DISCRIMINATING GROUPS BY AUDIO FEATURE ANALYSIS WITH openSMILE

Jon Clucas[1*], Jake Son[1], Michael P. Milham[2,3], Arno Klein[1]
jon.clucas@childmind.org
[1]MATTER Lab, Child Mind Institute [2]Center for the Developing Brain, Child Mind Institute [3]Nathan Kline Institute

INTRODUCTION

This preliminary exploration indicates real, measurable differences in the sounds produced by individuals and that these differences can potentially distinguish between children with selective mutism and typically developing children. While these results indicate that an audio recording in which most of the vocalizations are produced by the subject of interest can be sufficiently robust to other voices and environmental sounds to model these group differences, the sources of differential signals was not determinable by our methods.

Selective mutism (SM) is a condition in which afflicted individuals fail to speak in certain social environments but not others [1]. Currently, the mental health community lacks sufficient objective, quantifiable measures for SM diagnosis and treatment monitoring [9]. An individual's diagnosis is largely dependent on subjective parent/teacher reports, complicating analysis without standard instruments or measures used to compare symptoms at the population level or individually over time. This condition, with its definitional relation to voicing, is a ripe target for automated audio analysis.

METHODS

Voice data are abundant and relatively inexpensive to collect, providing researchers with the potential to classify psychiatric groups and to track individual psychiatric changes over time. openSMILE (**open-S**ource **M**edia **I**nterpretation by **L**arge feature-space **E**xtraction) is one analysis tool that automatically extracts low-level audio features, originally developed as an "acoustic emotion recommendation engine and keyword spotter" [3] and is capable of extracting thousands of low-level audio features from recorded sound files.

We analyzed audio files captured during a previously conducted response paradigm [4][5]. We selected two prebuilt openSMILE configuration files: emobase.conf and ComParE_2016.conf. emobase, with "998 acoustic features for emotion recognition" [3], is the openSMILE configuration file with the most robust documentation; ComParE_2016 is the most recent prebuilt openSMILE configuration file available.

scikit-learn's random forest regressor [7][8] with 2,000 estimators was then run with the openSMILE output features as the independent variable values and each participant's selective mutism diagnostic status as the dependent variable values.

The code used in preparation of this paper is available on GitHub, including a Jupyter notebook set up to replicate these analyses and explore the full range of models, predictions and outputs (Link 1), and all of the data (excluding the original sound files) are available on Open Science Framework (Link 2).

A differential signal is a difference between child voices and adult voices instead of or in addition to a difference between selectively mute voices and typically developing voices. To try to untangle these differences, we manually checked each file for audible adult vocalizations and marked those segments for removal or replacement, marking boundaries of the relevant segments using Audacity, a freely available digital audio editor.

For details about replacement methods considered and tested, see this poster's article in this conference's proceedings [2]. We replaced all of the noted adult vocalizations with each of the four following replacement methods: 1) silenced (time unchanged); 2) removed (time reduced); 3) replaced with computer-generated same-duration pink noise; and 4) replaced with randomly-selected same-duration low-amplitude segment from same recording. We re-ran our initial analysis on all four versions of our cleaned sound files and on the isolated adult sound files.

RESULTS

Initial random forest analyses on the openSMILE output features resulted in predictive values above 0.5 (chance level) for each openSMILE configuration file for both vocal conditions, and surprisingly for both button-press conditions, in which no vocalization was included in the protocol. Listening to the button-press conditions with the highest probability of voicing revealed the presence of adult voices (both parents and experimenters) in some of the recordings.

