Abstract ID :3636662
Session :5pSC

Emotion Recognition with Speech Articulatory Coordination Features

Yashish M. Siriwardena, Nadee Seneviratne and Carol Espy-Wilson
University of Maryland College Park, MD, USA

Introduction

- Mental health is about the functioning of the brain. It involves the processing of all of the information we encounter.
- Emotion is more about the feelings provoked by the information we processed.
- Changes in mental health affect neuromotor processes. For example, it is well known that major depressive disorder (MDD) results in psychomotor slowing which affects speech, ideation and motility.
- In previous work, we have studied the sensitivity of the Articulatory Coordination Features (ACFs) to changes in neuromotor processing caused by depression (MDD) and schizophrenia (SZ).
- In this study, we investigate whether ACFs are also predictive of emotions.

Previous Work on MDD and Schizophrenia

Figure 1 shows the steps needed to derive articulatory coordination features

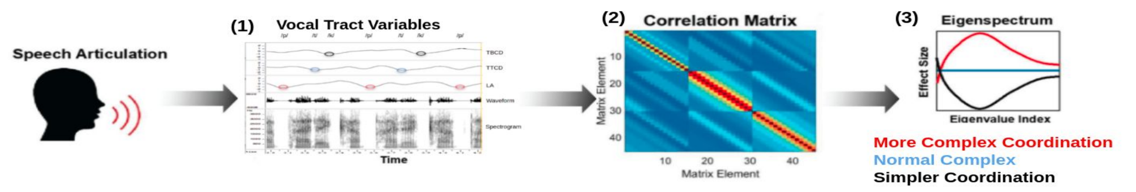


Figure1: Overview of Articulatory Coordination Features

(1) Vocal Tract Variables (TVs)
- Based on Articulatory Phonology

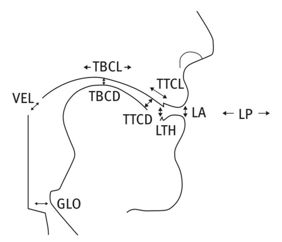


Table 1: List of vocal tract variables (TVs)

Constriction Organ	Tract Variable	Articulators
Lip	Lip Aperture (LA) Lip Protrusion (LP)	Upper Lip, Lower Lip, Jaw
Tongue Body	Tongue body constriction degree (TBCD) Tongue body constriction location (TBCL)	Tongue Body, Jaw
Tongue Tip	Tongue tip constriction degree (TTCD) Tongue tip constriction location (TTCL)	Tongue Body, Tip, Jaw
Velum	Velum (VEL)	Velum
Glottis	Glottis (GLO)	Glottis

DNN Based Speech Inversion System [1]

APP Detector (Aperiodicity /Periodicity) [2]

(2) Articulatory Coordination Features (ACFs)
- A channel-delay correlation matrix is computed from feature vectors at a specified delay scale (1,3,7,15)
- Previously computed using proxies for underlying articulatory coordination (formants, MFCCs etc)

$$R_j = \begin{bmatrix} \begin{bmatrix} r_{1,1}(j) & \dots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \dots & r_{N,N}(j) \end{bmatrix}_{1,1} & \cdots & \begin{bmatrix} r_{1,1}(j) & \dots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \dots & r_{N,N}(j) \end{bmatrix}_{1,M} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} r_{1,1}(j) & \dots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \dots & r_{N,N}(j) \end{bmatrix}_{M,1} & \cdots & \begin{bmatrix} r_{1,1}(j) & \dots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \dots & r_{N,N}(j) \end{bmatrix}_{M,M} \end{bmatrix}$$

(3) Eigenspectra Computation from ACFs
- Rank-ordered eigenspectra from correlation matrix
- Magnitude of eigenvalues represent the average correlation in the direction of corresponding eigenvectors

