Multi-modal Biomarkers to Discriminate Cognitive State*

Thomas F. Quatieri [1], James R. Williamson[1], Christopher J. Smalt[1],

Joey Perricone, Tejash Patel, Laura Brattain, Brian S. Helfer, Daryush D. Mehta, Jeffrey Palmer

Kristin Heaton[2], Marianna Eddy[3], Joseph Moran[3]

[1]MIT Lincoln Laboratory, Lexington, Massachusetts, USA
[2]USARIEM, [3]NSRDEC
[quatieri,jrw]@ll.mit.edu

1. Introduction

Multimodal biomarkers based on behavorial, neurophysiolgical, and cognitive measurements have recently obtained increasing popularity in the detection of cognitive stress- and neurological-based disorders. Such conditions are significantly and adversely affecting human performance and quality of life for a large fraction of the world's population. Example modalities used in detection of these conditions include voice, facial expression, physiology, eye tracking, gait, and EEG analysis. Toward the goal of finding simple, noninvasive means to detect, predict and monitor cognitive stress and neurological conditions, MIT Lincoln Laboratory is developing biomarkers that satisfy three criteria. First, we seek biomarkers that reflect core components of cognitive status such as working memory capacity, processing speed, attention, and arousal. Second, and as importantly, we seek biomarkers that reflect timing and coordination relations both within components of each modality and across different modalities. This is based on the hypothesis that neural coordination across different parts of the brain is essential in cognition (Figure 1). An example of timing and coordination within a modality is the set of finely timed and synchronized physiological components of speech production, while an example of coordination across modalities is the timing and synchrony that occurs across speech and facial expression while speaking. Third, we seek multimodal biomarkers that contribute in a complementary fashion under various channel and background conditions. In this chapter, as an illustration of this biomarker approach we focus on cognitive stress and the particular case of detecting different cognitive load levels. We also briefly show how similar feature-extraction principles can be applied to a neurological condition through the example of major depression disorder (MDD). MDD is one of several neurological disorders where multi-modal biomarkers based on principles of timing and coordination are important for detection [11]-[22]. In our cognitive load experiments, we use two easily obtained noninvasive modalities, voice and face, and show how these two modalities can be fused to produce results on par with more invasive, "gold-standard" EEG measurements. Vocal and facial biomarkers will also be used in our MDD case study. In both application areas we focus on timing and coordination relations within the components of each modality.

* Distribution A: public release.This work is sponsored by the Assistant Secretary of Defense for Research & Engineering under Air Force contract #FA8721-05-C-0002. Opinions,interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

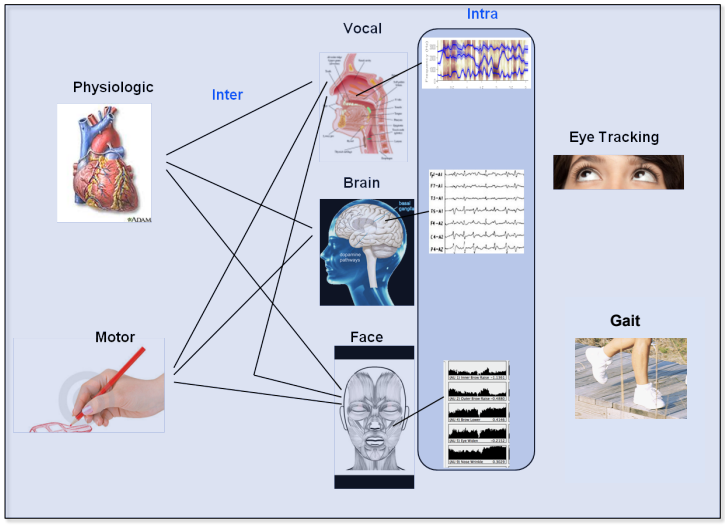

Figure 1: Biomarkers reflect elements of cognitive status as well as timing and coordination within and across modalities.

The ease of obtaining vocal and facial features (e.g., via mobile tablets or smartphones) greatly increases global accessibility to an automated method for cognitive assessment. Certain vocal and facial features have been shown to change with a subject's mental and emotional state, under numerous conditions including cognitive load and neurological conditions. For voice, these features include characterizations of prosody (e.g., fundamental frequency and speaking rate), spectral representations (e.g., mel-cepstra), and glottal excitation flow patterns, such as flow shape, timing jitter, amplitude shimmer, and aspiration [23]-[35]. For facial features, these include spectral representations and facial action units [36][37]. There are many examples of each of the three modalities being used to detect cognitive stress [1]-[10] and of voice or face modalities being used to detect a variety of neurological conditions such as depression and Parkinson's disease [23]-[35]. Voice has been used in cognitive load by Yin et al [2] who achieved 77% accuracy using standard vocal features (e.g., mel-cepstra, delta-delta mel-cepstra, and shifted mel-cepstra) to discriminate three cognitive load levels in a read story (and through several questions about the story), and in the Stroop test. Facial action units [53][54] have been used to predict neuropsychiatric disorders, while EEG entropy and power have been used to discriminate multiple cognitive load levels by Zarjam et al [9][10].

In this chapter, for all modalities, we use a common approach that deviates from the standard one. While we begin with standard "low-level" features that are used in the approaches listed above, we build upon these using "high-level" timing and coordination features. For voice, the low-level features are phoeneme boundaries, formant (vocal tract resonance) tracks, delta mel-cepstra coefficients, and creakiness (vocal-fold irregularity). For face, the low-level features are automatically extracted facial action units (FAUs) [59]. For EEG, the low-level features rely on spectral power from EEG channels following standard artifact removal. The high-level timing features include (from voice) phoneme-based measures of rate, duration, pitch dynamics, and pause information, and (from face) FAU-based measures of rate information. The high-level coordination features for all modalities are based on eigenspectra analysis of covariance, correlation, and coherence matrices that are constructed from sets of low-level features. Various subsets of these features have been used effectively at MIT Lincoln Laborarory in cogntive stress [17] and neuro-cognitive contexts such as in detection of depression, Parkinson's disease, traumatic brain injury, and dementia [11]-[16],[18]-[22], thus perhaps forming a common feature basis for neurocognitive change.

In this chapter, detection of cognitive load is used as a case study. However, the algorithms described provide a more general framework for detection of changes in neurocognitive health status from multiple sensing modalities. In Section 2, we open with a novel cognitive load data collection protocol that taxes auditory working memory by eliciting sentence recall under varying levels of cognitive load. In Section 3, we then describe how our signal processing methodology for vocal, facial, and EEG features are applied for detection of cognitive load under this protocol. Section 4 summarizes cognitive load detection results using a Gaussian classifier. Section 5 then shows how we can use our principles of timing and coordination to detect major depression disorder from vocal and facial signals. Lastly, Section 6 closes with conclusions and projections toward future work.

2. Design of a Multimodal Platform for Cognitive Load

Cognitive load is defined loosely as the mental demand experienced for a particular task. Demand can increase or decrease depending on the task and the degree of working memory required [1][2]. Efficient and effective methods are needed to monitor cognitive load under cognitively and physically stressful situations. In many scenarios, environmental and occupational stressors can produce cognitive overload, thereby degrading task performance and endangering safety. Examples of mental stressors are repetitive and/or intense cognitive tasks, psychological stress, and lack of sleep. Physical stressors include intense long-duration operations and/or heavy loads. Both stressors can cause cognitive load, and often contribute simultaneously to load. Applications for cognitive load assessment include individualized detection of cognitive load in an ambulatory, field, or clinical setting. In clinical applications, the objective is often to find and measure the specific causes of load. In operational settings, the objective is often to quickly assess cognitive ability and readiness under loaded conditions, regardless of their etiology. In designing a multimodal voice/face/EEG database protocol that reflects these typical cognitive load conditions, we employed the hypothesis that speech, and the corresponding facial movements that occur while speaking, are complex motor activities requiring precise neural timing and coordination, and that manipulating cognitive load level systematically alters this complex motor activity in a measurable way.

Subjects gave informed consent to our working memory-based protocol approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). Audio data are collected with a DPA acoustic lapel microphone (with a Roland Octa-Capture audio interface), facial video with a Canon high-definition video camera, and EEG signals with a 64-element Neuroscan device. An illustration of the collection platform is shown in Figure 2.

Figure 2: Multimodal platform for recording of speech, facial expression, EEG, and physiological signals.

Following setup and training, each subject engages in the primary task of verbally recalling sentences with varying levels of cognitive load, as determined by the number of digits being held in working memory [56]-[58]. Specifically, a single trial of the auditory working memory task comprises: the subject hearing a string of digits, then hearing a sentence, then waiting for a tone eliciting spoken recall of the sentence, followed by another tone eliciting recall of the digits. This task is administered with three difficulty levels, involving 108 trials per level. The same set of 108 sentences is used in each difficulty level. The order of trials (sentences and difficulty level) is randomized. The entire protocol, approximately two hours in duration, is illustrated in Figure 3. The multi-talker PRESTO sentence database is used for sentence stimuli [55]. We recorded 17 subjects but used 11 subjects from whom robust recordings were obtained in all three modalities.

