# Multimodal Depression Classification Using Articulatory Coordination Features and Hierarchical Attention Based Text Embeddings

Nadee Seneviratne, Carol Espy-Wilson
University of Maryland – College Park, USA

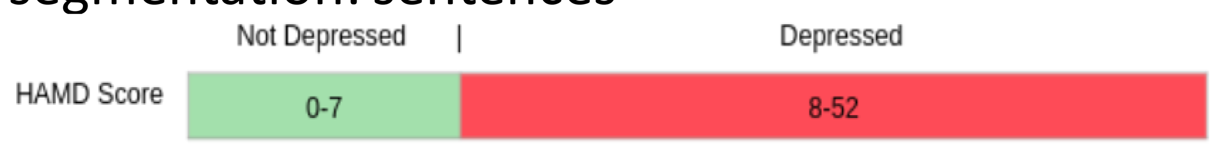
## 1. INTRODUCTION

- Major depressive disorder (MDD) is a leading cause of disability worldwide with an estimated 3.8% of the population affected
- Monitoring and providing treatments heavily rely on human intervention
  - Shortage of clinicians limits the timely access to treatments
  - 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
- Language 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)
- We improve the performance by aggregating the segment-level classification outputs via a recurrent neural work for session-level classification

## 2. DATASET PREPARATION

| Database | MD-1 [1] | MD-2 [2] |
|---|---|---|
| Study Type | Observational | Clinical Trial |
| Longitudinal | 6 Weeks | 4 Weeks |
| # Subjects | 20 F, 15 M | 104 F, 61 M |
| Demography | 31 Caucasian | 125 Caucasian |
| | 1 African American | 26 African American |
| | 1 Bi-racial | 4 Asian |
| | 1 Greek, 1 Hispanic | 10 Other |
| Assessment | HAMD-CL: Bi-weekly | HAMD-CL, QIDS-CL: Weeks 1,2,4 |
| Recording Type | Interactive Voice Response Technology (8kHz) | |
| Lengths | Min: 2.5s, Max: 156.8s | Min: 2.6s, Max: 181.2s |

- Audio segments were created by applying a 20s window with 5s shifts and the resultant segments longer than 10s were included
- Text data segmentation: sentences