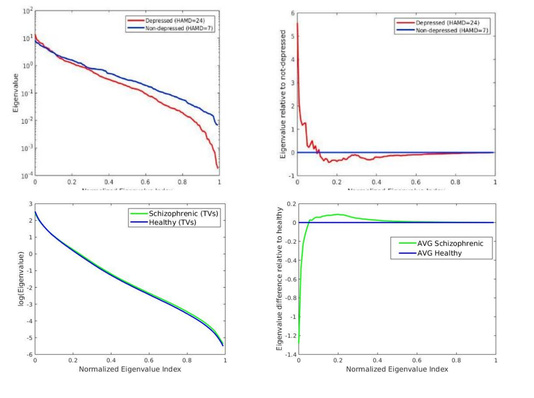


Figure 2: Eigenspectra and difference plots for MDD (top row) and SZ (bottom row)

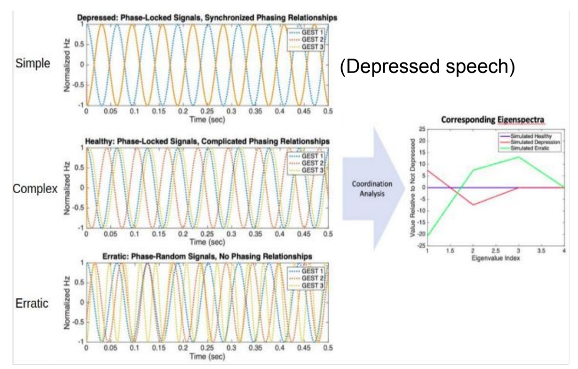


Figure 3: Simulated Eigenspectra to Different Coordination Patterns [3]

Classification of MDD and SZ using Dilated CNNs

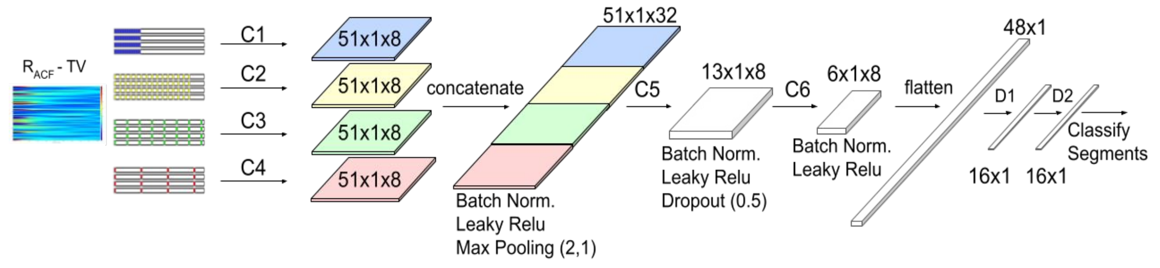


Figure 4: Unimodal architecture for ACF classification

MDD Classification Results (Session-Wise)[4]

Features	AUC ROC	UWA
TV ACFs	0.8246	0.8024
MFCC ACFs	0.7016	0.6451
Formant ACFs	0.75	0.7238
eGeMAPS	0.6673	0.6612

SZ Classification Results (Subject wise) [5]

Features	F1(SZ)/F1(H)	UWA
TV ACFs	0.62/0.78	0.7222
MFCC ACFs	0.55/0.80	0.7222
FAU ACFs	0.77/0.89	0.8333
eGeMAPS	0.54/0.78	0.7012

Simpler coordination due to depression results in less coarticulation and undershoot