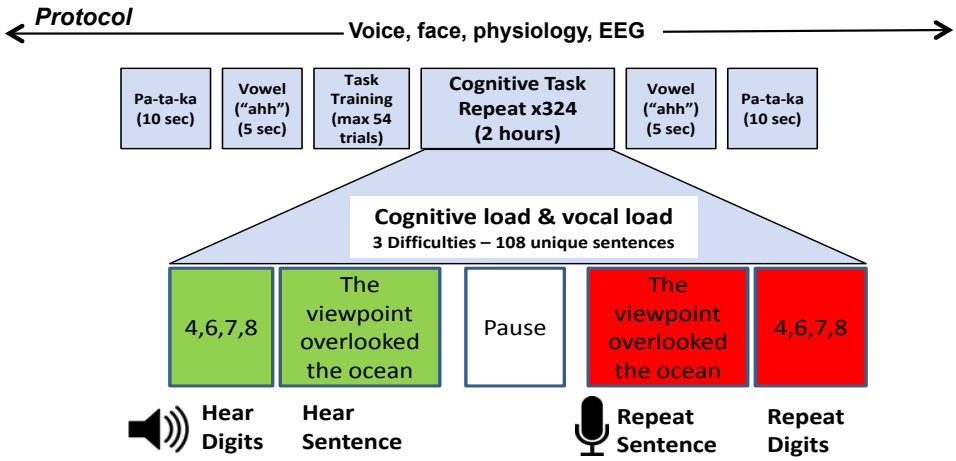

Figure 3. Auditory working memory protocol. Audio and video are analyzed during sentence recall while the EEG is analyzed during the pause interval to avoid motion and muscle artifacts.

The working memory task is split into a training and a testing phase. During training, the maximum number of digits that a subject can accurately recall is estimated using an adaptive tracking algorithm [56]. This number n_c, is used to determine the three difficulty levels in the test phase, which were typically set as: d_n = {ceil(n_c), ceil(n_c)-1, ceil(n_c)-2}. Despite some minor protocol changes among early subjects, this common load assessment test was used for most subjects, and used in 10 of the 11 subjects analyzed in this chapter. We later will define a binary detection problem of discriminating *high load* (max number) from low load (max number minus two). The range of digit spans across all subjects was 2–5 for low load and 4–7 for high load.

Finally, as seen in Figure 3, with our protocol we also measure skin conductance, temperature, and pulse oxygenation level. These are not a focus of the present study but pave the way to investigating a broader suite of mult-modal biomarkers.

3. Feature Extraction

Biomarkers for monitoring and detecting cognitive stresses, as well as neurological disorders, comprise behavioral, physiologic, and cognitive modalities. Three particular modalities that are gaining popularity include speech, facial expression, and EEG signals. The effectiveness of predicting cognitive load from these modalities is based on the hypothesis that manipulating cognitive load level systematically alters the underlying neural motor activation required in speech production and facial expression due to a competition for mental resources between motor activity

and working memory. This neural activation is reflected in the EEG measurements which is sometimes considered a "gold standard" in viewing the effect of working memory demand [8]-[10].

Our features from these three modalities (voice, face, EEG) are based on common principles of timing and coordination within components of each modality. In each case, we extract first *low-level* (standard) features followed by *high-level* (novel) features as functions of the low-level features. For voice, feature vectors are extracted only from the single spoken sentence component of each trial in the test phase of the auditory memory task of Section 2. Low-level vocal features comprise measures of phoneme and pseudosyllable durations, pitch dynamics, spectral (formant) dynamics, and vocal-fold irregularity (creak). We construct high-level features that capture timing and inter-relationships across the low-level features. The feature sets are derived under the hypothesis that differences in cognitive load produce detectable changes in speech production timing and coordination within and across articulatory and vocal fold components. For facial expression, analyzed during the same time interval as audio, the extracted low-level features are Facial Action Units [36][37][59], which are followed by correlation-based measures as high-level features. For the EEG, during the pause interval to avoid motion and muscle artifacts**,** we perform preprocessing to extract low-level EEG signals free of many typical artifacts, followed by correlation and frequency-dependent coherence and power measures. In this section, we describe key details of the various extraction methods and illustrate extraction through the cognitive load database decribed in Section 2.

3.1 Vocal Features

We exploit dynamic variation and inter-relationships across speech production systems by computing features that reflect complementary aspects of the speech vocal-fold source, vocal tract system, and prosody [75]. We describe a broad suite of features which are used in detection of cognitive load in Section 4, and in detection of depression in Section 5, as well as in other cognitive stress and neurological disorders [11]-[22].

3.1.1 Low-level vocal feature extraction

In this section we introduce a set of low-level features some of which are used in the cognitive load scenario of this chapter, while others are used in the depression example of Section 5, as well as in our other cognitive load and neurological detection efforts [11]-[22].

Voice Source

Harmonics-to-noise ratio (HNR): A spectral measure of harmonics-to-noise ratio was performed using a periodic/noise decomposition method that employs a comb filter to extract the harmonic component of a signal [60][61]. The harmonics-to-noise ratio is the ratio, in dB, of the power of the decomposed harmonic signal and the power of the decomposed speech noise signal and was computed every 10 ms.

Cepstral peak prominence (CPP): Several studies have reported strong correlations between cepstral peak prominence (CPP) and overall dysphonia perception [62]-[64], breathiness [65]-[66], and vocal fold kinematics. CPP is defined as the difference, in dB, between the magnitude of the highest peak and the noise floor in the power cepstrum for quefrencies greater than 2 ms (corresponding to a range minimally affected by vocal tract–related information) and was computed every 10 ms.

Creak voice quality: A creaky voice quality (vocal fry, irregular pitch periods, glottalization, etc.), is characterized using acoustic measures of low-frequency/damped glottal pulses [71]. The creak measure builds on metrics of short-term power, intra-frame periodicity, inter-pulse similarity [72], and two measures of the degree of sub-harmonic energy (reflecting the presence of secondary glottal pulses) and the temporal peakiness of glottal pulses [73]. These values are input into an

artificial neural network to yield creak posterior probabilities on a frame-by-frame basis every 10 ms [74].

Speech System

Formant frequencies: A Kalman filter technique is used to characterize vocal tract resonance dynamics by smoothly tracking the first three formant (resonant) requencies, while also smoothly coasting through non-speech regions [70].

Mel-frequency cepstral coefficients (MFCCs): 16 delta MFCCs [75] are used to characterize velocities of vocal tract spectral magnitudes, typical in speech-related recognition applications [77]. Delta MFCCs [75] are computed using regression with the two frames before and after a given frame.

Speech Prosody

Phonemes: Using an automatic phoneme recognition algorithm [38], phonetic boundaries are detected, with each segment labeled with one of 40 phonetic speech classes (see Figure 17 in Section 3.1.2).

Pitch: The fundamental frequency (pitch) was estimated using a time-domain autocorrelation method over 40 ms Hanning windows every 1 ms [75].

3.1.2 High-level vocal feature extraction

Our high-level features are designed to characterize properties of timing and coordination from the low-level features.

Correlation Structure: Measures of the structure of correlations among low-level speech features have been applied in the estimation of depression [13][22] and Parkinson's disease [19], the estimation of cognitive performance associated with dementia [18][11], the detection of changes in cognitive performance associated with mild traumatic brain injury [14], and in our earlier cognitive load effort [17]. The details for this approach are in [20], where the method was first introduced for analysis of EEG signals for epileptic seizure prediction.

Channel-delay correlation and covariance matrices are computed from multiple time series channels of vocal parameters. Each matrix contains correlation or covariance coefficients between the channels at multiple time delays. Changes over time in the coupling strengths among the channel signals cause changes in the eigenvalue spectra of the channel-delay matrices. The matrices are computed at multiple "time scales" corresponding to separate sub-frame spacings. Features at each time scale consist of the eigenvalue spectra of channel-delay correlation matrices, as well as covariance power (logarithm of the trace) and entropy (logarithm of the determinant) from channel-delay covariance matrices. This methodology is illustrated in Figure 4 with the generation of formant track correlation matrices.

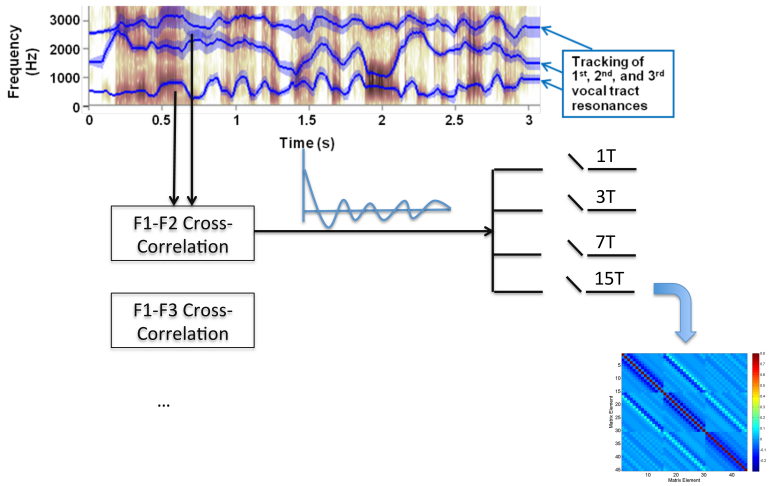

Figure 4. Diagram of cross-correlation analysis of articulatory coordination, as performed through formant-based features using channel-delay correlation matrices at multiple delay scales. A channel-delay matrix from one scale is shown.

In the cognitive load application, parameters were used to extract correlation structure features from three different low-level speech sources: formant frequency tracks, creak probabilities, and delta MFCCs. Sub-frame spacings of 1, 3, and 7 are used and, due to the 10-ms frame interval of the low-level features, these correspond to time spacings of 10, 30, and 70 ms, respectively. Each matrix (for each scale) is constructed using 15 time delays. The number of correlation-based features is the number of signal channels times the number of scales (i.e., number of sub-frame spacings) times the number of time delays (15) per time scale. The number of covariance-based features is the number of time scales (entropy features) plus one log power feature, as power is invariant across scale. Parameters are similar to those of previous studies [11]-[22], with a constant number (four) of principal components used for all sensor modalities to avoid overfitting. Specifics are given in Section 5 on cognitive load detection. An example comparison of formant-based correlation matrices for low and high load conditions for one subject is shown in Figure 5, indicating more complexity in the high-load correlation matrix. Differences of matrix eigenvalues from these matrices are shown at he right. Similar differences are observed in correlation matrices corresponding to the other speech features described above.

