IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 11, 2018, accepted June 30, 2018, date of publication July 18, 2018, date of current version August 15, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2856478

Speech-Based Automated Cognitive Impairment Detection From Remotely-Collected Cognitive Test Audio

BEA YU[1], JAMES R. WILLIAMSON[1], JAMES C. MUNDT[2], AND THOMAS F. QUATIERI[1], (Fellow, IEEE)
[1]MIT Lincoln Laboratory, Lexington, MA 02421, USA
[2]Sand Ridge Secure Treatment Center, Mauston, WI 53948, USA

Corresponding author: Bea Yu (bea.yu@ll.mit.edu)

This work was supported in part by the National Institute on Aging under Grants 2U19AG010483 and 1R41AG044218, in part by the National Institute of Health through the Air Force Contract under Grant FA8721-05-C-0002, and in part by the Assistant Secretary of Defense for Research and Engineering through the Air Force Contract under Grants FA8721-05-C-0002 and/or FA8702-15-D-0001. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

ABSTRACT Remote-automated cognitive impairment (CI) monitoring has the potential to facilitate care for the elderly with mobility restrictions. In particular, CI detection based on speech features from audio data collected for remote cognitive testing holds significant promise to improve remote cognitive health monitoring. This requires no additional testing for speech analysis and, combined with cognitive test scores, can improve CI detection over using cognitive test scores alone. This paper builds on previous work with an expanded set of speech features extracted from a larger suite of remotely administered cognitive tests. The speech features tested include measures of phoneme characteristics, pitch, and articulation. The relative merits of using speech features, a common cognitive test score, and both combined for CI prediction are also explored. The best performing system uses a combination of speech features and the cognitive test score, obtaining a performance outcome of area under the ROC curve (AUC) $= 0.77$. This outcome is better at the 5% significance level than that obtained using the speech features alone (AUC $= 0.74$) or the cognitive test alone (AUC $= 0.54$). Additionally, the influence of validation methodology on performance estimation is addressed in detail. Learning statistical models for speech-based CI diagnosis is challenging due to limited availability of audio data from subjects with clinical CI diagnoses. Rigorous validation methods for model learning are important in this context. The stringent validation methodology developed in this paper produces more conservative, and likely, more generalizable performance estimates compared with methodologies used in prior art.

INDEX TERMS Mild cognitive impairment, motor coordination, vocal biomarkers, formant frequencies, phoneme durations, cross-validation, feature selection, remote health care monitoring.

I. INTRODUCTION

Constraints on elderly mobility and human resources for elder care have spawned an active area of research in technology to enable remote, automated monitoring as part of an assisted senior living system. Several tests to assess cognitive functioning in the elderly can be administered remotely over a telephone or the internet. Such tests involve the collection of audio responses from a participant to be either manually or automatically scored. In general, cognitive tests for mild cognitive impairment (MCI) and dementia of the Alzheimer's type (DAT) include tests of episodic memory, executive functioning, language, spatial skills, and attention [1]. Clinicians typically administer several different tests to facilitate MCI/DAT diagnosis. In category fluency tests, a subject is asked to name as many examples from a category (animals, vegetables, items in a grocery store, etc.) as possible within a short time period, often 30 seconds or one minute [1]. Category fluency tests have been used with some success to detect DAT [2]–[7] and MCI [8]–[12]. In another example test, the East Boston memory

test (EB), participants are told a story and asked to summarize the content of the story immediately after hearing it and again after a specific delay. Category fluency tests and the East Boston memory test are well suited for remote administration.

While the linguistic content of cognitive test audio samples is primarily leveraged for cognitive assessment, there is also active research in extracting information from the acoustic speech signal itself to detect DAT and MCI. The coupling of speech and language with dementia [13], [14] entails correlations between vocal features reflecting prosody, voice quality, and linguistic complexity and various degrees of cognitive impairment [12], [15], [16], [21], [22]. This idea is not unique to the DAT and MCI research community. There is a growing consensus that non-linguistic vocal features are powerful indicators of multiple dimensions of mental condition and emotional state. These features, which include characterizations of prosody (e.g., fundamental frequency and speaking rate), spectral representations (e.g., mel cepstra), articulation (e.g., formant frequency correlation structure) and glottal excitation (e.g., timing jitter, amplitude shimmer, and aspiration) have been applied to detect neurocognitive and neuromotor changes from a variety of causes such as age-related cognitive impairment, major depressive disorder, mild traumatic brain injury, Parkinson's disease, and cognitive workload [15]–[23].

