Speech-based Multimodal Assessment of Schizophrenia

Gowtham Premananth & Carol Espy-Wilson
Department of Electrical and Computer Engineering
University of Maryland, College Park, USA

Introduction

- The varying symptoms of schizophrenia, like delusions, hallucinations, diminished emotional expression, and poverty of speech, affect speech production.
- Our work focuses on utilizing these changes in speech for the detection of schizophrenia and assessment of symptom severity.
- The complex symptom nature of schizophrenia also affects the linguistic nature of speech and the changes in facial expressions.
- Combining text-based and video-based features with speech-based features has proven to be beneficial in both the detection and assessment of schizophrenia symptoms.

Dataset

- The dataset used in this study was collected at the School of Medicine, University of Maryland contains audio and video recordings of interview sessions.

No.of Subjects	39
No.of Sessions	140
Hours of Speech	34.45 hours
Symptoms based subclasses	Healthy control, Positive-Schizophrenia, Mixed-Schizophrenia
BPRS-based severity score range	19-62

Feature Extraction

Video-based Features: Facial Action Units (FAUs)

- The facial action coding system is a comprehensive anatomically based system that tracks facial movements and converts FAUs.
- Openface 2.0 toolkit was used to extract these FAUs.

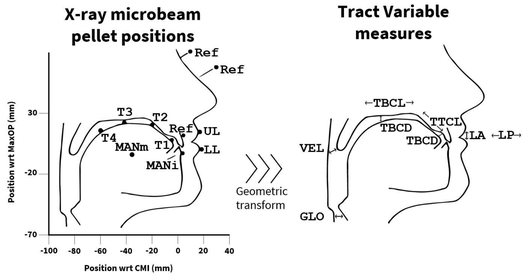
Text-based Features

- Glove Embeddings and BERT embeddings were used as text-based features.
- Glove embeddings produce a fixed-size vector representation for each word in the text.
- BERT embeddings produce both word-level and sentence-level vector representations.

Audio-based Features: Vocal Tract Variables (TVs)

- An Acoustic-to-Articulatory speech inversion system [4] is used to extract the TVs.
- An Aperiodicity, Periodicity, Pitch (APP) detector [5] is used to extract source features to be used as proxy to the Glottis TV.

Constrictor	TV
Lip	Lip Aperture (LA)
	Lip Protrusion (LP)
Tongue Body	Tongue Body Constriction Location (TBCL)
	Tongue Body Constriction Degree (TBCD)
Tongue Tip	Tongue Tip Constriction Location (TTCL)
	Tongue Tip Constriction Degree (TTCD)
Velum	Velum (VEL)
Glottis	Glottis (GLO)



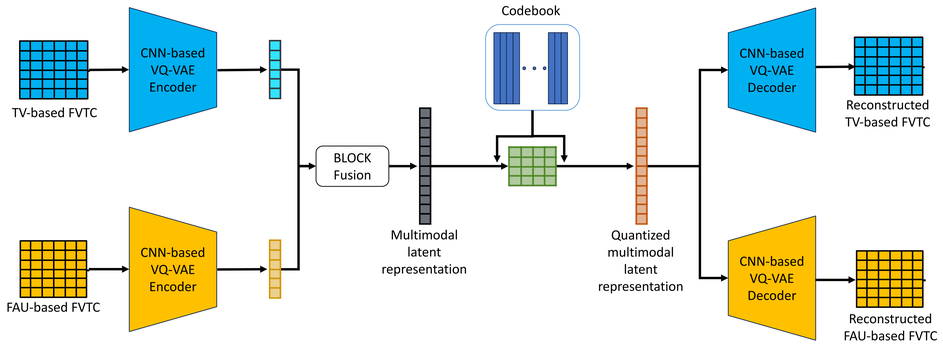
Absolute positions of articulators *dependent* on speaker dimensions