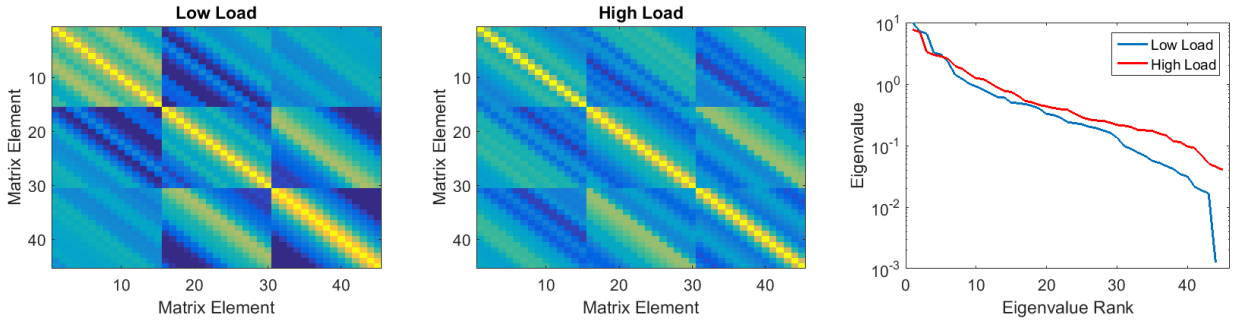

Figure 5: Example comparison of formant-based correlation matrices for low and high cognitive load for one subject. Matrix eigenvalues from these matrices are shown at right.

The differences in eigenspectra patterns due to high versus low cognitive loads provide indications about the effect of load on speech. In Figure 6, averages across all subjects of normalized (z-scored) eigenvalues from formant, creak, and delta-MFCC signals are shown for low load (blue)

and high load (red). The eigenvalues are ordered, left to right, from largest to smallest. So, in all three cases there is greater power in the medium level eigenvalues during higher cognitive load. This indicates greater dynamical complexity in formant frequencies, creak, and spectral content during higher cognitive load. The normalized eigenvalues are plotted in units of standard deviation.

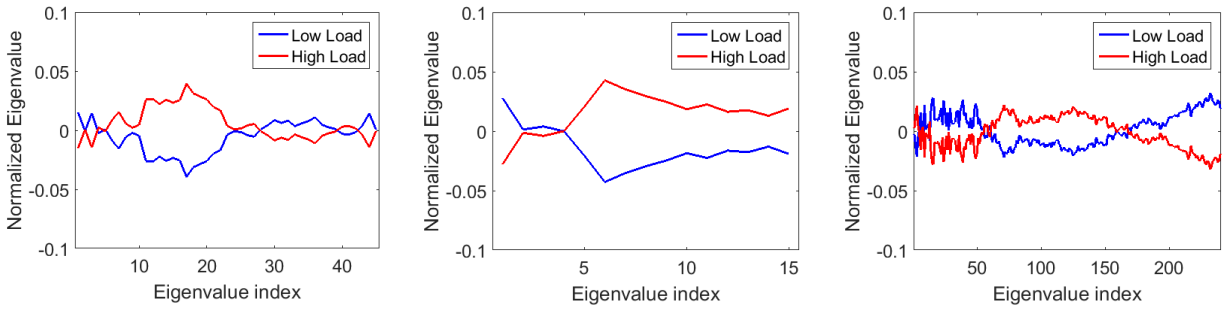

Figure 6. Correlation structure features: Average normalized eigenvalues from all subjects for low and high cognitive loads, based on formant frequencies (left), creak (center), and delta mel-cepstra (right).

Coherence Structure and Power: We have also introduced a feature set that characterizes the structure of signal coherence and power at multiple frequency bands. The coherence between channels, indicating the amount of cross-channel power in a frequency band relative to the amount of within-channel power, provides a measure of how closely related the signals are within a frequency band. The power and cross-power are computed among three formant frequency channels in three different frequency bands, and a 3 × 3 coherence matrix is constructed for each band. The eigenspectra of the coherence matrices indicate the structure of coherence across the channels. PCA is used to project these features into lower dimensional representations.

The differences in coherence and power features due to high versus low cognitive load provide indications about the effect of load on speech. In Figure 7 (left), averages across all subjects of normalized coherence eigenvalues from the middle frequency band (1.0–2.0 Hz) are shown for low load (blue) and high load (red). As before, the eigenvalues are ordered, left to right, from largest to smallest. Similar to the correlation structure results shown in Figure 7, these results indicate greater power in the mid-level eigenvalue for the higher load condition. In Figure 7 (right), it is shown that the higher load condition is also associated with more power (i.e., variability) in all three formant tracks. The normalized eigenvalues and power features are plotted in units of standard deviation.

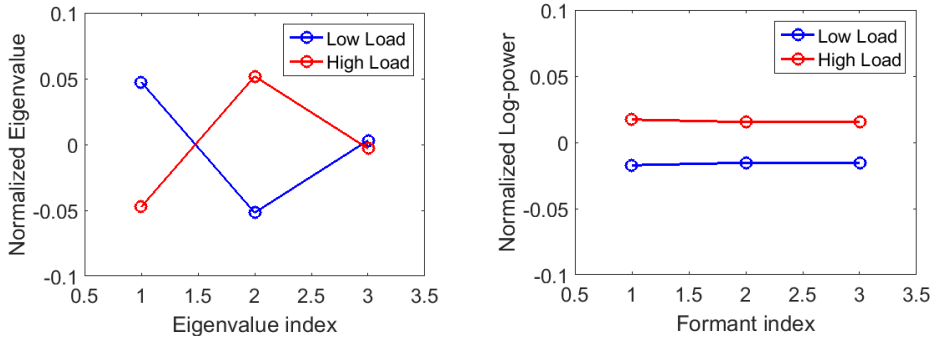

Figure 7. Left: Average normalized log eigenvalues from coherence matrix at frequency band 1.0–2.0 Hz for low and high cognitive loads from formant frequencies. Right: Normalized log power for the three formant frequency tracks at frequency band 1.0–2.0 Hz.

3.2 Video Features

3.2.1 Low-level features

Although Facial Action Units provide a formalized method for identifying changes in facial expression frame-by-frame [36], their extraction in large quantities of data has been impeded by the need for trained annotators to mark individual frames of a recorded video session. For this reason, the University of California San Diego has developed a computer expression recognition toolbox (CERT) for the automatic identification of FAUs from individual video frames [59]. Table 1 lists the FAUs output by CERT used for the video-based facial expression analysis.

Following CERT, all frames marked as invalid by the program and values considered outliers are removed. In addition, each frame of data is retained only if it is marked valid across all 20 FAUs. If the duration of the remaining FAU time series was less than 30s or 40% of their original length, the entire set of FAUs for that recording was not used.

Table 1. The 20 facial action units from CERT.

#	Description	#	Description
1	Inner Brow Raise	11	Lip Stretch
2	Outer Brow Raise	12	Cheek Raise
3	Brow Lower	13	Lids Tight
4	Eye Widen	14	Lip Pucker
5	Nose Wrinkle	15	Lip Tightener
6	Lip Raise	16	Lip Presser
7	Lip Corner Pull	17	Lips Part
8	Dimpler	18	Jaw Drop
9	Lip Corner Depressor	19	Lips Suck
10	Chin Raise	20	Blink/Eye Closure

Each FAU feature was converted from a support vector machine (SVM) hyperplane distance to a posterior probability using a logistic model trained on a separate database of video recordings [68]. Henceforth, the term FAU refers to these frame-by-frame estimates of FAU posterior probabilities.

3.2.2 High-level features

Our high-level features are designed to characterize properties of timing and coordination from the low-level features. Facial coordination features are obtained by applying the correlation structure technique to the FAU time series using the same parameters that were used to analyze the vocal-based features. Because of the 30 Hz FAU frame rate, the spacings of 1, 3, and 7 data points for the three time scales corresponds to time sampling in increments of approximately 33 ms, 100 ms, and 234 ms. Examples of correlation matrices for low and high loads (for one of the 11 subjects) are shown in Figure 8 (left and center) while average (across all 11 subjects) normalized eigenvalues from all subjects' trials are shown in Figure 8 (right). Once again, more power is found in the middle-level eigenvalues during high cognitive load.

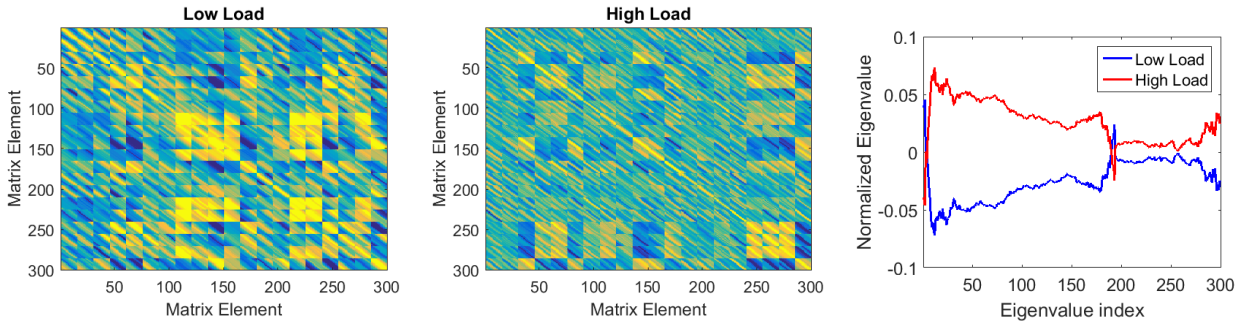

Figure 8: Example correlation matrix for low load (left) and high load (center) condition; Average normalized eigenvalues for correlation matrix for FAU features (right).