In the particular domain of speech-based CI detection, several results have been published in the past decade. Satt *et al.* [21], [22] achieved an $18\% \pm 6\%$ equal error rate (EER) for MCI/dementia prediction using speech features from Greek speech and a $20\% \pm 6\%$ EER for MCI prediction using a different dataset in French. In both cases, their audio data was collected in person and consisted of at least three different speech tasks designed for the purpose of extracting non-linguistic speech features to detect MCI and DAT. Roark *et al*. [17] explored combining semantic and speech features from cognitive test audio collected in person for CI detection. They used various combinations of English speech features, natural language processing features and nine different cognitive test scores for classification of MCI and obtained an AUC of 0.86 with their best set of 17 features. Two of our earlier works used vocal features from English speech remotely collected in the Alzheimer's Disease Cooperative Study (ADCS) to predict cognitive impairment as measured by clinical evaluation [15] and the animal fluency cognitive test score as a metric [16]. In [16], the best performing regression model was a second-order model that combined speaking rate and formant features, resulting in a correlation (R) of 0.61 and a root mean squared error (RMSE) of 5.07 with respect to a 9-34 score range. Vocal features provided a reduction by about 30% in RMSE from a baseline (mean score) in predicting cognitive performance derived from the animal fluency assessment. In [15], a small set of six vocal features showed promise for cognitive impairment classification. A SVM classifier with 10-fold cross validation obtained an EER of 13.5%.

In this paper, new combinations of articulatory and phoneme-dependent rate features from the ADCS database of remotely collected speech are explored for CI detection. These feature combinations are computed across several different cognitive testing protocols within ADCS, including the East Boston Immediate testing protocol. This protocol was not used in [15] or [16] and the best-performing system tested in this work is derived from the East Bostion Immediate (EBi) testing protocol. With a rigorous, repeatable, out-of-sample feature selection and cross-validation approach, this model achieves prediction accuracy, characterized by the area under the receiver operating characteristic curve (AUC) [24] of $\text{AUC} = 0.74$, using all feature types from EBi audio. When these features are combined with animal fluency cognitive test scores, also a component of the ADCS, the model yields a result of $\text{AUC} = 0.77$. Because multiple, different approaches to validation are used in the prior art described above, it is difficult to compare results on an equal footing. A general analysis of the validation methods used in this and other speech-based CI research is developed here to provide context for interpreting performance estimates. Each of these methods is applied to the dataset from this work to show a specific instance of how a validation approach can significantly change performance estimates, as a consequence of over-fitting.

The rest of the paper is organized as follows. Section II details data collection procedures and participant demographics in the ADCS, the source of the speech and clinical data used in this work. In Section III, signal-processing methodologies for obtaining speech features from phoneme-based and pseudo-syllable-based speaking rates, pitch variability, and formant frequency correlation structure are described. In Section IV, the evaluated feature sets are described along with the supervised learning and CI detection methods used in this work. In section V, the performance evaluation approach and results are presented. Section VI compares different cross-validation methodologies for feature selection and supervised learning. Finally, Section VII closes with conclusions and projections toward future work.

II. DATA DESCRIPTION

The Alzheimer's Disease Cooperative Study (ADCS) coordinated a 4-year longitudinal data collection, entitled ''Multi-Center Trial to Evaluate Home-Based Assessment (HBA) Methods for Alzheimer's Disease Prevention Research in People over 75 Years Old.'' The purpose of the data collection was to enable evaluation of different technology platforms for administering home-based assessments outside of clinic visits. All participants completed comprehensive in-person medical and neurological diagnostic evaluations at study baseline. Eligible participants were randomized to three different study arms, one of which was a speech-enabled, computer-automated telephone system using interactive voice response (IVR) technology [25]. From this study arm, a 214-subject audio database of audio was compiled. The sample comprises 72 male and 142 female participants.

