Multimodal Depression Classification Using Articulatory Coordination Features $\hat{\mathcal{S}}$ and Hierarchical Attention Based Text Embeddings

INTRODUCTION

- [\
- [\
- L
- A

jor depressive diso an estimated 3.8% nitoring and provid hortage of clinicians l digital health techno D is accompanied condition of slowed eation, and motility a guage conveys a gen ultimodal binary of Articulatory coordinar gestures (speech mod lierarchical attention- improve the perfor	order (MDD) is a lea 6 of the population ding treatments he imits the timely acce ology can help clinicia by Psychomotor sl neuromotor output nd a necessary featu reat amount of infe depression classifie tion features to quar clality) based text embeddi ormance by aggreg	ading cause of di affected avily rely on hur ss to treatments ns monitor patien owing that manifests cha re of MDD ormation on a p r using: htify changes in tim ngs (language mon ating the segme	 Based on Psychomotor Slowing (PMS) [5,6] Altered coordination and timing across articula ACFs proposed by Huang et al. [7] that utilizes dil incorporating more delays to the correlation matr ⁶ TVs (Degree and Place of Constriction) Speech Inversion System Aperiodicity, Periodicity and Pitch Detector Source (glottal) Information 	
sification outputs vi	ia a recurrent neur	al work for sessi	on-level	5. Model Architectures
sification				Inimodal Audio Classifier
2.	DATASET PR	EPARATION		
DatabaseStudy TypeOkLongitudinal6 N# Subjects20Demography311 A1 A1 C1 C	MD-1 [1] Diservational Weeks F, 15 M Caucasian African American Bi-racial Greek, 1 Hispanic	MD-2 Clinical Trial 4 Weeks 104 F, 61 M 125 Caucasian 26 African American 4 Asian 10 Other		$\begin{array}{c} R_{ACF} - TV \\ \hline \\ $
Assessment HA Recording Type	AMD-CL: Bi-weekly Interactive Void	HAMD-CL, QIDS-CL: We	eeks 1,2,4 (8kHz)	
Lengths	Min: 2.5s, Max: 156.8s	Min: 2.6s, N	lax: 181.2s	
udio segments were created by applying a 20s window with 5s shifts d the resultant segments longer than 10s were included				Multimodal Classifier
ext data segmentat N HAMD Score	tion: sentences ot Depressed 0-7	Depressed 8-52		Audio Segment audio segment embeddings
3. \	Jocal Tract Va	riables (TVs	5)	Hierarchical Attention Network
Based on Articu	latory Phonology	/ [3,4]		
Constriction Organ	Tract \	Tract Variable		In the set of the
C	Lip Aperture Lip Protrusion		Upper Lip, Lower Lip, Jaw	Atte Sent Ren Ren La Field
ngue Body	Tongue body const	riction degree	Tongue Body, Jaw	9. REF
ngue Tip	Tongue tip constric	tion degree	Tongue Body, Tip, Jaw	 [1] J. C. Mundt, P. J. Snyder, M. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity 64, 2007. [2] Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R.," Vocal acoustic biomarkers of depression severity and treation for speaker independent of the sp
				jul 2019.