An alternative FAU-based feature set, *the facial activation rate*, can be obtained by computing mean FAU values (an estimate of percent time present via posteriori probabilities) over each passage and combining several of these into a fused FAU rate measure.

3.3 EEG Features

3.3.1 Low-level features

EEG signals were measured at 500 Hz with a 64-element Neuroscan system, followed by high-pass filtering and standard artifact removal. Measurements were made during the sentence listening and pause region of the protocol (Figure 3) to avoid motion and muscle artifacts during speaking.

3.3.2 High-level features

As with formant correlation structure, we have introduced feature sets that characterize the broadband correlation structure and coherence structure and power at multiple frequency bands. The correlation structure is computed similarly to above, except that a larger number of delay scales are used, with five delays per scale and with delay spacings of 5, 11, 23, 47, and 95. The coherence and power features are computed in five standard frequency bands (delta, theta, alpha, beta, gamma), and a coherence matrix is constructed for each band. Example correlation matrices for the low and high conditions from one subject are shown in Figure 9 (left and center), and average normalized eigenvalues across all subjects' trials in Figure 9 (right). Average normalized EEG coherence eigenvalues and channel log-power at the beta frequency band (16–31 Hz) are shown in Figure 10 . As with the voice and face biomarkers above, greater power is found in the middle to small EEG eigenvalues in the high load condition. Unlike the formant power features in Figure 7, we find that high load is associated with lower levels of EEG power.

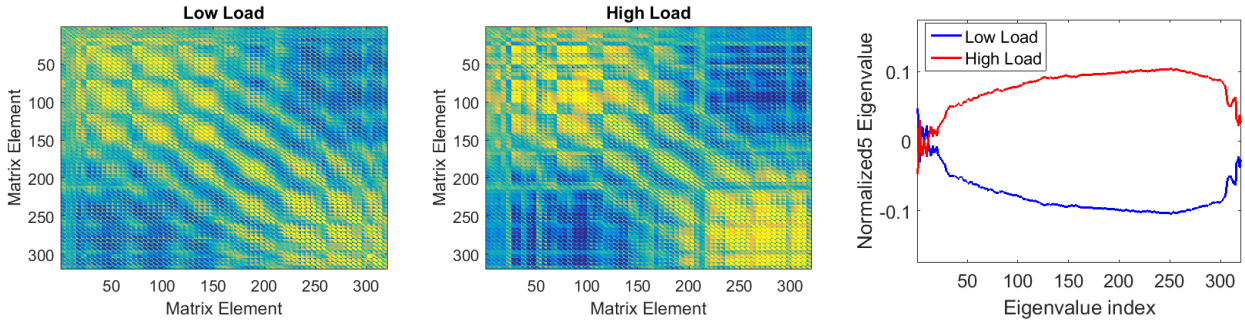

Figure 9: Example comparison of EEG-based correlation matrices for low and high cognitive load for one subject. Average normalized eigenvalues for correlation matrix for FAU features (right).

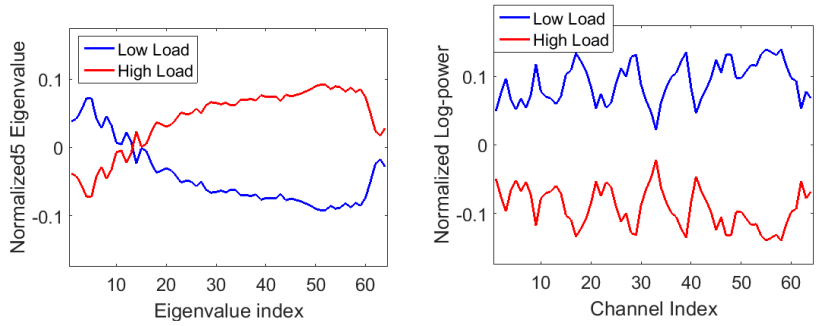

Figure 10: Average normalized log eigenvalues (left) for EEG coherence matrices and channel log power (right) at the beta frequency band (16–31 Hz) for low and high cognitive loads.

4. Results for cognitive load

Our goal is to detect differences in cognitive load from voice and face measurements and compare with EEG signal analysis. To evaluate detection performance, for each subject the 108 feature vectors (one vector per spoken sentence and load condition) from the max-digit condition is assigned to the *high* load class, and the 108 vectors from the max-digit-minus-two condition is assigned to the *low* load class. Leave-one-subject-out cross validation is used, with a classifier trained on the data from 10 held out subjects used to discriminate between high and low load on a test subject.

A key processing step is individualized feature normalization. This involves, for each subject (whether in the training or test set), subtracting the mean from each feature across both load conditions. This processing step is done to remove inter-subject feature variability, and implies that the ability to discriminate load conditions requires some knowledge of a subject's baseline features.

Load discrimination is done with a Gaussian classifier (GC), where the Gaussians are centered on the two class means, and a common covariance matrix is used based on the data across both load conditions. In each trial, the GC produces a load score (log-likelihood ratio of high versus low load). A receiver operating characteristic (ROC) curve is obtained by varying a detection threshold to characterize the sensitivity/specificity tradeoff. For each subject, 216 scores are obtained (108 for each load). A single ROC curve derived from scores of all 11 subjects characterizes total performance, with the area under the curve (AUC) serving as a summary statistic.

4.1 AUC as a summary statistic: Single trial case

In certain applications, vectors comprising our correlation- and coherence-based eigenspectra and covariance-based entropy and power have been concatenated into a single feature vector and then projected, using principal component analysis (PCA), into a lower-dimensional representation. In the current cognitive load application, better discriminative value was found by applying PCA separately to the multi-scale correlation-, covariance-, and coherence-based features.

Tables 2-4 list the number of features used by the Gaussian classifier for each feature set, and the AUC results. We see that the EEG modality achieves an AUC of 0.67, outperforming audio with an AUC of 0.56 and video with an AUC of 0.55. Table 5 summarizes various combinations of the features. The best overall performance of AUC = 0.68 is obtained by combining (via class fusion) all of the feature sets.

Table 2. Summary of area under ROC curve (AUC) results for detecting high cognitive load from a single trial (sentence) for the EEG modality. Coherence and channel features cover five frequency bands: delta, theta, alpha, beta, gamma.

Feature sets	description	# PCA features	AUC
1	Covariance and correlation structure	Cov-struct: 4 Corr-struct: 4	0.67
2	Coherence structure	3 for each band	0.53
3	Channel power	3 for each band	0.60
Combined			**0.67**

Table 3. Summary of area under ROC curve (AUC) results for detecting high cognitive load from a single trial (sentence) for the audio modality. Coherence and channel features are applied to formant tracks at three freq. bands: 0.25-1.0, 1.0-2.0, 2.0-4.0 Hz

Feature sets	description	# PCA features	AUC
1	Covariance and correlation structure of formants	Cov-struct: 4 Corr-struct: 4	0.52
2	Covariance and correlation structure of delta-MFCC	Cov-struct: 4 Corr-struct: 4	0.54
3	Covariance and correlation structure of creak	Cov-struct: 4 Corr-struct: 4	0.52
4	Coherence structure	3 for each band	0.53
5	Channel power	3 for each band	0.52
Combined			**0.56**

Table 4. Summary of area under ROC curve (AUC) results for detecting high cognitive load from a single trial (sentence) for the video modality.

Feature sets	description	# PCA features	AUC
1	Covariance and correlation structure	Cov-struct: 4 Corr-struct: 4	**0.55**

Table 5. Summary of area under ROC curve (AUC) results for detecting high cognitive load from a single trial (sentence) for the various feature combinations from audio, video, and EEG modalities.

Feature combinations	AUC
Audio + Video	0.57
EEG + Audio	0.68
EEG + Video	0.67
EEG + Audio + Video	0.68

4.2 Detection versus false alarm results

Although our protocol involves feature processing of single spoken sentences, the ability to detect load after fusing evidence across multiple sentences can be assessed by combining the Gaussian classifier scores from different trials, provided that the trials involve the same load condition. This was done by randomly selecting, from the same subject, a number of trials of either high load or low load, and summing their Gaussian classifier scores. For each subject, load condition and combination number, 200 randomly chosen sets of trials were used to determine the fused scores across multiple sentences.

Figures 11-12 summarize the ROC results (detection versus false alarm) for each modality alone and in combination. In Figure 11, we see a comparison of ROCs across each modality alone. We observe that the EEG-based detector rapidly converges up to a limit that is due to prediction failure on 1 out of the 11 subjects. With multiple trials, the EEG is getting 100% correct on 9 subjects, about 60% on one, and 0 % on another. Thus the ROC across all 11 subjects is limited by those last two subjects. The audio modality is converging more slowly to successful prediction on 10 of the 11 subjects, while the video modality is converging even more slowly to successful prediction on all 11 subjects. As seen in Figure 12, combining audio and video after 6 minutes only slightly underperforms EEG alone, while combining all three modalities provides a reasonable gain over any one modality.