Relative measures of constrictions *independent* of speaker dimensions

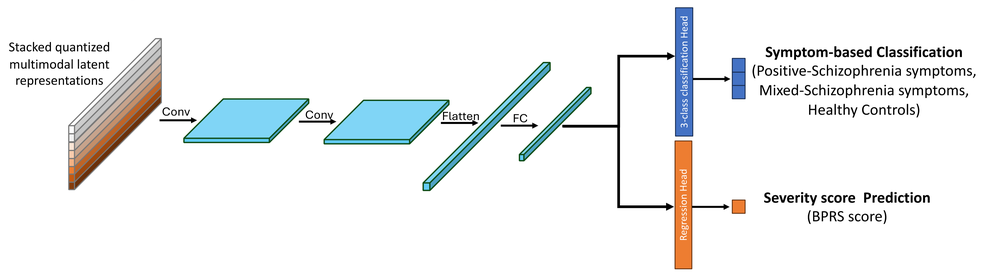
Coordination Features

- The FAUs and TVs are used to calculate FAU-based and TV-based coordination features using a channel-delay correlation mechanism [6].

Multimodal representation learning framework



Multi-Task Learning (MTL) Framework for Schizophrenia Assessment



Results

Modalities	Model	Classification			Severity Estimation
		Acc.	F1	AUC_ROC	MAE
A,V	CNN-LSTM model with Attention [1]	52.17	49.33	68.53	-
A, V, T	CNN-LSTM model with Attention [1]	56.52	51.62	74.21	-
A,V	CNN-LSTM with mGMU [2]	60.87	60.28	77.35	-
A,V,T	CNN-LSTM with mGMU [2]	65.22	65.47	82.14	-
A,V	MM-VQ-VAE representations based classification [3]	54.96	71.04	57.62	-
A,V	MM-VQ-VAE representations based regression [3]	-	-	-	8.81
A,V	MM-VQ-VAE representations based MTL model [3]	**75.00**	**76.41**	**91.52**	**7.19**

Acoustic and Articulatory Feature Fusion

- Previous works on speech-based mental health assessments have only either used acoustic features or articulatory features as inputs.
- In our latest work [7], we have developed a speech-based schizophrenia assessment system that fuses self-supervised speech features (WavLM) and TV-based coordination features and provides a performance improvement.

Conclusion and Future Work

- Combining the MTL paradigm with TV and FAU-based multimodal representations enhances the performance in symptom-based classification.
- Our model with the MTL paradigm provides improved performance in symptom severity estimation when compared with a standalone regression model.
- Data scarcity limits multimodal representation learning to a single dataset. Therefore, we plan to integrate diverse data sources for more generalizable representation learning models.
- The fusion of articulatory and acoustic features provides a performance improvement in speech-only settings.
- We plan to expand our work to assess individual symptom severity across the schizophrenia spectrum.

References

[1] G. Premananth, Y. M. Siriwardena, P. Resnik, & C. Y. Espy-Wilson, 'A multi-modal approach for identifying schizophrenia using cross-modal attention', ArXiv, vol. abs/2309.15136, 2023.
[2] G. Premananth, Y. M. Siriwardena, P. Resnik, S. Bansal, D. L.Kelly, & C. Espy-Wilson, 'A Multimodal Framework for the Assessment of the Schizophrenia Spectrum', Interspeech 2024, 2024
[3] G. Premananth and C. Espy-Wilson, "Self-supervised Multimodal Speech Representations for the Assessment of Schizophrenia Symptoms," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025
[4] Y. M. Siriwardena and C. Y. Espy-Wilson, 'The Secret Source: Incorporating Source Features to Improve Acoustic-To-Articulatory Speech Inversion', ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2022.
[5] Deshmukh, O., Espy-Wilson, C., Salomon, A. & Singh, J. Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech. IEEE Transactions On Speech And Audio Processing. 13, 776-786 (2005)
[6] Huang, Z., Epps, J. & Joachim, D. Exploiting Vocal Tract Coordination Using Dilated CNNS For Depression Detection In Naturalistic Environments. ICASSP 2020. pp. 6549-6553 (2020)
[7] Premananth, G., & Espy-Wilson, C. (2024). Speech-Based Estimation of Schizophrenia Severity Using Feature Fusion. arXiv [Eess.AS]. Retrieved from http://arxiv.org/abs/2411.06033

Acknowledgment

This work was supported by the National Science Foundation grant numbered 2124270.