Speech recordings (sampled at 8 kHz) were collected over standard home telephones either quarterly (50% of subjects) or annually (50% of subjects).

No personally identifying information is available through the ADCS Data Core. All information is linked through anonymous participant ID numbers. Individually identifiable information is stored at the investigative sites and secured to protect participant identities in accordance with the oversight of the institutional IRBs. The research here does not apply to U.S. Department of Health and Human Services 45 Code of Federal Regulations part 46 and is not considered human research. All data used was stored and handled accordingly.

A. CLINICAL AND REMOTE DATA COLLECTION PROCEDURES

Clinical evaluation consisted of a medical examination, a neurological examination with specific questions about memory complaint, and a neuropsychological battery taken from the Uniform Data Set of the National Alzheimer Coordinating Center. The tests included: Logical Memory, Immediate and Delayed; Digit Span: Forward and Backward; Category Fluency: Animal and Vegetable; Trail Making Test: Parts A and B; Digit Symbol Substitution; and Boston Naming Test. In addition, the clinician administered a 24-item ADCS-MCI Activities of Daily Living and used this assessment battery to 1) exclude participants with dementia at the beginning of the study, and 2) categorize eligible participants as normal or MCI based on evidence of memory impairment from the interview and neuropsychological evaluation. A similar process occurred at the end of the study for those still participating. CI assessment occurred additionally during the four-year study if a participant reported changes deemed significant enough to warrant a follow-up exam.

Cognitive functioning in HBA participants was assessed remotely using a variety of tasks including the animal fluency test and the East Boston memory test. In the animal fluency memory task (abbreviated as AF), participants list as many animals as possible during a one-minute interval. In the East Boston memory test, participants are told a story and asked to summarize the content of the story immediately (East Boston Immediate or EBi) after hearing it and again after a specific delay (East Boston Delayed or EBd) (Table 1). All speech data used in this study to predict on-site, clinical cognitive impairment diagnosis is derived from audio recordings of these tasks. The subset of observations used was selected based on two criteria: 1) temporal proximity of the audio collection to a clinical cognitive assessment, and 2) no evidence of confounding factors such as alcoholism or depression in the participant.

B. PARTICIPANT DEMOGRAPHICS AND DATA PREPROCESSING

Participants lacking audio recordings in sufficient temporal proximity to the clinical evaluation were not included in this work. 285 audio samples from 168 participants were collected within three months of a clinical evaluation. Several audio samples were not suitable for analysis. For example, in one AF audio sample a woman stops in the middle to answer her door and never finishes the task. 284 observations remain from 167 participants after removing those of poor quality. 21 observations are from study participants diagnosed with some form of cognitive impairment: dementia (2 observations), Amnestic MCI single domain (14 observations), and Amnestic MCI multiple domain (5 observations) (Table 2). Dementia criteria for this study are for Alzheimer's Disease-based dementia. In amnestic MCI, memory is significantly impaired. In single domain (SD) amnestic MCI only memory is impaired, whereas in multiple domain (MD) amnestic MCI, memory and at least one other cognitive ability is affected (e.g., visual-spatial skills and/or executive functioning). 263 observations are from participants diagnosed with normal cognitive functioning status. In this work, observations were binned into two diagnostic classes. The *normal* class consists of participants with neither an MCI nor a dementia diagnosis. The *cognitive impairment* (CI) class consists of participants diagnosed with either some form of amnestic MCI or with dementia (Table 2).

ADCS data was collected in a longitudinal study with multiple observations of many participants. There are five observations of one participant, four observations of seven

TABLE 1. Cognitive test audio names and descriptions.

Cognitive Test Audio Name (acronym)	Description
East Boston Immediate (EBi)	Participants are told a story and asked to summarize the content of the story immediately
East Boston Delayed (EBd)	Participants are told a story and asked to summarize the content of the story after a delay (approximately 10 minutes)
Animal Fluency (AF)	Participants list as many animals as possible during a one-minute interval

TABLE 2. Distribution of participant clinically determined cognitive status in the data. The cognitive impairment class (CI) includes all participants with dementia and amnestic MCI. Single domain (SD) means only memory is impaired in amnestic MCI while multiple domain (MD) means memory and at least one other cognitive ability is affected.