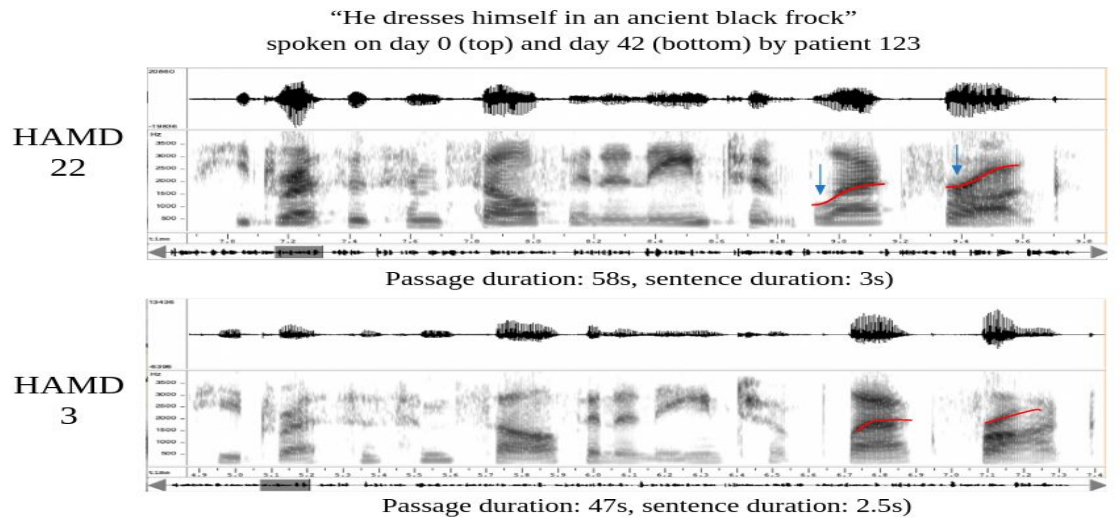


Figure 5: Words "black frock" from the read speech of an utterance from the Grandfather passage for a subject when depressed and in remission

Current Study on Emotion Recognition using TV ACFs

(1) **Database**
- IEMOCAP [6] database
  -Consists of 5 sessions. In each session, two actors act out scenarios (scripted or improvised)
  -Used 5530 utterances (~7 hours of speech) : neutral (1708), angry (1103), sad (1083), and happy (1636) which have majority agreement of annotators

(2) **Feature Extraction**
- Acoustic features
  - 6 TVs and 2 glottal TVs from Table 1
  - 12 Mel Frequency Cepstral Coefficients (MFCCs)
  - 3 Formants
  - 23 Geneva Minimalistic Acoustic Parameter Set (eGeMAPS taken from the OpenSmile feature set)
- Text features
  - 100 dimensional GloVe embeddings generated from the IEMOCAP transcriptions

Eigenspectra Analysis on IEMOCAP

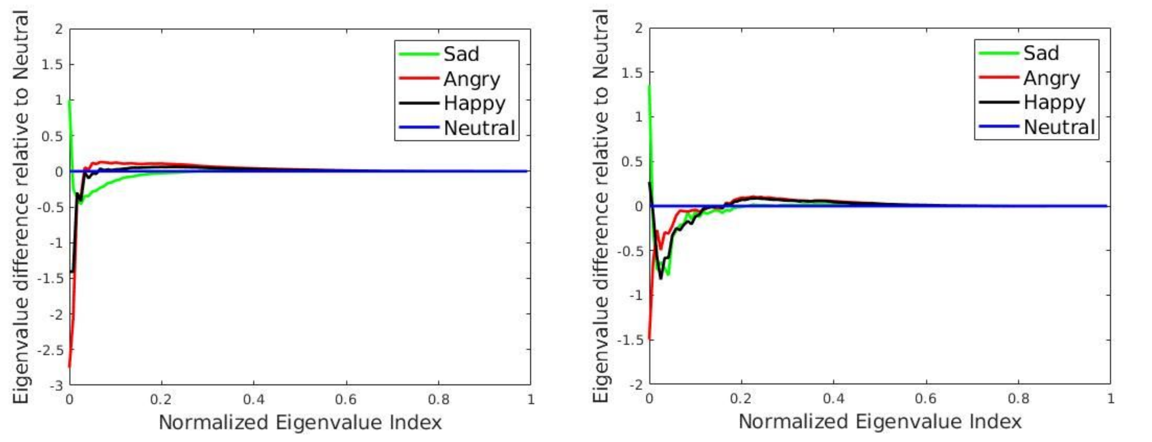
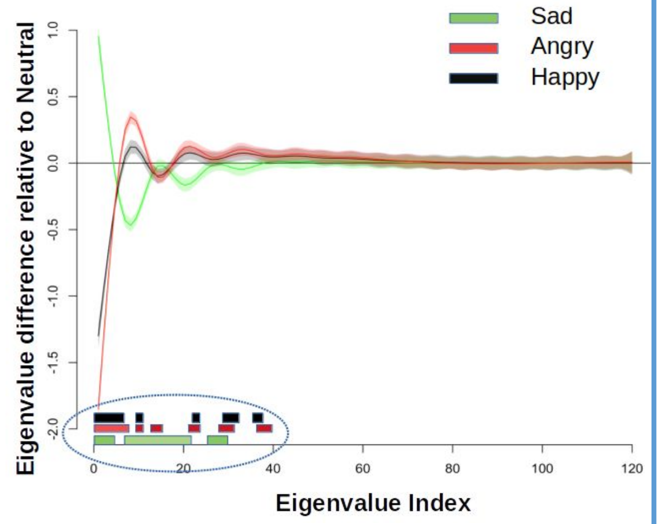


