Discriminating Groups by Audio Feature Analysis with openSMILE

Jon Clucas
Jake Son
MATTER Lab
Child Mind Institute
New York, NY 11102, USA
jon.clucas@childmind.org
jake.son@childmind.org

Michael P. Milham
Center for the Developing Brain
Child Mind Institute
New York, NY 11102, USA
Nathan Kline Institute
Orangeburg, NY 10962, USA
michael.milham@childmind.org

Arno Klein
MATTER Lab
Child Mind Institute
New York, NY 11102, USA
arno.klein@childmind.org

Abstract

Voice data is abundant and almost certainly rich with relevant signs for mental health assessment and longitudinal tracking. MATTER Lab is building a multimodal analysis framework including the mhealthx software pipeline [5]. openSMILE is a software package for audio signal processing. As a preliminary exploration of openSMILE for inclusion in our framework, random forest regressors were applied to features extracted by the software from audio files collected during a pediatric psychiatric study. The initial analysis suggests a need for quality control, to identify spurious factors and experimental design flaws. Discovered extraneous voices (parents, researchers and 1 sibling) were semiautomatically replaced in each file where present, using four different methods. The random forests analysis was repeated on each of the resulting datasets. Surprisingly, the analysis still better than chance for nearly all conditions. While replacing the extraneous voices moved the control condition predictions closer to chance and the experimental conditions further from chance, the results demonstrate lingering confounds after all attempts at correction. As voice data is rarely if ever perfectly separable from environmental context, exploring the limits of decontextualization will be invaluable.

Author Keywords
audio; feature analysis; voice; pediatrics; psychiatry; openSMILE

CCS Concepts
•Applied computing → Health informatics;

Introduction
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" [2], p. 6 and is capable of extracting thousands of low-level audio features from recorded sound files.

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 [8]. 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.

Initial Methods
We analyzed audio files captured during a previously conducted response paradigm [3, 4]. We selected two prebuilt openSMILE configuration files: emobase.conf and ComParE_2016.conf. emobase, with "998 acoustic features for emotion recognition" [2], 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 [6, 7] with 2,000 estimators was then run with the openSMILE output features as the independent variable values ('X') and each participant's selective mutism diagnostic status as the dependent variable values ('Y').

Initial Results
Initial random forest analyses on the openSMILE output features resulted in predictive values above 0.5 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.

Audio Cleanup Methods
A possible experimental confound is a difference between child voices and adult voices rather than between selectively mute voices and typically developing voices. To try to get a closer estimate of the latter difference, 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.

We considered nine methods to compensate for removing a segment of a sound file: 1) removal without replacement, where the audio clip before and after the removed segment are simply joined, "timeshifting" the resulting audio. For the other methods we replaced the removed segment with 2) silence, 3) white noise, 4) pink noise, 5) brow-

nian noise, 6) ambient sound adjacent to the removed clip temporally stretched over the duration of the removed clip, 7) a clip generated from the spectral profile of the overall sound, 8) a clone mask and 9) a clone mask applied not just to the removed segment but to the entire recording. The clone masks were created by identifying low-amplitude portions of the given sound file, randomly selecting one of the identified clips, appending a reversed copy to the end of the clip zero or more times until the mask is at least as long as necessary, then trimming the clip to the exact duration required. We used Audacity to create the replacement clips for methods 3–5 and for the other methods we used ffmpeg wrapped in Python with Pydub and SciPy (https://github.com/ChildMindInstitute/selective-mutism-response-paradigm-sound-analysis/tree/ad75e12162322cbc9ccb3e7ef44663faebb6a8be/SM_openSMILE/openSMILE_preprocessing/noise_replacement/noise_replacement.py).

With a single, randomly selected file from our sample, we removed a randomly selected clip of ambient noise and tested each of the above nine correction methods. After running each of the modified sound files through both of our openSMILE configuration files, we summed the median absolute deviations from the original for each feature in the outputs.

From these results, we selected three correction methods to test across all of our sound files: 1) high-efficacy replacement clone mask, and quick options 2) silence and 3) deletion ("timeshift"). Applying the same comparison as in our single-file test, we found a similar pattern of fidelity to the original sound's low-level audio properties. We replaced all of the noted adult vocalizations with each of these three replacement methods, as well as pink noise replacement. We compared the openSMILE outputs of each of these files with those of their respective original sound files; we also compared the openSMILE outputs of the isolated adult vocalizations to those of their respective original sound files. Based on these comparisons, we re-ran our initial analysis on all four versions of our cleaned sound files and on the isolated adult sound files.

Audio Cleanup Results

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 that that of the original sound files in three of the four experimental conditions (see Table 2).

Discussion

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. These results also 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, though as expected, the signal-to-noise ratio appears to be diminished in the presence of extraneous voices.

Future Directions

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 (https://github.com/ChildMindInstitute/selective-mutism-response-paradigm-sound-analysis/releases/tag/v0.1.0), and all of the data (excluding the original sound files) are available on Open Science Framework (https:

openSMILE config file			emobase		ComParE_016	
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 w/ 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	0.785740	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.

//osf.io/ut59y/). Having established some degree of confidence that measurable differences exist in the sounds produced by individuals, we can employ more sophisticated analyses to develop useful, automated, objective measures for diagnosis and longitudinal evaluation.

Acknowledgements

We thank Bonhwang Koo (Center for the Developing Brain, Child Mind Institute) and Helen Xu (Sidney Kimmel Medical College, Thomas Jefferson University; Center for the Developing Brain, Child Mind Institute) for their assistance in conducting this research.

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. 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
3. Erica J. Ho, Lindsay M. Alexander, Nicolas Langer, Maki S. Koyama, Helen Y. Xu, Renee K. Jozanovic, Bonhwang Koo, Lei Ai, Jacob Stroud, Grace Russo, Rachel Busman, Michael P. Milham, and Simon P. Kelly. 2016A. Novel Techniques for Elucidating Neurophysiological Mechanisms of Selective Mutism. (oct 2016A).
4. Erica J. Ho, Lindsay M. Alexander, Nicolas Langer, Maki S. Koyama, Helen Y. Xu, Renee K. Jozanovic, Grace Russo, Rachel Busman, Michael P. Milham, and Simon P. Kelly. 2016B. Novel Techniques for Elucidating Neurophysiological Mechanisms of Selective Mutism. *Journal of the American Academy of Child & Adolescent Pyschiatry* 55, 10S (oct 2016B), S231. DOI: http://dx.doi.org/10.1016/j.jaac.2016.09.401
5. Child Mind Institute MATTER Lab. 2018. mhealthx software pipeline. (2018). http://matter.childmind.org/mhealthx.html
6. Fabian Pedregosa, GaÃńl Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and ÃL'douard Duchesnay. 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
7. 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
8. Helen Xu, Jacob Stroud, Renee Jozanovic, Jon Clucas, Jake Son, Bonhwang Koo, Juliet Schwarz, Arno Klein, Rachel Busman, and Michael P Milham. 2018. Passive Audio Vocal Capture and Measurement in the Evaluation of Selective Mutism. *bioRxiv* 250308 (jan 2018). DOI:http://dx.doi.org/10.1101/250308