	Normal	CI		
		Amnestic MCI/SD	Amnestic MCI/MD	Dementia
Count	260	15	5	2

participants, three observations of seventeen participants, two observations of 58 participants, and one observation of 84 participants. 12 of the 83 participants with multiple observations transition from normal to MCI within the study, with one participant transitioning further to dementia. Cross-observational (i.e., cross-session) averaging is one way to address the complex missing-data profile of this study and transform it into data resembling a cross-sectional study with independent data for inductive learning. Longitudinal data collected from individuals can also have additional structure relative to cross-sectional data due to within person correlations, such as baseline cognitive ability and non-random changes that occur in an individual over time due to clinical progression of disease and aging. Such correlations, in addition to providing feature information, can influence classifier performance. Under these conditions it is challenging to isolate the predictive power of a feature set [26]. Factors such as short time-scale changes in stress or fatigue, which can occur independently of long-term cognitive status changes, contribute to intersession speech variability in a participant [27]. These factors are a source of noise that can degrade CI classification accuracy.

Cross-observation averaging of speech features obtained from multiple observations with an unchanging CI diagnosis also mitigates these effects. For example, for a participant with four normal clinical assessments, averages of speech features obtained from the four audio collections are used in the feature vector for that participant. Features across sessions where a transition from normal to MCI or from MCI to dementia takes place are not averaged, under the assumption that these speech samples were generated from different cognitive states. This results in 12 participants in the dataset with multiple observations across different CI diagnoses (constituting 25 total observations in the data). The remaining 155 participants have either one observation or one averaged observation. The resultant dataset includes 20 CI observations and 160 normal observations for model testing and training.

Age, education level and sex are potential confounding factor in this study. CI and dementia are known to increase with age in the elderly [28], and speech features could also change with these covariates. To be of clinical use, speech features need to provide additional predictive capability for CI beyond that predicted with these demographic variables. Age and education level histograms of the participants in this dataset demonstrate their influences on CI diagnosis. Figure 1 shows frequencies of the age and education demographic variables for normal (tan) and CI (green) participants in split bean plots. For the ADCS data, the histograms do not appear drastically different in shape and there is no strong bias toward higher ages or lower education levels in the CI population.

The effect size of these demographic variables on CI is quantified in Table 3 with Cohen's d [30] for the age and education variables and the log odds ratio, appropriate for binary variables, for the gender variable. Given the sample sizes of CI $=$ 20 and normal $=$ 160, age and education

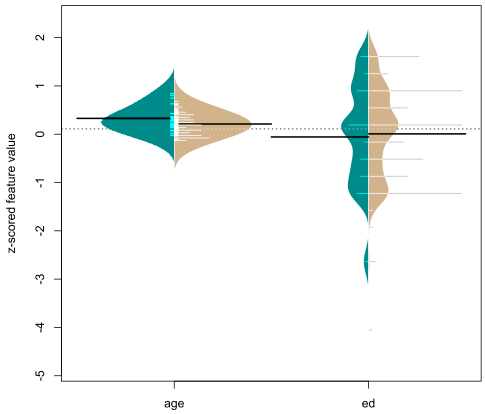

FIGURE 1. Split bean plots showing z-scored feature value frequencies of age and education (ed) covariates for CI (green) and Normal (tan) observations in the averaged ADCS data set. The solid black lines are mean values for each covariate for CI and Normal observations while the dotted line represents an overall average for all covariates for both Normal and CI observations. Thin cyan and white lines indicate feature values with observations. Green and tan envelopes depict the shape of the frequency distribution over all observed feature values.

TABLE 3. Effect size metrics quantifying how much age, gender, and education differ for the normal vs. CI class. Corresponding 95% confidence intervals and p-values are shown as well. As suggested in the split bean plots in Figure 1, there is negligible information in these covariates from this dataset for CI detection.