Figure 6: Difference plot computed from TVs for IEMOCAP for the the 5530 utterances with majority agreement among annotators (left) and for 2032 utterances with all annotator agreement (right)

- The emotion "Sad" shows a simpler articulatory coordination pattern relative to "Neutral" (like MDD)
- The emotion "Angry" shows a more complex articulatory coordination relative to "Neutral" (like SZ)
- These results motivated us to check if articulatory coordination features derived from TVs can be used for effective emotion recognition

Statistical significance of eigenspectra patterns

- Generalized Additive Mixed Models (GAMMs) to compare the resulting averaged eigenspectra across all subjects
- The test shows statistical significance for the following difference patterns at marked low rank eigenvalue ranges



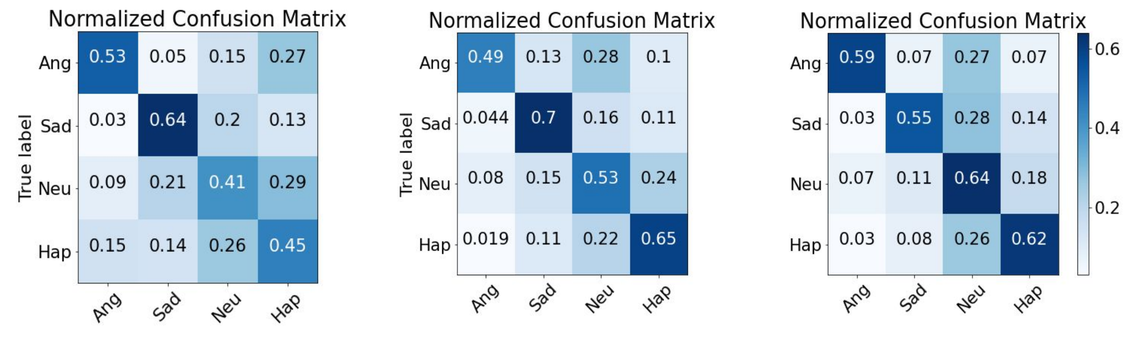
Emotion recognition on IEMOCAP data

- Speaker-based normalization to reduce speaker specific effects using only the neutral speech [7]
- Leave one session out cross validation

Deep Learning Based Models

(1) Unimodal systems (IEMOCAP)
- TVs Model : Dilated CNN [8]
- eGeMAPS model :LSTM model with two LSTM layers

Model	TVs	eGeMAPS	GloVe
UWA (%)	50.81	55.80	61.16



TVs Model | eGeMAPS Model | GloVe Model

(2) Multimodal systems (IEMOCAP)

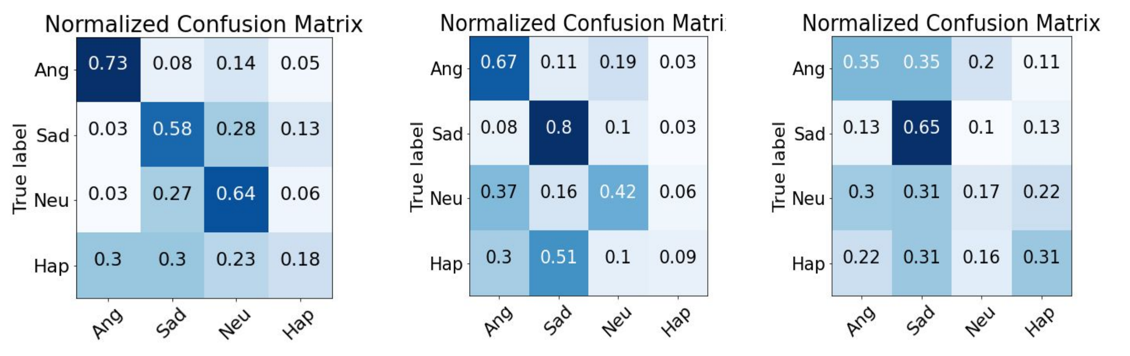
Model	TVs+Glove	eGeMAPS+GloVe
UWA (%)	62.70	68.18 [9]