After removing the audible adult vocalizations, the predictive power of the random forests regressor increased in all four experimental conditions regardless of replacement method or configuration file (see Table 1). The predictive value of the isolated adult vocalizations was also greater than that of the original sound files in three of the four experimental conditions, the exception being vocal response condition with stranger presence (see Table 2).

openSMILE config file			emobase		ComParE_2016	
adult vocalizations	**experimental condition**		button press	vocal response	button press	vocal response
silenced	stranger presence	yes	0.785714	0.902423	0.714281	0.853655
		no	0.738079	0.833334	0.738103	0.809551
removed		yes	0.809530	0.878049	0.809555	0.853617
		no	0.785707	0.833337	0.714306	0.809535
replaced with computer-generated same-duration pink noise		yes	0.809538	0.853657	0.809542	0.853616
		no	0.738095	0.809523	0.809529	0.785745
replaced with randomly-selected same-duration low-amplitude segment from same recording		yes	0.809538	0.878038	0.78574	0.853635
		no	0.809518	0.809531	0.761923	0.785756

Table 1: Random forests out-of-bag predictive confidence values of SM vs. control.

openSMILE config file		emobase		ComParE_2016	
experimental condition		button press	vocal response	button press	vocal response
stranger presence	yes	0.827581	0.538182	0.827556	0.307381
	no	0.965512	0.793098	0.965507	0.758636

Table 2: Random forests out-of-bag predictive confidence values for isolated adult vocalizations.

FUTURE WORK

These results indicate audible differentiability between groups of children with and without a diagnosis of selective mutism, in the voices of the children themselves and in the way adults speak to and in the presence of these children. To further identify the relevant signals, more work is needed.

Comparing ambient sounds with children from each group and from the room without anyone within would reduce the possibility that ambient sounds during data collection provide a confounding artifactual signal. Comparing the voices of parents in both the presence and the absence of their children would help to identify the signal differentiating the groups in 3/4 of the conditions in Table 2. Comparing the voices of study coordinators when speaking to children in each group, parents in each group, and in the absence of participants would also help to identify the differential signal in Table 2. Comparing voices recorded in additional contexts would help to identify differential signal and to identify confounding noise.

REFERENCES

1. American Psychiatric Association. 2013. Selective mutism. In *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American Psychiatric Association, Arlington, VA, 194–197.
2. Jon Clucas, Jake Son, Michael P Milham, and Arno Klein. 2018. Discriminating Groups by Audio Feature Analysis with openSMILE. *Proceedings of the 3rd Symposium on Computing and Mental Health*. http://mentalhealth.media.mit.edu/wp-content/uploads/ sites/46/2018/04/CMH2018_paper_41.pdf
3. Florian Eyben, Felix Weninger, Martin Wöllmer, and Björn Schuller. 2016. *openSMILE:) open-Source Media Interpretation by Large feature-space Extraction version 2.3, November 2016* (2.3 ed.). Gilching, Germany. http://www.audeering.com/research-and-open-source/files/openSMILE-book-latest.pdf
4. Erica J. Ho, Lindsay M. Alexander, Nicolas Langer, et al. 2016A. Novel Techniques for Elucidating Neurophysiological Mechanisms of Selective Mutism. (Oct 2016A).
5. Erica J. Ho, Lindsay M. Alexander, Nicolas Langer, et al. 2016B. Novel Techniques for Elucidating Neurophysiological Mechanisms of Selective Mutism. *Journal of the American Academy of Child & Adolescent Psychiatry* 55, 10S (Oct 2016B), S231. DOI:10.1016/j.jaac.2016.09.401
6. Child Mind Institute MATTER Lab. 2018. mhealthx software pipeline. (2018). http://matter.childmind.org/mhealthx.html
7. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (Oct 2011), 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
8. scikit-learn developers. 2017. Random Forests. In *scikit-learn User Guide*. 1.11.2.1. http://scikit-learn.org/stable/modules/ensemble.html#random-forests
9. Helen Xu, Jacob Stroud, Renee Jozanovic, et al. 2018. Passive Audio Vocal Capture and Measurement in the Evaluation of Selective Mutism. *bioRxiv* 250308 (Jan 2018). DOI:10.1101/250308

LINKS

1. https://github.com/ChildMindInstitute/selective-mutism-response-paradigm-sound-analysis/releases/tag/v0.1.0
2. https://osf.io/ut59y