	Age	Education	Gender
Effect Size	-0.13 (Cohen's d)	-0.06 (Cohen's d)	0.23 log(odds ratio)
95% Confidence Interval (p-value)	[-0.6 , 0.34] (0.59)	[-0.5 , 0.4] (0.79)	[-0.8 , 1.3] (0.67)

exhibit negligible effect sizes that are not significant at the 5% or 10% level. Gender has a slightly larger effect size that is also not significant at the 5% or 10% significance level. In light of these weak, insignificant associations with CI, these covariates were not accounted for in subsequent models including other features.

III. VOCAL FEATURE EXTRACTION

The speech features used in this study are based primarily on phonemic, pseudo-syllable, and articulatory measures. A motivation behind investigating this particular suite of features is that neurophysiological changes associated with dementia affect motor timing and coordination and therefore involve the disruption of articulatory control and kinematics [9], [10]. More specifically, the approach explored here is based on the hypothesis that general psychomotor slowing due to dementia affects speaking rate and articulatory coordination. Two recent methods investigate these disruptions in articulation and speaking rate: a phoneme-based rate measure (including pause information) [18] and articulatory coordination using formant-track cross correlations [29].

These types of feature sets have also been found effective in detecting depression and Parkinsons disease, which can be characterized by psychomotor retardation [20], [31], and thus potentially represent a common feature basis for certain neurological impairments. Before averaging over longitudinal observations (i.e. audio collection sessions), the 87 speech features obtained from the AF test, the Ebi test, and the EBd test, are extracted for each observation, for a total of 261 features per observation. These are then averaged over longitudinal observations as described above.

A. AVERAGE SPEECH-RATE FEATURES

Measures of speech rate are derived from the counts and durations of individual phonemes. These are derived from a phone recognition algorithm based on a Hidden Markov Model (HMM) approach, reported with a phoneme-recognition accuracy of about 80% [32]. This model was trained with English speech but not elderly speech in particular. Accurate phoneme classification, however, is not as important as consistency in the phoneme labeling and accuracy in the phoneme boundary demarcations. It is interesting to note that the same phoneme recognition algorithm was used successfully as a basis for estimating major depressive disorder in German speech [19] and Parkinson's disease severity in Spanish speech [20], despite the fact that the algorithm is intended for phoneme recognition in English speech.

The number of phonemic speech units per second over the entire duration of a single participant's session is used to compute two features.*Speaking rat*e refers to the average phoneme rate with pauses included, whereas *articulation rate*refers to average phoneme rate with pauses excluded. Speaking rate and articulation rate measures are also based on pseudo-syllables [33] and are computed using an automatic phoneme recognition system that, as above, first detects individual speech sounds. These phonemes are then combined such that each vowel forms the nucleus of its own segment, with all of the preceding consonants grouped with it. For example, ''V,'' ''CV,'' and ''CCV'' are all valid pseudo-syllables.

B. PHONEME-SPECIFIC FEATURES

These include the average duration and the total count of 40 individual phonemes. The phoneme dictionary includes 'sil', the so-called silence phoneme, which is used to estimate pauses between speech segments. The large amount of unobserved phoneme data in the ADCS audio was approached differently here compared to [15]. In [15], analysis was restricted to the subset observations containing a positive number or a zero for the average duration of 20 phonemes. All phonemes that lacked one or more observations in at least one audio sample, depicted by 'NaN', were discluded from the study. To use the entire set of 40 phonemes in the feature selection process in the current work, an average duration of zero is attributed to a phoneme if it is unobserved in an audio sample.

C. FORMANT FREQUENCY CORRELATION STRUCTURE (Xcorr)

To assess coordination of speech articulators, the dynamics of vocal resonances (formant frequencies) are measured based on the structure of correlations of formant tracks. This feature extraction approach has been successfully applied to vocal signals to predict symptom severity in major depressive disorder [19]. A detailed description of this feature analysis approach, in the context of epileptic seizure prediction from multichannel EEG, is provided in [29]. In summary, the approach computes channel-delay correlation matrices from the first three formant tracks. Each matrix contains correlation coefficients, computed at multiple relative time delays, between formant tracks obtained from the audio samples. The formant correlation structure is characterized using the rank-ordered matrix eigenspectra. Changes in the coupling strengths among the formant tracks cause changes in the eigenvalue spectra of the channel-delay matrices. For this work, matrices are computed with four different sub-frame intervals of 1, 3, 7, 15 for the four scales, with 10 time-delays used per scale. The first principal component of the concatenated eigenvalue spectra from the multiple time scales is used as the feature.