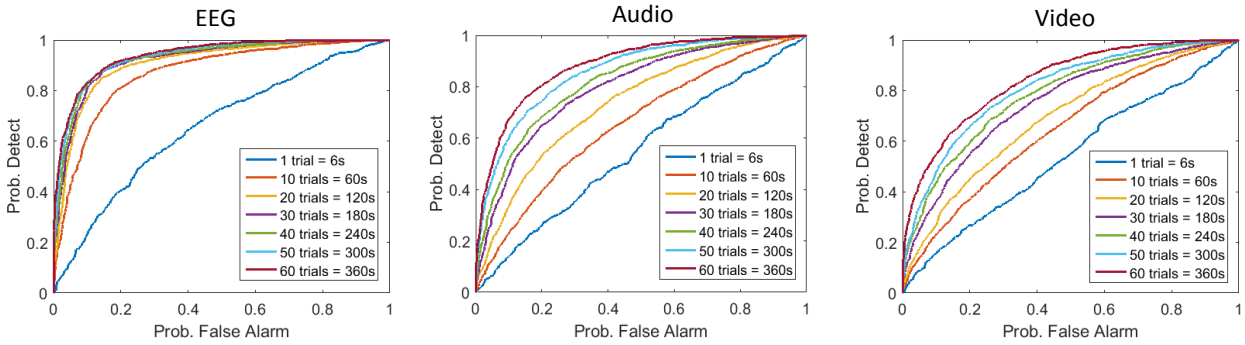

Figure 11. Independent probability of detection versus false alarm for EEG, audio, and video modalities. Each panel gives ROCs as a function of increasing number of trials from 1 to 360, corresponding to 6 s to 360 s (6 minutes) for low and high cognitive loads.

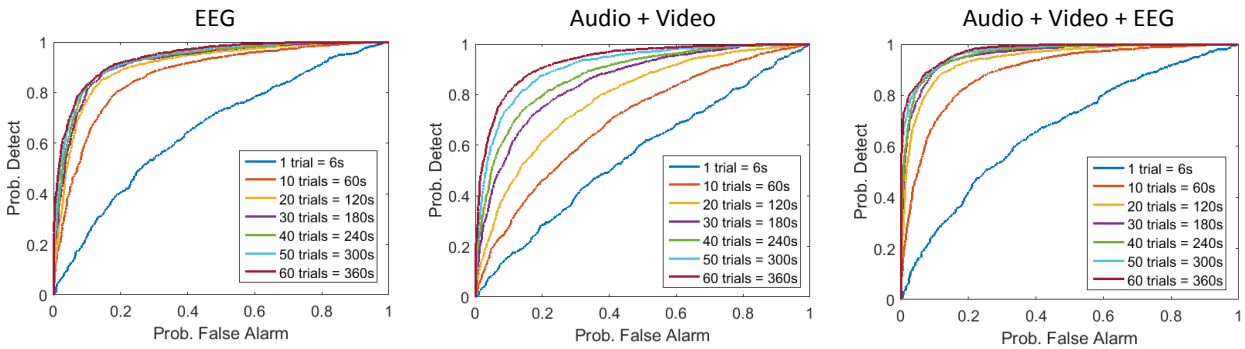

Figure 12. Fused probability of detection versus false alarm for combinations of EEG, audio, and video modalities. Each panel gives ROCs as a function of increasing number of trials from 1 to 60, corresponding to 6 s to 360 s (6 minutes) for low and high cognitive loads.

4.3 Analysis of convergence

Figure 13 contains boxplots (mean and variance) summarizing the AUC values for the 11 subjects within each modality, given combinations of 1, 5, 10, ..., 60 trials. The median AUC value for each modality increases as a function of the number of trials, with EEG detection accuracy quickly converging to 100% for 9 of 11 subjects, and audio and video converging much more slowly. Figure 14 contains boxplots comparing the EEG AUC values with two combined-modality results, which correspond to the ROC plots of Figure 12. Observe in Figures 13 and 14, however, outliers marked with red crosses that not included in calculation of the box plots. As noted earlier in our ROC discussion, with multiple trials, the EEG is getting 100% correct on 9 subjects, about 60% on one, and 0 % on another. Because more outliers tend to occur with the EEG than with the audio or video modality, one must consider this in drawing conclusions in comparing convergence and relative performance.

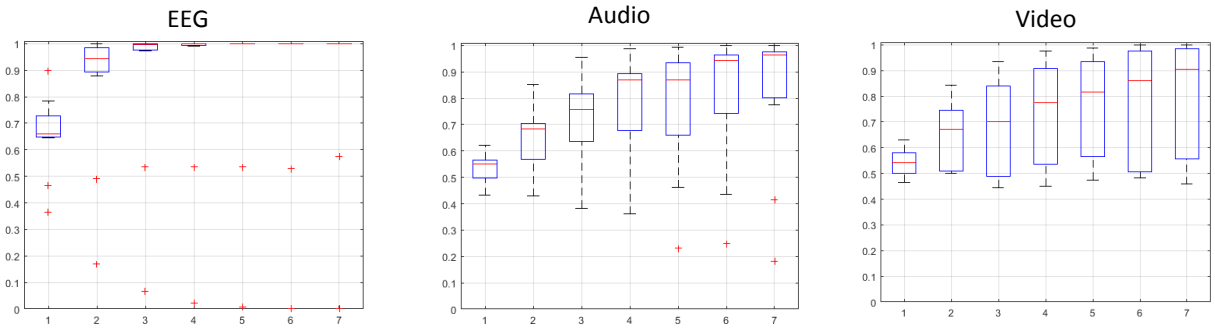

Figure 13. Independent AUC variance results across 11 subjects as a function of number of combined trials with same load for each modality. Outliers are marked with red + symbols and are not included in the variance calculation.

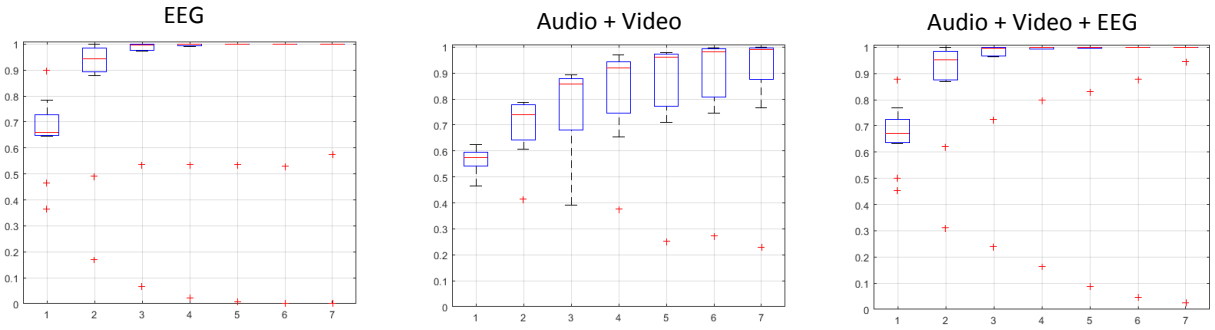

Figure 14. Fused AUC variance results across 11 subjects as a function of number of combined trials with same load for EEG, audio + video, and audio + video + EEG modalities.

5. Timing- and Coordination-based multimodal features in other conditions: Major Depression Disorder Case Study

In individuals with major depressive disorder, neurophysiological changes often alter motor control and thus affect the mechanisms controlling speech production and facial expression. These changes are typically associated with psychomotor retardation, a condition marked by slowed neuromotor output that is behaviorally manifested as altered coordination and timing across multiple motor-based properties. As with cognitive load, changes in motor outputs can be inferred from vocal acoustics and facial movements as individuals speak. Correspondingly our novel multi-scale correlation structure and timing feature sets from audio-based vocal features and video-based facial

action units have been shown to be effective. The feature sets enable detection of changes in coordination, movement, and timing of vocal and facial gestures that are potentially symptomatic of depression. Combining complementary features in Gaussian mixture model and extreme learning machine classifiers, our multivariate regression scheme predicts self-reported Beck Depression Inventory (BDI) ratings with a root-mean-square error of 8.12 and mean absolute error of 6.31 using a dataset from the 2014 Audio-Video Emotion Challenge (AVEC). In this section, we briefly summarize use of our principles of timing and coordination in vocal and facial expression.

5.1 Audio-video AVEC depression database

The 2014 Audio/Video Emotion Challenge (AVEC) uses a depression corpus that includes audio and video recordings of depressed subjects performing a human-computer interaction task [75]. Data were collected from 84 German subjects, with a subset of subjects recorded during multiple sessions: 31 subjects were recorded twice and 18 subjects were recorded three times. The subjects' age varied between 18 and 63 years, with a mean of 31.5 years and a standard deviation of 12.3 years.

Subjects performed two speech tasks in the German language: (1) reading a phonetically-balanced passage and (2) replying to a free-response question. The read passage (NW) was an excerpt of the fable *Die Sonne und der Wind* (*The North Wind and the Sun*). The free speech section (FS) asked the subjects to respond to one of a number of questions (prompted in written German), such as "What is your favorite dish?" "What was your best gift, and why?" and "Discuss a sad childhood memory." The NW and FS passages ranged in duration from 00:31 to 01:29 (mm:ss) and 00:06 to 03:50 (mm:ss), respectively.

Video of the subjects' face was captured using a webcam at 30 frames per second and a spatial resolution of 640 x 480 pixels. Audio was captured with a headset microphone connected to a laptop soundcard at sampling rates of 32 kHz or 48 kHz using the AAC codec. For each session, the self-reported BDI score was available. The recorded sessions were split into three partitions (training, development, and test) with 50 recordings in each set. We combined the training and development sets into a single 100-session data set, which is henceforth termed the Training set.