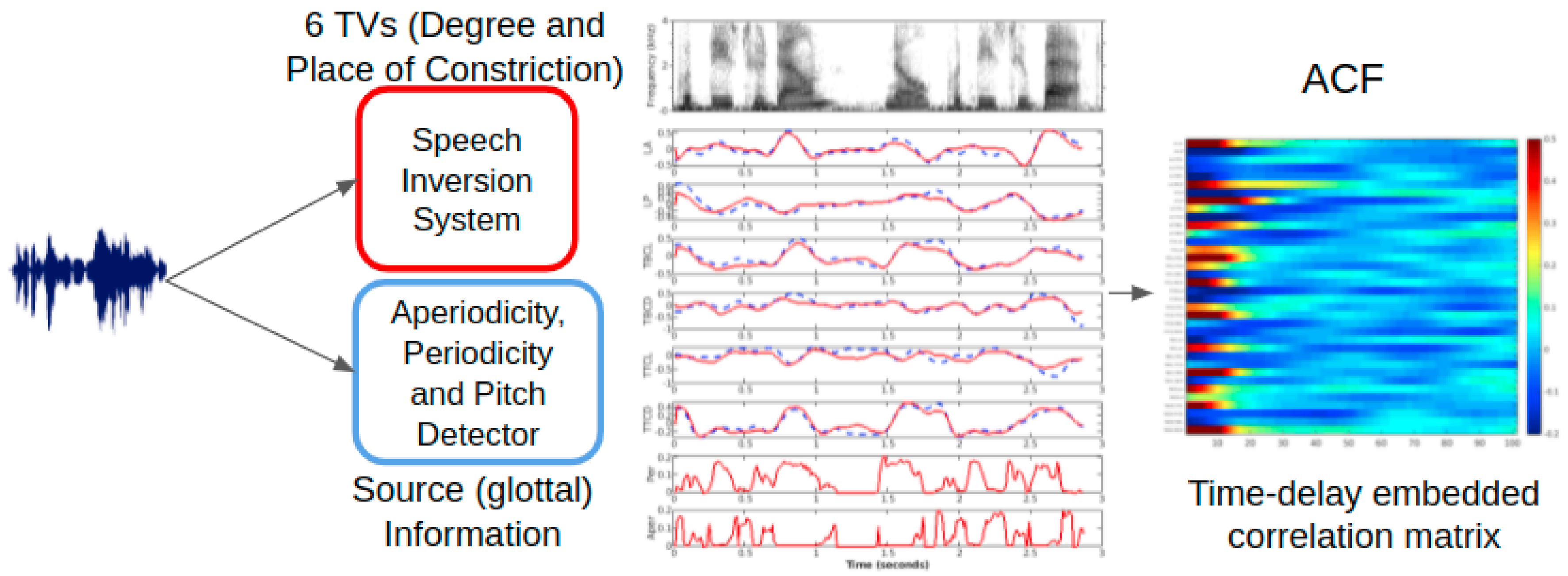

## 3. Vocal Tract Variables (TVs)

- Based on Articulatory Phonology [3,4]

| Constriction Organ | Tract Variable | Articulators |
|---|---|---|
| Lip | Lip Aperture<br>Lip Protrusion | Upper Lip, Lower Lip, Jaw |
| Tongue Body | Tongue body constriction degree<br>Tongue body constriction location | Tongue Body, Jaw |
| Tongue Tip | Tongue tip constriction degree<br>Tongue tip constriction location | Tongue Body, Tip, Jaw |
| Velum | Velum | Velum |
| Glottis | Glottis | Glottis |

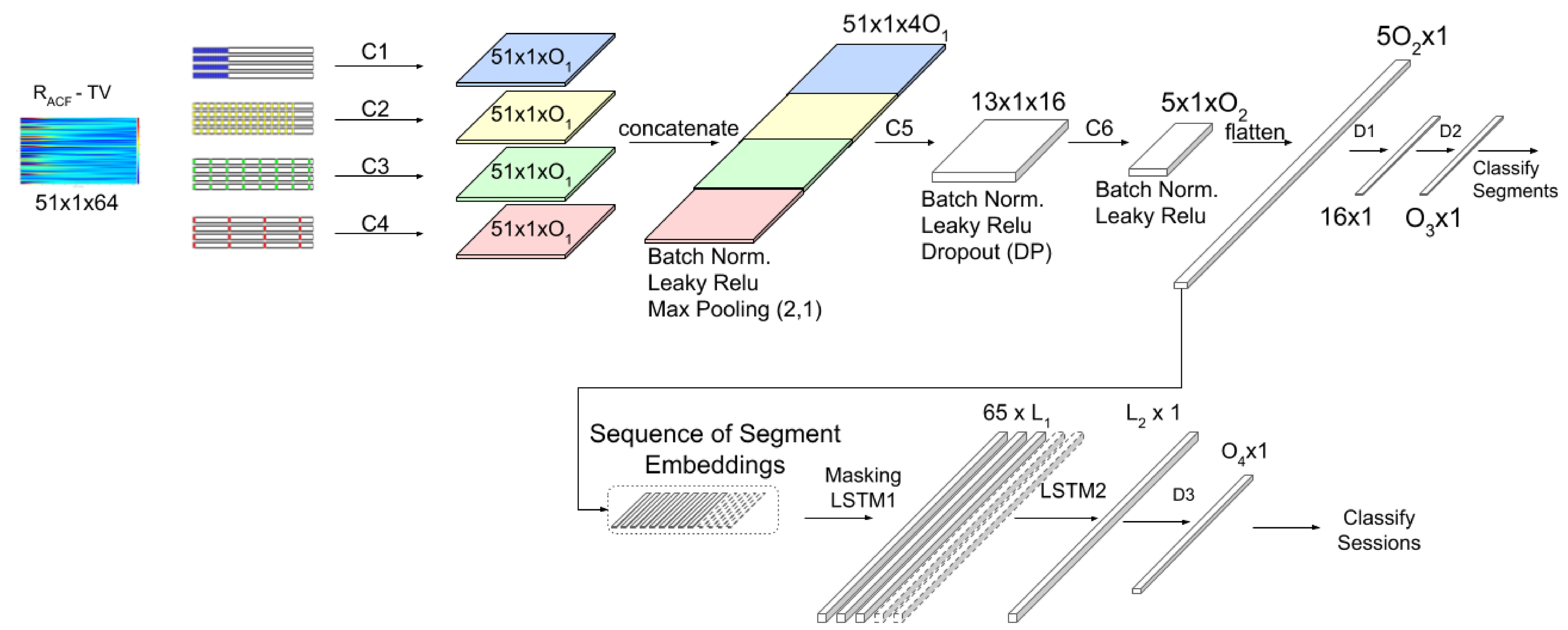
## 4. Articulatory Coordination Features