D. PITCH VARIABILITY

Pitch estimation is performed in voiced regions using a sinusoidal-based algorithm [33]. For each speech sample in the database, the pitch variance is estimated as the mean-squared pitch deviation from the mean.

IV. FEATURE SELECTION AND CLASSIFICATION

The feature selection process in [15] involved some manual tuning and was performed using information from the entire dataset. The implications of using information from the entire dataset in feature selection is addressed in section VI. This work uses a more automated, generalizable pipeline, from feature selection to performance evaluation, enabling reproducibility. For each of the three speech tasks, EBd, EBi and AF, the following feature selection protocol is used. First, features are grouped into five categories based on the type of feature in the group (Table 4). This grouping was done to enable a global perspective on the effect of feature type on CI detection performance for exploratory insights. Cohen's d effect size measure is used to rank order the effect-sizes of features within a feature type if it includes more than one feature (e.g. the forty average phoneme durations, the forty phoneme counts and a global feature set including six of the average features described above). For feature selection, the most discriminative feature out of a given feature type that has more than one feature is chosen.

The total number of features per session sums to 87. This step is done on training data only. Selected features from the training data are then used to learn a support vector machine (SVM) and a Gaussian classifier (GC), also using training

TABLE 4. Feature groups used in the feature selection process. The feature with the largest effect size, as measured by Cohen's d, from fset3, 4 and 5 is selected. fset1 is a fixed scalar that is not selected and fset2 is the first principal component of the formant-based, cross-correlation features, where PCA is done on training data only. The total number of features per session sums to 87.

Feature Set (fset)	Description
1	Animal fluency score
2	First principal component of formant-based, cross-correlation features
3	Average pitch, pitch variability, pseudo syllable-based articulation rate, pseudo syllable-based speaking rate, phoneme-based articulation rate and phoneme-based speaking rate
4	Average phoneme durations for 40 standard English phonemes
5	Total phoneme counts for 40 standard English phonemes

data only. This classifier is then applied to unseen data to produce predictions.

V. RESULTS AND DISCUSSION

Using the above vocal features, the primary goal is to produce accurate performance estimates while also gaining exploratory insights that can be used in a follow-up confirmatory analysis with new data. Therefore, CI detection accuracy is measured as a function of three dimensions: audio type (EBd, EBi and AF), feature set type (fsets1-5) and classifier type (GC vs. SVM). The leave-pair-out cross-validation (LPOCV) design is conservative in that a sample from the smaller CI class (20 observations) is always held out in the test data. This avoids optimistically biased performance estimates resulting from selecting only the larger normal class (160 observations) in most of the test samples [37].

While optimistic biases are unlikely in this analysis, a negative bias, indicated by AUC values below 0.50, was present. This is very likely due the small size of the ADCS dataset and the fact that the Normal/CI class distribution is imbalanced. Parker *et al*. (2007), show that AUC estimates from small, imbalanced, low-signal datasets, such as the ADCS dataset, suffer from negative bias due to stratification. Therefore, our AUC estimates are likely smaller on average than if derived from a larger, balanced dataset sampled from the same population. While other performance metrics are not as sensitive to stratification bias, we use AUC in this study to compare our results with those in [17]. Performance results from prior art derived from small datasets potentially suffer from a similar bias, given that speech-based CI detection appears to be a generally low-signal regime.

Significance of differences between AUC values were determined using a *t-test* for differences between sample means computed with standard errors (SE) defined in [24]. Results are summarized in Table 5a for the Gaussian classifier and in Table 5b for the SVM. Gray boxes correspond to AUC significantly better than random, at the 5% level.