5.2 Feature extraction

As with the cognitive load problem, our high-level voice and face features are designed to characterize properties of coordination and timing from the low-level features of Section 3. The measures of coordination use assessments of the multi-scale structure of correlations among the low-level features. As before, this approach is motivated by the observation that auto- and cross-correlations of measured signals can reveal hidden parameters in the stochastic-dynamical systems that generate the time series. For vocal-based timing features we use cumulative phoneme-dependent durations and pitch slopes, obtained using estimated phoneme boundaries. For facial-based timing features, we use FAU rates obtained from their estimated posterior probabilities.

5.2.1 Speech features

Speech correlation structure

Referring back to Section 3, low-level speech features selected are based on articulatory (formant) correlations, source (vocal fold irregularity) correlations, and articulatory-to-source correlations. As an example, Figure 15 shows the correlation structure matrix for the source correlation of cepstral-peak-prominence and harmonic-to-noise ratio (CPP–HNR). The matrix is based on vectors that consist of 88 elements (2 channels, 4 scales, 15 delays per scale, and 2 covariance features per

scale) comparing one control and depressed subject. Figure 15 also shows the top 20 eigenvalues per scale corresponding to these subjects and the average normalized eigenvalues across all subjects for four different depression severity ranges. Similar correlation structures and eigenvalue spreads are found for formant correlation features, delta MFCC correlation features, and formant-CPP (articulatory-to-source) features.

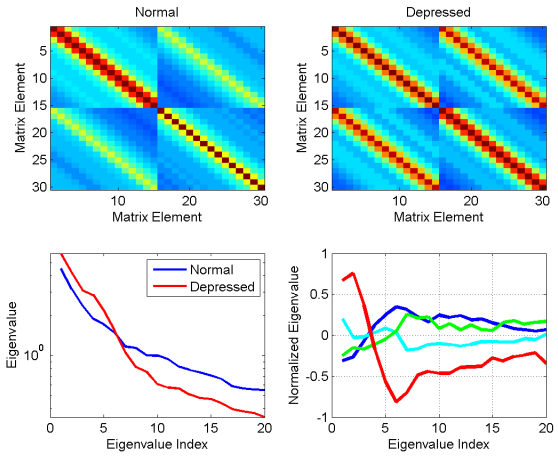

Figure 15. CPP–HNR correlation features. Top: Channel-delay correlation matrices from NW passage for a normal and a depressed subject. Bottom: Eigenvalues for these subjects (left) and average normalized eigenvalues for four Beck assessment ranges in the training set (right).

Phoneme-based featured

Phoneme-dependent features:

Phoneme durations**:** Based on phoneme boundaries introduced as low-level features in Section 2, we find that computing phoneme-specific characteristics, rather than the more typical average measures of speaking rate, can reveal stronger relationships between speech rate and depression severity [12][15]. Figure 16 shows the example of average phoneme durations which can be used themselves as features or as a basis for other features such as when the most highly correlating with a disorder assessment are combined.

Pitch slopes**:** From the pitch estimate above, within each phone segment, a linear fit is made to the pitch values, yielding a pitch slope feature (ΔHz/s) associated with each instance of phonetic speech units. As with phoneme durations, these average values can be used themselves as features or as a basis for other features such as when the phoneme-dependent pitch slopes that are most highly correlating with a disorder are combined.

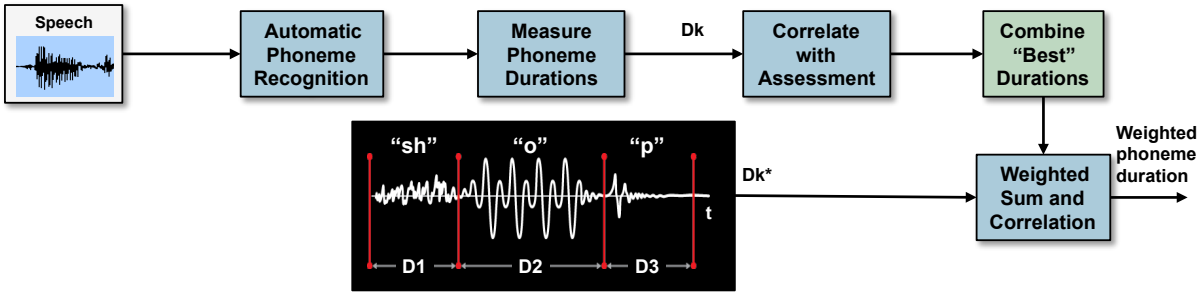

Figure 16: Phoneme recognizer provides boundaries. In one feature set, average phoneme durations are correlated against severity of a disorder and combined according to the highest correlations.

Based on estimated average durations for each phoneme, the summed average durations of certain phonemes are linearly combined to yield fused phoneme duration measures. A subset of phonemes whose summed durations are highly correlated with BDI scores on the training set are selected to create these fused measures, with weights based on the strength of their individual correlations. Table 6 lists the selected phonemes for the North Wind passage (left) and the first six of the ten selected phonemes for the Free Speech passage (right), along with their individual BDI correlations. The correlations of the fused measures for each passage are shown at the bottom. The linear combination used to obtain the fused measures is more fully described in [13].

Table 6. Correlation coefficients (R, $p < 0.01$) between fused phoneme durations and BDI scores in the training set. Fusion is done using linear combinations of phoneme durations. Only 6 of the 10 Free Speech phonemes are shown.

North Wind		**Free Speech**	
Phoneme	***R***	**Phoneme**	***R***
'l'	0.50	'ng'	0.38
'ah'	0.45	't'	0.34
'n'	0.41	'hh'	0.33
'ih'	0.34	'ey'	0.32
'b'	0.34	'ow'	0.28
'ow'	0.34	'er'	0.27
Fused	**0.54**	**Fused**	**0.57**

A fused phoneme-dependent pitch slope measure is also obtained using essentially the same procedure as described above. For each phoneme, we compute the sum of *valid* pitch slopes across all instances of that phoneme. Invalid slopes are those with absolute value greater than eight, resulting in the exclusion of most slopes that are computed from discontinuous pitch contours. For each passage, the set of phonemes with the highest correlating summed pitch slopes are then selected. The summed pitch slopes are combined to obtain fused measures for the NW and FS passages. Using 20 phonemes for NW and 15 phonemes for FS, these fused measures have BDI correlations of $\boldsymbol{R}=0.63$ (NW) and $\boldsymbol{R}=0.51$(FS).

5.2.3 Video features

Facial Correlation Structure

Facial coordination features are obtained by applying the correlation technique to the FAU time series using the same parameters that were used to analyze the vocal-based features. Because of the 30-Hz FAU frame rate, spacing for the four time scales correspond to time sampling in increments of approximately 33 ms, 100 ms, 234 ms, and 500 ms. Figure 17 (top) shows example FAU channel-delay matrices at a single time scale from the same normal and depressed subjects that were used for illustration in Figure 15. These matrices are derived from the FS passage. As with the correlation-based speech features, Figure 17 (bottom-left) shows that the eigenspectra of the depressed subject contain less power in the small eigenvalues. This effect is observed across a spectrum of BDI scores in all 83 free-response training set recordings with valid FAU features. The facial-based eigenvalue differences are similar to those found in the correlation-based speech features.

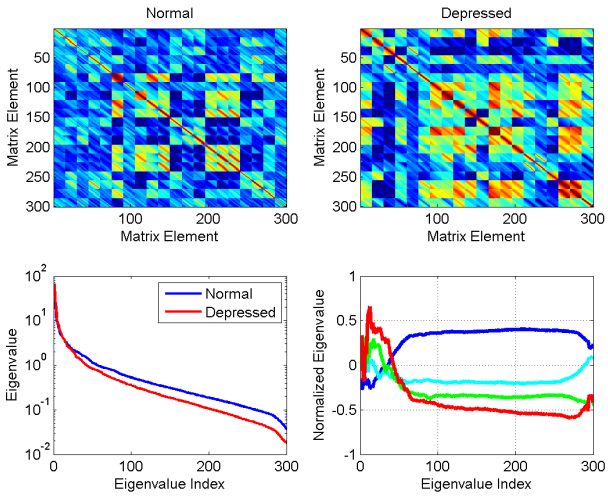

Figure 17. FAU correlation features. Top: Channel-delay correlation matrices from FS passage for a normal and a depressed subject. Bottom: Eigenvalues for these subjects (left) and average normalized eigenvalues for four BDI ranges in Training set (right).

FAU Rate

We also used FAU rate as described in Section 3. Weights are based on FAU correlations with BDI scores using the combination rule based on highest correlation [13]. Correlation coefficient (R) between mean FAU posterior probabilities and BDI in the training set ($p < 0.05$ for all $|R| \geq 0.21$). Fusion is done using linear combinations of the mean FAU posterior probabilities.

5.2.4 Dimensionality reduction

The correlation feature vectors typically contain highly correlated elements. To obtain lower-dimensional uncorrelated feature vectors for machine learning techniques, we apply principal component analysis (PCA). Table 6 lists the number of principal components we chose for each correlation feature type, along with phonetic and FAU rate features. The number of principal components in each case was determined empirically by cross-validation performance.

Table 6. Total number of dimensions (# Dim.) and number of dimensions selected after principal component analysis (PCA #) for each of the eight features sets.