- Based on Psychomotor Slowing (PMS) [5,6]
  - Altered coordination and timing across articulators
- ACFs proposed by Huang et al. [7] that utilizes dilated CNNs and incorporating more delays to the correlation matrix

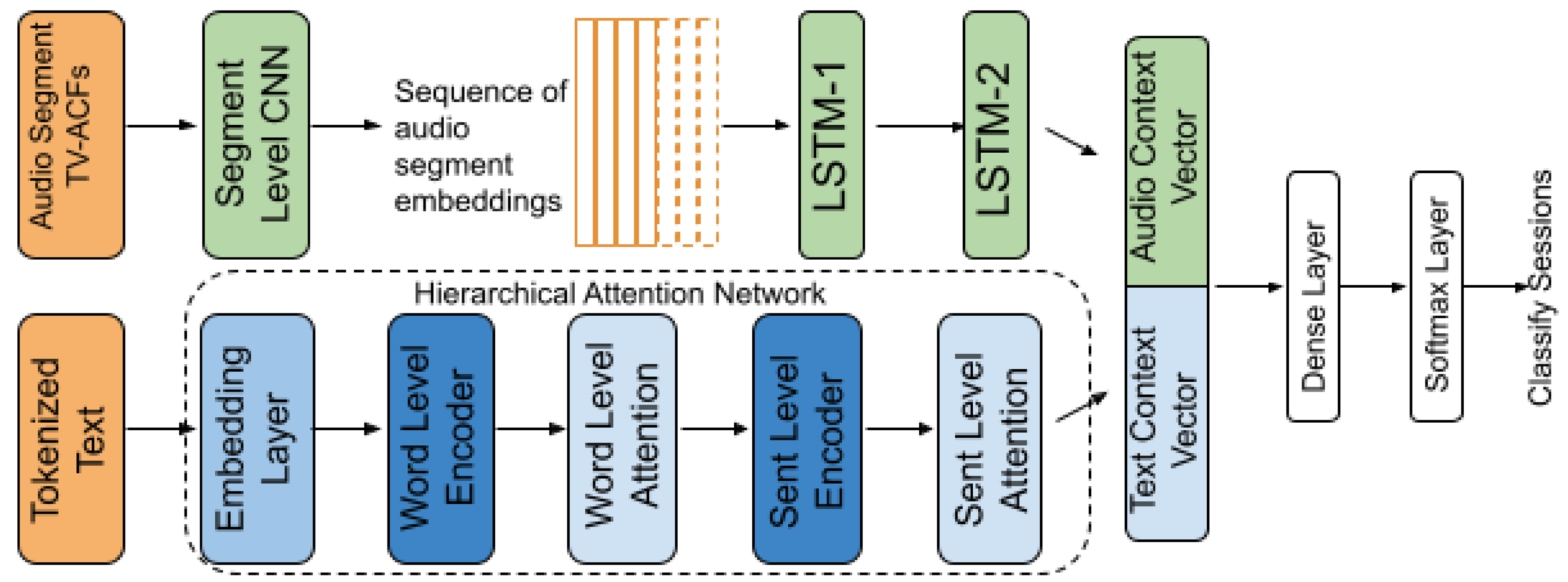



## 5. Model Architectures

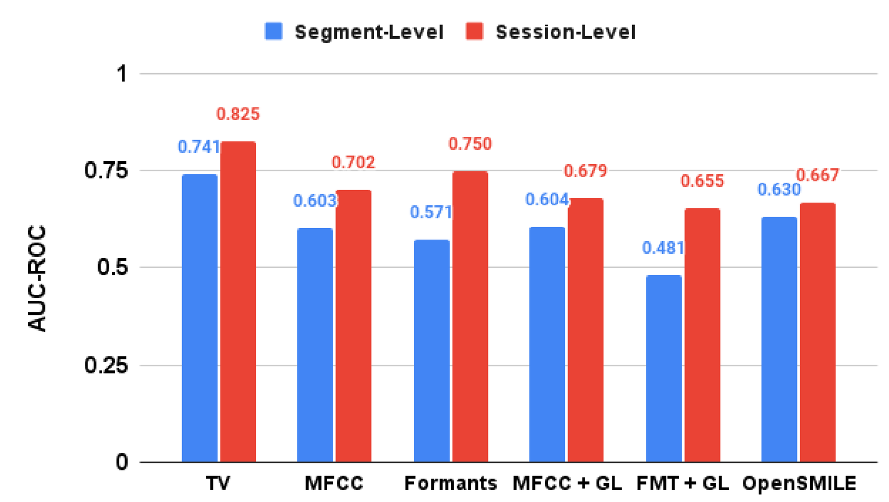
- Unimodal Audio Classifier

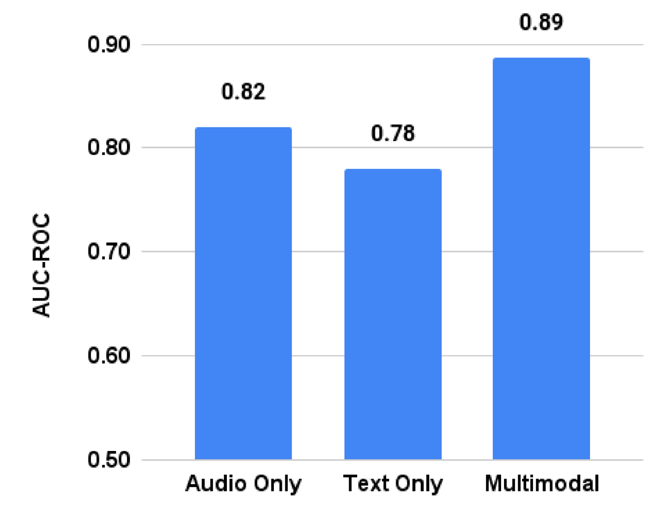



- Multimodal Classifier




## 6. RESULTS




- 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

## 7. ERROR ANALYSIS

- Reasons for errors of the text model
  - **Complex sentence structures (often with mixed sentiments)**
    - "**I'm not feeling guilty** and **feeling like I cannot do anything like before**" [Ground-truth: Not depressed]
  - **Excessive use of negation**
    - "I do not feel like I do not want to do anything" [Ground-truth: Not depressed]
- Misclassified sessions by all models
  - Sentiment not agreeing with the severity score assigned by the clinician which was used to determine the ground-truth label
  - Quasi-numerical nature of HAMD scores leading to ambiguity at the depression severity threshold boundaries

## 8. CONCLUSION

- 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

[1] J. C. Mundt, P. J. Snyder, M. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology," Journal of Neurolinguistics, vol. 20, pp. 50–64, 2007.
[2] Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R.," Vocal acoustic biomarkers of depression severity and treatment response". Biological psychiatry, 72(7), 580–587, 2012.
[3] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," The Journal of the Acoustical Society of America, vol. 146, no. 1, pp. 316–329, jul 2019.
[4] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 776–786, 9 2005.
[5] J. R. Williamson, D. Young, A. A. Nierenberg, J. Niemi, B. S. Helfer, and T. F. Quatieri, "Tracking depression severity from audio and video based on speech articulatory coordination," Computer Speech & Language, vol. 55, pp. 40–56, 2019
[6] Christina Sobin and Harold Sackeim. "Psychomotor symptoms of depression". In: The American journal of psychiatry (1997)
[7] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated CNNS for depression detection in naturalistic environments," in Proc. of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing

## ACKNOWLEDGEMENTS

*This work was supported by the UMCP & UMB Artificial Intelligence + Medicine for High Impact Challenge Award and the National Science Foundation grant numbered 2124270. We thank Dr. James Mundt for the depression databases MD-1&2 [1, 2] and Dr. Thomas Quatieri and Dr. James Williamson for granting access to the MD-2 database which was funded by Pfizer*