A. RESULTS FROM THE PERFORMANCE EVALUATION

Two feature set/audio combinations stand out for both the GC and SVM. fset3 from AF audio achieves an AUC of 0.73-0.75 for both classifiers. The selected feature from fset3 that produced this high classification performance in the leave-pair-out cross validation was phone rate for all phones. The performance improvement obtained by adding the AF score is not significant at the 5% level. This feature also had a strong effect size computed over the entire dataset (testing and training data): Cohen's d $= -0.79$, 95% confidence interval $= [-1.3, -0.3]$.

fset2+fset3+fset4+fset5 for the EBi audio achieve an average AUC of 0.74-0.75 for both the SVM and GC in the leave-pair-out cross validation. AUC value increases to 0.77 if the animal fluency score (fset1) is combined with the speech features for the GC but decreases to 0.72 for the SVM. The first principal component of the EBi cross-correlation feature (fset2), determined from training data and computed for test data in the LPOCV, as well as the animal fluency cognitive test score (fset1), are not selected using the Cohen's d metric in the cross-validation trials. However Cohen's d values for them are computed here over the entire dataset (testing and training data) to quantify their effect size for future confirmatory models built from this dataset. They demonstrate a moderate effect size not significant at the 5% level using the Bonferroni correction for the 261 independent feature comparisons in our feature selection process (Cohen's d $= 0.49$), 95% confidence interval $= [0.02, 0.96]$ and Cohen's d $= -0.41$, 95% confidence interval $= [-0.9, 0.1]$), respectively. Features from fsest3, fset4 and fset5 are selected using Cohen's d in the LPOCV trials. Unlike the AF audio, in EBi audio, the selected feature from fset3, pitch variance, does not have a strong, significant effect size (Cohen's d $= -0.24$, 95% confidence interval $= [-0.7, 0.2]$) when computed using the entire dataset. However, using the entire EBi audio dataset, the average duration of the <th> phoneme and the count of the <eh> phoneme were selected in the feature selection process from fset5 and fset4, respectively, and have strong, significant effect sizes, Cohen's d $= 0.86$, 95% confidence interval $= [0.4, 1.3]$, and 0.785, 95% confidence interval $= [0.3, 1.3]$, respectively.

Figure 2 shows CI (green) and Normal (tan) z-scored distributions of these features, taken from the entire dataset, as split bean plots. Cyan and white lines indicate observed feature values, green and tan contours indicate the general shape of the frequency distribution of the feature observations, black, solid lines indicate mean feature values for CI

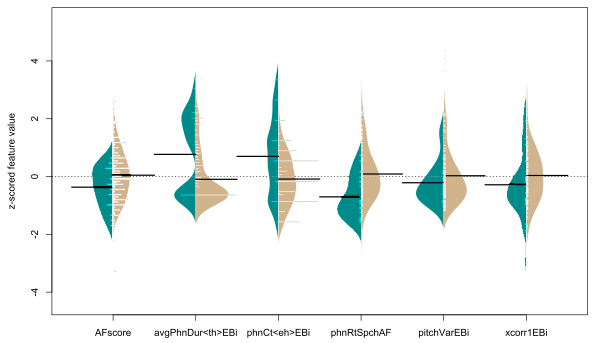

FIGURE 2. Split bean plots of z-scored feature distributions over all data for features from the two best performing audio/feature sets: AF fset3 and EBi fset1+fset2+fset3+fset4+fset5. Black, solid lines are mean values of each feature distribution while the dotted line is the mean value of all features. From the AF audio, the average phoneme-based speaking rate from all phonemes (phnRtSpchAF) was selected from fset3 as the feature with the strongest effect size for CI detection. From the EBi audio, the average duration of the <th> phoneme (avgPhnDur<th>) and the count of the <eh> phoneme (phnCt<eh>EBi) were selected as having the strongest effect size from fset4 and fset5, respectively. Generally, Cohens' d above |0.7| is considered a strong effect size. The Cohen's d for these features is strong and statistically significant at the 5% level. In contrast, in EBi audio, the Cohen's d for the pitch variance is weak and not significant, indicating that fset3 from EBi could be removed in future CI prediction models. Although the animal fluency score (AFscore), from fset1, and the formant-based cross correlation feature (xcorr1EBi from), from fset2, are not subjected to feature selection, the feature histograms for these features are also shown. They each demonstrate a statistically insignificant, moderate effect size (Cohen's d = 0.48, 95% confidence interval = [0.0, 1.0] and Cohen's d = −0.41, 95% confidence interval = [−0.9, 0.1] for xcorr1 and AFscore, respectively).