Feature Set	Data	Feature Type	# Dim.	PCA #
1	NW	Formant–CPP *xcorr*	248	4
	NW	CPP–HNR *xcorr*	88	2
	NW	Delta MFCC *xcorr*	968	5
2	NW	Phoneme duration	1	1
3	NW	Pitch slope	1	1
4	NW	FAU rate	1	1
5	FS	FAU *xcorr*	1208	6
6	FS	Phoneme rate	1	1
7	FS	Pitch slope	1	1
8	FS	FAU rate	1	1

5.3 Multivariate fusion and prediction

Our next step involves mapping the features described in Section 5.2 into univariate scores that can be easily mapped into BDI predictions. To do this, we use both generative Gaussian mixture models (GMMs), which have been widely used for automatic speaker recognition [77] and have recently been extended to vocal-based depression classification [13][22], and discriminative extreme learning machines (ELMs), a single layer feedforward neural network architecture with randomly assigned hidden nodes [78][79].

Gaussian staircase: To train the generative GMMs, we utilize the *Gaussian staircase* approach in which each GMM is comprised of an ensemble of Gaussian classifiers [13][22]. The ensemble is derived from six partitions of the training data into different ranges of depression severity for low (Class 1) and high (Class 2) depression. Given a BDI range of 0 to 45, the Class 1 ranges for the six Gaussian classifiers are: 0−4, 0−10, 0−17, 0−23, 0−30, and 0−36, with the Class 2 ranges being the complement of these. The Gaussian classifiers comprise a single, highly regularized GMM classifier, with feature densities that smoothly increase in the direction of decreasing (Class 1) or of increasing (Class 2) levels of depression. Additional regularization of the densities is obtained by adding 0.1 to the diagonal elements of the normalized covariance matrices.

Subject-based adaptation: Individual variability in the relationships between features and BDI are partially accounted for within the GMMs using Gaussian-mean subject-based adaptation. Motivated by GMM adaptation methods in automatic speaker recognition [77], if one or more sessions in the Training set have the same subject ID as the Test subject and are in the same BDI-based partition, the mean of the Gaussian for that partition is assigned to the mean of the data from that subject only, rather than the mean of the data from all subjects within the partition [13][22].

Fusion: A separate GMM classifier is used for each Feature Set, outputting a log-likelihood ratio score for Class 1 (Normal) and Class 2 (Depressed) [13]. Separate ELM classifiers are used for Feature Sets 1 and 2. Initial BDI predictions are obtained from three Predictors 1, 2,3, which use different combinations of the eight Feature Sets and two types of classifiers. Within each Predictor, the classifier outputs from Feature Sets are summed together. Following this, a univariate regression model is created from the Training set and applied to the classifier output from the Test data. The resulting univariate regression output is the intitial BDI score prediction from each Predictor. For Predictors 1 and 2, subject-based adaptation is then applied to adjust this initial prediction by correcting for consistent biases seen in the BDI Training set predictions of the same subject. If there are any Training sessions from a given Test subject, then the average Training set error from that subject is used to adjust the prediction. Details of our fusion methodology are described in [13].

5.4 Results

The prediction system described above was used in our winning system in the AVEC 2014 Challenge, with test RMSE = 8.12 and MAE = 6.31. These results are an improvement on our winning submission in the AVEC 2013 competition, which was test RMSE = 8.50 and MAE = 6.52. The 2013 result was obtained using voice only and a read passage (*Homo Faber*) that was much longer than the NW passage made available in 2014. Introduction of vocal and facial features helped improved performance in 2014 despite the relative lack of data in this challenge. A different perspective on these results is shown in Figure 18 giving the ROC curve (probability of detection versus false alarm) for both the AVEC 2013 and 2014 databases**.** In this binary detection problem, two classes are mild/moderate and moderate/severe severity levels. With significantly less data (2014) using voice and face, performance is comparable to a voice-only system (2013).

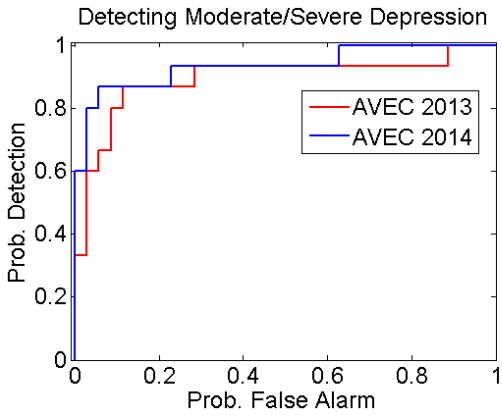

Figure 18. Probability of detection versus false alarm of depression using AVEC 2013 and 2014 databases. With significantly less data (2014) using voice and face, performance is comparable to a voice-only system (2013).

6. Conclusions and Discussion

In this chapter, we demonstrated the power of a multimodal approach using speech and facial features to discriminate between high and low cognitive load conditions, and illustrated its generality to cognitive disorders using a case study in major depressive disorder (MDD). Our vocal features capture timing and inter-relationships among phoneme durations, pitch dynamics, articulation, spectral dynamics, and creak, while facial features capture relations and rate of facial action units underlying facial muscle activity. In our cognitive load study, as a reference, we extracted EEG features that reflect relations across EEG channels. Using a database consisting of audio, video, and EEG from 11 subjects and recalled sentences and pauses prior to recalling a digit span, we effectively applied classification models of cognitive load and explored tradeoffs with audio and video in comparison with the EEG "gold standard". We found that by merging audio and video brought us close to EEG-based performance, thus providing a simple noninvasive alternative to a complex 64-channel EEG. In illustrating the generalizability of our timing and coordination features to neurological disease, we briefly described a predictor of depression state based on vocal and facial modalities using the 2014 Audio/Video Emotion Challenge (AVEC). On-going work involves expansion of our approach to a larger suite of modalities (facial video, physiology, eye tracking, gait, EGG) as well as biomarkers that reflect timing and correlation both within and across these modalities

7. References

1. Lively, S.E., Pisoni, D.B., Van Summers, W., Bernacki, R.H., "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences," J Acoust Soc Am. 1993 May; 93(5): 2962–2973.
2. Yin, B., Chen, F., Ruiz, N., Ambikairajah, E., "Speech-based cognitive load monitoring system, ICASSP 2008.
3. Yin, Bo, and Fang Chen. "Towards automatic cognitive load measurement from speech analysis." *Human-Computer Interaction. Interaction Design and Usability*. Springer Berlin Heidelberg, 2007. 1011-1020.

4. Khawaja, M.A., Ruiz, N., Cheng, F., “Think before you talk: An empirical study of relationship between speech pauses and cognitive load,” OZCHI 2008, December 8-12, 2008.
5. Le, P., J. Epps, Choi, H.C., and Ambikairajah, E., “A study of voice source- and vocal tract-based features in cognitive load classification,” Proceedings of International Conference on Pattern Recognition, 2010, pp. 4516-4519.
6. Boril, H., Sadjadi, O., Kleinschmidt, T., and J. Hansen, Analysis and detection of cognitive load and frustration in drivers’ speech,” Proceedings of Interspeech, 2010, pp. 502-505.
7. Yap, T.F., Speech Production Under Cognitive Load: E˙ects and Classification, PhD Thesis, The University of New South Wales School of Electrical Engineering and Telecommunications Sydney, Australia, Sept. 2011.
8. P. Zarjam, J. Epps, et al., "Characterizing working memory load using EEG delta activity." in the 19th Eusipco Conference, pp. 1554-1558, 2011.
9. P. Zarjam, J. Epps, et al., "Spectral EEG features for evaluating cognitive load.", in the 33rd EMBS Conference, pp. 3841-3844, 2011.
10. P. Zarjam, J. Epps, et al., "Evaluation of working memory load using EEG signals.", in the 2nd APSIPA Conference, pp. 715–719, 2011.
11. Yu, B., Quatieri, T.F., Williamson, J.W., and Mundt, J., "Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers." Fifteenth Annual Conference of the International Speech Communication Association. 2014.
12. Quatieri, T.F. and Malyska, N. 2012. Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity. Interspeech (2012).
13. Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., & Mehta, D. D. (2014, November). Vocal and Facial Biomarkers of Depression based on Motor Incoordination and Timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (AVEC), pp. 65-72. ACM. (Winning paper in (AVEC) Depression Challenge)
14. Helfer, B. S., Quatieri, T. F., Williamson, J. R., Keyes, L., Evans, B., Greene, W. N., Palmer, J., & Heaton, K. (2014). Articulatory Dynamics and Coordination in Classifying Cognitive Change with Preclinical mTBI. In *Fifteenth Annual Conference of the International Speech Communication Association.*
15. Trevino, A., Quatieri, T. F. and Malyska, N., “Phonologically-based biomarkers for major depressive disorder,” EURASIP Journal on Advances in Signal Processing: Special Issue on Emotion and Mental State Recognition from Speech, 42:2011–2042, 2011
16. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T.F. Quatieri, A review of depression and suicide risk assessment using speech analysis, Speech Communication, Vol. 17, July 2015, Pages 10–49.
17. T.F. Quatieri, J.R. Williamson, C.J. Smalt, T. Patel, J. Perricone, D.D. Mehta, B.S. Helfer, G. Ciccarelli, D. Ricke, N. Malyska, J. Palmer, K. Heaton, M. Eddy, J. Moran, Vocal biomarkers to discriminate cognitive load in a working memory task, Interspeech 2015.
18. B. Yu, T.F. Quatieri, J.R. Williamson, J.C. Mundt,˙ Cognitive impairment prediction in the elderly based on vocal biomarkers, Interspeech 2015.
19. J.R. Williamson, T.F. Quatieri , B.S. Helfer, J. Perricone, S.S. Ghosh, G. Ciccarelli, D.D. Mehta, Segment-dependent dynamics in predicting Parkinson’s disease, accepted, Interspeech 2015.
20. Williamson, J.R., Bliss, D.W. and Browne, D.W. 2011. Epileptic seizure prediction using the spatiotemporal correlation structure of intracranial EEG. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (2011), 665–668.
21. Williamson, J.R., Bliss, D.W., Browne, D.W. and Narayanan, J.T. 2012. Seizure prediction using EEG spatiotemporal correlation structure. Epilepsy & Behavior. 25, 2 (2012), 230–238.

22. Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B. and Mehta, D.D. 2013. Vocal biomarkers of depression based on motor incoordination. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (2013), 41–48.
23. Moore, E., Clements, M., Peifer, J. and Weisser, L. 2003. Analysis of prosodic variation in speech for clinical depression. Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE (2003), 2925–2928.
24. Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K. and Geralts, D.S. 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of neurolinguistics. 20, 1 (Jan. 2007), 50–64.
25. Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K. and Wilkes, D.M. 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. Biomedical Engineering, IEEE Transactions on. 51, 9 (2004), 1530–1540.
26. Darby, J.K., Simmons, N. and Berger, P.A. 1984. Speech and voice parameters of depression: A pilot study. Journal of Communication Disorders. 17, 2 (1984), 75–85.
27. Dejonckere, P. and Lebacq, J. 1996. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. ORL. 58, 6 (1996), 326–332.
28. Fava, M. and Kendler, K.S. 2000. Major depressive disorder. Neuron. 28, 2 (2000), 335–341.
29. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M. and Wilkes, D.M. 2000. Acoustical properties of speech as indicators of depression and suicidal risk. Biomedical Engineering, IEEE Transactions on. 47, 7 (2000), 829–837.
30. Greden, J.F. and Carroll, B.J. 1981. Psychomotor function in affective disorders: An overview of new monitoring techniques. The American journal of psychiatry. (1981).
31. Orozco-Arroyave, J., Arias-Londoño, J., Vargas-Bonilla, J., González-Rátiva, M., and Nöth, E. (2014) New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease, in *Proc. LREC*, 342–347.
32. Canter, G.J., 1963. Speech characteristics of patients with Parkinson's disease: I. Intensity, pitch, and duration. *J. Speech Hear. Disord.* 28 (3), 221–229.
33. Canter, G.J., 1965a. Speech characteristics of patients with parkinson's disease: II. Physiological support for speech. *J. Speech Hear. Disord.* 30 (1), 44–49.
34. Canter, G.J., 1965b. Speech characteristics of patients with Parkinson's disease: III. Articulation, diadochokinesis, and over-all speech adequacy. *J. Speech Hear. Disord.* 30 (3), 217–224.
35. Logemann, J.A., Fisher, H.B., Boshes, B., Blonsky, E.R., 1978. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *J. Speech Hear. Disord.* 43(1), 47.
36. Ekman, P., Freisen, W.V. and Ancoli, S. 1980. Facial signs of emotional experience. Journal of personality and social psychology. 39, 6 (1980), 1125.
37. Gaebel, W. and Wölwer, W. 1992. Facial expression and emotional face recognition in schizophrenia and depression. European archives of psychiatry and clinical neuroscience. 242, 1 (1992), 46–52.
38. Shen, W., White, C., Hazen, T.J., "A comparison of query-by-example methods for spoken term detection," in Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (2010)
39. Mehta, D. D., Rudoy, D. and Wolfe, P. J., "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," The Journal of the Acoustical Society of America, 132(3):1732–1746, 2012.
40. Singer, J.D. and Willett, J.B., "Applied longitudinal data analysis: Modeling change and event occurrence," Oxford university press, 2003.

41. Park et al., 2010 H. Park, R. Felty, K. Lormore, D. Pisoni PRESTO: perceptually robust English sentence test: open set—design, philosophy, and preliminary findings J. Acoust. Soc. Am., 127 (2010), p. 1958
42. Levitt, H., 1971. Transformed up-down methods in psychoacoustics. J. Acoust. Soc. Am. 49, 467–477.
43. Le, P.N., Ambikairajah, E, Choi, H.C., and J. Epps, "A non-uniform sub-band approach to speech-based cognitive load classification," Proceedings of ICICS, 2009, pp. 1-5.
44. Harnsberger, James D., Richard Wright, and David B. Pisoni. "A new method for eliciting three speaking styles in the laboratory." *Speech communication* 50.4 (2008): 323-336.
45. Rouas J., "Automatic Prosodic Variations Modeling for Language and Dialect Discrimination, IEEE Trans.Audio, Speech, and Language Proc., Vol. 15, Nop. 6, August 2007.
46. Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences. 17, (1993), 97–110.
47. Gerratt, B. R., and Kreiman, J., "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, no. 4, pp. 365-381, 2001.
48. Ishi, C.T., Sakakibara, K.I., Ishiguro, H., and Hagita, N., "A method for automatic detection of vocal fry," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 47-56, 2008.
49. Kane, J., Drugman, T., and Gobl, C., "Improved automatic detection of creak," *Computer Speech & Language*, vol. 27, no. 4, pp. 1028-1047, 2013.
50. http://tcts.fpms.ac.be/~drugman/Toolbox/.
51. Williamson, J.R., Bliss, D., Browne, D.W., and Narayanan, J.T., "Seizure prediction using EEG spatiotemporal correlation structure," *Epilepsy Behav.*, vol. 25, no. 2, pp. 230–238, 2012
52. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, Jul. 2015.
53. Wang, P., Barrett, F., Martin, E., Milonova, M., Gurd, R.E., Gur, R.C., Kohler, C., and Verma, R.: 'Automated video-based facial expression analysis of neuropsychiatric disorders', Journal of Neuroscience Methods, 2008, 168, pp. 224–238.
54. Schmidt, K.L., Bhattacharya, S., and Denlinger, R.: 'Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises', Journal of Nonverbal Behavior, 2009, 33, pp. 35–45.
55. H. Park, R. Felty, K. Lormore, D. Pisoni, "PRESTO: Perceptually robust English sentence test: open set—design, philosophy, and preliminary findings," J. Acoust. Soc. Am., 127 (2010), p. 1958.
56. Levitt, H., 1971. "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. 49, 467–477
57. Le, P.N., Ambikairajah, E, Choi, H.C., and J. Epps, "A non-uniform sub-band approach to speech-based cognitive load classification," Proceedings of ICICS, 2009, pp. 1-
58. Harnsberger, J.D., Wright, R., and Pisoni, D.B., "A new method for eliciting three speaking styles in the laboratory," Speech communication, 50.4 (2008): 323-336.
59. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. and Bartlett, M. 2011. The computer expression recognition toolbox (CERT). Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on (2011), 298–305.

60. Jackson, P.J. and Shadle, C.H. 2000. Performance of the pitch-scaled harmonic filter and applications in speech analysis. Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on (2000), 1311–1314.
61. Jackson, P.J. and Shadle, C.H. 2001. Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. Speech and Audio Processing, IEEE Transactions on. 9, 7 (2001), 713–726.
62. Dejonckere, P. and Lebacq, J. 1996. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. ORL. 58, 6 (1996), 326–332.
63. Heman-Ackah, Y.D., Heuer, R.J., Michael, D.D., Ostrowski, R., Horman, M., Baroody, M.M., Hillenbrand, J. and Sataloff, R.T. 2003. Cepstral peak prominence: a more reliable measure of dysphonia. Annals of Otology Rhinology and Laryngology. 112, 4 (2003), 324–333.
64. Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N. and De Bodt, M. 2010. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. Journal of voice. 24, 5 (2010), 540–555.
65. Heman-Ackah, Y.D., Michael, D.D. and Goding Jr, G.S. 2002. The relationship between cepstral peak prominence and selected parameters of dysphonia. Journal of Voice. 16, 1 (2002), 20–27.
66. Hillenbrand, J. and Houde, R.A. 1996. Acoustic Correlates of Breathy Vocal QualityDysphonic Voices and Continuous Speech. Journal of Speech, Language, and Hearing Research. 39, 2 (1996), 311–321.
67. Low, L.-S., Maddage, M., Lech, M., Sheeber, L. and Allen, N. 2010. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (2010), 5154–5157.
68. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (2010), 94–101.
69. Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N. and De Bodt, M. 2010. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. Journal of voice. 24, 5 (2010), 540–555.
70. Mehta, D.D., Rudoy, D. and Wolfe, P.J. 2012. Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. The Journal of the Acoustical Society of America. 132, 3 (Sep. 2012), 1732–46.
71. Gerratt, B. R., and Kreiman, J., "Toward a taxonomy of nonmodal phonation," Journal of Phonetics, vol. 29, no. 4, pp. 365-381, 2001.
72. Ishi, C.T., Sakakibara, K.I., Ishiguro, H., and Hagita, N., "A method for automatic detection of vocal fry," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 47-56, 2008.
73. Kane, J., Drugman, T., and Gobl, C., "Improved automatic detection of creak," Computer, Speech and Language, vol. 27, no. 4, pp. 1028-1047, 2013.
74. http://tcts.fpms.ac.be/~drugman/Toolbox/.
75. Quatieri, T. F. (2002). Discrete-time speech signal processing: principles and practice. Pearson Education.
76. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R. and Pantic, M. 2013. AVEC 2014–3D Dimensional Affect and Depression Recognition Challenge. (2013).

77. Reynolds, D.A., Quatieri, T.F. and Dunn, R.B. 2000. Speaker verification using adapted Gaussian mixture models. Digital signal processing. 10, 1 (2000), 19–41.
78. Guang-Bin Huang, Hongming Zhou, Xiaojian Ding and Rui Zhang 2012. Extreme Learning Machine for Regression and Multiclass Classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 42, 2 (Apr. 2012), 513–529.
79. Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K. 2006. Extreme learning machine: Theory and applications. Neurocomputing. 70, 1-3 (Dec. 2006), 489–501.