Major depressive disorder (MDD) is a leading cause of disability worldwide •Based on Psychome itle are estimated 2.0% of the set	otor Slowing (PMS
 A digital health technology can help clinicians monitor patients between visits A digital health technology can help clinicians monitor patients between visits A digital health technology can help clinicians monitor patients between visits MDD is accompanied by Psychomotor slowing A condition of slowed neuromotor output that manifests changes in speech, ideation, and motility and a necessary feature of MDD anguage conveys a great amount of information on a person's mental state A multimodal binary depression classifier using: Articulatory coordination features to quantify changes in timing across articulatory gestures (speech modality) Hierarchical attention-based text embeddings (language modality) 	Ation and timing acr Huang et al. [7] that delays to the corres egree and constriction: peech version ystem eriodicity, eriodicity, ad Pitch etector ce (glottal)
assification outputs via a recurrent neural work for session-level	5. Model Arch
 Unimodal Audio C 	Classifier
DataSet PREPARATION ∑udy Type Observational Longitudinal 6 Weeks 4 Subjects 20 F, 15 M 20 F, 15 M 104 F, 61 M Demography 31 Caucasian 1 African American 26 African American 1 Bi-racial 4 Asian 1 Greek, 1 Hispanic 10 Other Assessment HAMD-CL: Bi-weekly HAMD Score Min: 2.5s, Max: 186.2s Not Depressed Min: 2.5s, Max: 186.2s MAMD Score 07 Base Depressed MAMD Score 07 04 For the set Mark Score 04 For the set Mark Score	51x1x0 51x1x0 51x1x0 51x1x0 51x1x0 51x1x0 Batch Norm. Leaky Relu Max Pooling (2,1) Sequence of Segment Embeddings Masking LSTM1 Setting Masking LSTM1 Setting Masking LSTM1 Setting Masking LSTM1 Setting Masking LSTM1 Setting Masking LSTM1 Setting Masking LSTM1 Setting
3. Vocal fract variables (TVS)	ierarchical Attention Network
 Based on Articulatory Phonology [3,4] Constriction Organ Tract Variable Articulators 	ncoder ttention ncoder
Lip Aperture Lip Protrusion Lip, Jaw	" SĂ S™
Tongue BodyTongue body constriction degreeTongue Body, JawTongue body constriction locationTongue Body, Jaw	
Tongue TipTongue tip constriction degree Tongue tip constriction locationTongue Body, Tip, Jaw[1] J. C. Mundt, P. J. Snyder, M. Cannizzaro, K. Chap 64, 2007.[2] Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lender [3] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and iul 2019	appie, and D. S. Geralts, "Voice acoustic meas derking, W. R.," Vocal acoustic biomarkers of d nd C. Espy-Wilson, "Unsupervised speaker ada
Velum Velum [4] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, a [5] J. P. Williamson, D. Young, A. A. Nierenhorg, J. Young, Y.	n, and J. Singh, "Use of temporal information: (
Glottis Glottis [6] Christina Sobin and Harold Sackeim. "Psychon [7] Z. Huang, J. Epps, and D. Joachim, "Exploiting V	[6] Christina Sobin and Harold Sackeim. "Psychomotor symptoms of depression". In: The Ar [7] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated CNNS

Nadee Seneviratne, Carol Espy-Wilson University of Maryland – College Park, USA **4. Articulatory Coordination Features**

detection of periodicity, aperiodicity, and pitch in speech," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 776–78 ng depression severity from audio and video based on speech articulatory coordination," Computer Speech & Language, vol. 55, pp. 4 nerican journal of psychiatry (1997) for depression detection in naturalistic environments," in Proc. of 2020 IEEE International Conference on Acoustics, Speech and Signal

) [5,6] ross articulators at utilizes dilated CNNs and elation matrix



Classify Segments

16x1 O₂x1

•It can be shown for a binary classifier, if all the classes have a better than 50% recall in the segment-level classifier, plurality voting based session-level classifier would result in a better recall for all the classes

•More generalized results for RNN based session-level classifier

- Reasons for errors of the text model [Ground-truth: Not depressed]
 - Excessive use of negation depressed]
- Misclassified sessions by all models which was used to determine the ground-truth label depression severity threshold boundaries

•Improving the classification results by •Segment-to-session level classification with segment level classifier satisfying certain constraints •Effective multimodal systems can be developed using TV based ACFs and hierarchical attention-based text embeddings • Inter-learning among different modalities can compensate for the errors made by individual modalities

9. REFERENCES

Audic

sures of depression severity and treatment response collected via interactive voice response (ivr) technology," Journal of Neurolinguist

Softma

lepression severity and treatment response". Biological psychiatry, 72(7), 580–587, 2012.

aptation for speaker independent acoustic to articulatory speech inversion," The Journal of the Acoustical Society of America, vol. 146,





7. ERROR ANALYSIS

•Complex sentence structures (often with mixed sentiments)

•"I'm not feeling guilty and feeling like I cannot do anything like before"

•"I do not feel like I do not want to do anything" [Ground-truth: Not

•Sentiment not agreeing with the severity score assigned by the clinician •Quasi-numerical nature of HAMD scores leading to ambiguity at the

8. CONCLUSION

	ACKNOWLEDGEMENTS
ics, vol. 20, pp. 50–	This work was supported by the UMCP & UMB Artificial Intelligence + Medicine for High Impact Challenge Award
, no. 1, pp. 316–329,	and the National Science Foundation grant numbered 2124270. We thank Dr. James Mundt for the depression
86, 9 2005. 10 — 56, 2019	databases MD-1&2 [1, 2] and Dr. Thomas Quatieri and Dr. James Williamson for granting access to the MD-2
Processing	database which was funded by Pfizer