; Oliveira, C. & Lopes, C. "Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters". *Procedia Technology*, vol. 9, pp. 1112–1122, 2013

[10] P. A. McRae, K. Tjaden, and B. Schoonings, "Acoustic and perceptual consequences of articulatory rate change in Parkinson disease," *J. Speech Lang. Hear. Res*. Vol. 45, no.1, pp.35–50, 2002.

[11] G. S. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *J. Speech Hear. Res*, vol. 38, no.5, pp.1001–1013, 1995.

[12] G. Weismer, J.-Y. Jeng, J. Laures, R. D. Kent, and J. F. Kent, "Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders," *Folia Phoniatr. Logop,* vol. 53, pp. 1–18, 2001.

[13] M. Goudbeek, J. P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 1575–1578, 2009.

[14] S. Yildirim, M. Bulut, and C. Lee, "An acoustic study of emotions expressed in speech.," *Proc. InterSpeech*, pp. 2193–2196, 2004.

[15] C. Agurto, R. Norel, R. Ostrand, G. Bedi, H. D. Wit, M. J. Baggott, M. G. Kirkpatrick, M. Wardle, and G. A. Cecchi, "Phonological Markers of Oxytocin and MDMA Ingestion," *Interspeech 2017*, 2017.

[16] R. Corretge, "Praat vocal toolkit: overview.," 2012. [Online]. Available: http://www.praatvocaltoolkit.com/2012/11/cut-pauses.html. [Accessed: 05-Jan-2017].

[17] S. Sandoval, V. Berisha, R.L. Utianski, J.M. Liss, A. Spanias "Automatic assessment of vowel space area," *The Journal of the Acoustical Society of America*. Vol. 134, no.5, 2013.

[18] N. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behav. Res. Methods*, vol. 41, no. 2, pp. 385–90, 2009.

[19] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. 2009.