Analysis of Results by TV Model: Over half of the utterances have more than one speaker and many times the two speakers have different emotions Thus, using the ACFs for these utterances is problematic. Further, we wanted to see how well the system would work if we only considered the data where all annotators agreed on the emotion expressed (2032 utterances).

Reduced IEMOCAP dataset: To address the problem of multi-speaker utterances, we used speaker diarization. Prior to diarization, we excluded significantly silent utterances. This reduced the number of utterances from 2032 to around 1000. For this study, we listened to 500 in the reduced set and found that 145 of the utterances contained only a single speaker. Given this substantially reduced data set, we could only use a simpler classifier, a Support Vector Machine, and a smaller set of features (averaged eigenspectra) for experiments.

Support Vector Machine (SVM) models
- Eigenspectral features averaged across different regions (eigen indices) as features for the SVM.

Model	TVs	MFCCs	Formants
Averaged ranges	[0.0, 0.05], [0.05, 0.39], [0.39, 0.56], [0.56, 1.0]	[0.0, 0.07], [0.07, 0.09], [0.09, 1.0]	[0.0, 0.29], [0.29, 0.38], [0.38, 0.43], [0.43, 1.0]
UWA (%)	52.95	49.41	37.02



TVs Model | MFCCs Model | Formants Model

Discussion and Future work

- The eigenspectra pattern for "Sad" relative to "Neutral" is similar to that for "Depressed" relative to "Remission". Likewise, the eigenspectra pattern for "Angry" relative to "Neutral" is similar to "Schizophrenia" relative to "Healthy Controls".
- TV-based SVM model does better emotion recognition compared to MFCCs- and formant-based models
- "Clean" data (one speaker) has a significant impact on the effectiveness of ACFs. Further experiments with a larger emotion dataset and DNN-based classification models are planned.
- A multimodal system using ACFs and natural language processing gives better performance than unimodal systems.

References

[1] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, & C. Espy-Wilson, "Vocal tract length normalization for speaker
[2] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh. 2005. Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech. IEEE Transactions on Speech and Audio Processing 13, 5 (2005), 776–786.
[3] Adam C. Lammert,James Williamson,Nadee Seneviratne,Carol Espy-Wilson and Thomas F. Quatieri, A Coupled Oscillator Planning Account of the Speech Articulatory Coordination Metric With Applications to Disordered Speech in ISSP 2021
[4] Seneviratne, N., Espy-Wilson, C. (2021) Generalized Dilated CNN Models for Depression Detection Using Inverted Vocal Tract Variables. Proc. Interspeech 2021, 4513-4517, doi: 10.21437/Interspeech.2021-1960
[5] Yashish M. Siriwardena, Carol Espy-Wilson, Chris Kitchen, and Deanna L. Kelly. 2021. Multimodal Approach for Assessing Neuromotor Coordination in Schizophrenia Using Convolutional Neural Networks. In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), ACM
[6] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, 2008
[7] C. Busso, A. Metallinou, and S. S. Narayanan, "Iterative feature normalization for emotional speech detection," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011, pp. 5692–5695
[8] Zhaocheng Huang, J. Epps, and D. Joachim. 2020. Exploiting Vocal Tract Coordination Using Dilated CNNS For Depression Detection In Naturalistic Environments. ICASSP 2020
[9] Sahu, S., Mitra, V., Seneviratne, N., & Espy-Wilson, C.Y. (2019). Multi-Modal Learning for Speech Emotion Recognition: An Analysis and Comparison of ASR Outputs with Ground Truth Transcription. INTERSPEECH.

Acknowledgements

This research was supported by the NSF grant 2124270