and Normal observations and the black dotted line indicates a grand mean over all feature values plotted. The count of <eh> and average duration of <th> in the EBi audio is larger on average for CI relative to normal observations. In contrast, the phone rate over all phonemes from the AF audio is on average lower for the CI relative to normal observations. Also of note is the bimodality of the average duration of <th> for the CI observations. One cluster of CI participants resembles normal participants in average <th> duration. The other cluster demonstrates <th> durations much longer on average than normal participants.

B. DISCUSSION

Interestingly, at the 5% significance level, the AF score improves AUC using the GC but decreases AUC with the SVM with all feature types from EBi. In fact, with the GC, all speech feature sets perform better when paired with the AF score, at the 5% confidence level, except the set of global speech features from the AF audio in fset3. The feature selected from this set in all of the LPOCV trials and when all data is used is the phoneme-based speaking rate. The difference between AUC values from the model using this speech feature with and without the AF score is not statistically significant for GC or SVM.

This finding is consistent with observations in [16], in which strong correlations between phoneme-based and pseudo syllable-based rates and animal fluency scores were discovered at the 5% significance level, 0.57 and 0.58, respectively. Such correlations entail mutual information between the AF score and the phoneme-based speaking rate from the AF audio. However, the effect size for CI detection of the phoneme-based speaking rate from the AF audio is large but still not significant the 5% level, with the Bonferroni correction, Cohen's d = −0.79, 95%, confidence interval = [−1.3, −0.3], while that of the AF score is medium and not significant at the 10% level, Cohen's d = −0.41, 95%, confidence interval = [−0.9, 0.1], cf. feature histograms in Figure 2. Future confirmatory models for CI detection may benefit from using the phoneme-based speaking rate from AF audio instead of the AF score. Such models could be entirely speech-based, eliminate the need to manually or automatically score the AF test and potentially have better discrimination power.

Some of the speech features sets under the SVM actually perform worse at the 5% confidence level when paired with the AF score. This variability is possibly because the SVM using a radial basis kernel is a more complicated classifier than the Gaussian Classifier and may be less robust with small datasets commonly found in speech-based MCI/AD detection. A simple, linear kernel may be a better choice for the SVM in limited data regimes. It is also noteworthy in these results that features from the East Boston Immediate (EBi) audio strongly outperform those from the East Boston Delayed (EBd) audio, which was used in [15] and [16]. Therefore, in this dataset, speech from very short-term/working memory tasks carries a stronger signal in the feature dimensions we tested than that collected in the delayed task measured approximately ten minutes after the story was heard. There may be implications in this finding for what modes of psychomotor functioning, as manifested in speech about contents of working memory, are most strongly implicated in MCI and DAT and worthy of further study.

The general impact of averaging across longitudinal data is difficult to assess for this study because the majority of participants (96) had only one observation or two independent observations that were not averaged if they transitioned to CI during the study. Additionally, 58 participants had only two observations- one at the beginning and one at the end of the study. Conducting a rigorous analysis on the effects of cross-observation averaging on classification results would require uniform, repeated sampling profiles for each subject so that large differences in missing data per subject would not be a hidden, confounding factor in the analysis.

For the same reason, learning personalized models for CI prediction from repeated data, though a compelling approach, would be very difficult with this data. Most participants have only 1 or 2 observations and participants with multiple observations often do not transition to CI during the study, leaving no training samples for the CI class. Only one participant has five and seven have four observations. Of these, all but one participant, who transitioned from MCI to dementia, had only one or no observations taken in the CI condition. Learning meaningful, individualized models with no or so

TABLE 5. **(a). Performance results for the Gaussian classifier on the various audio/feature-set permutations. Pairing speech features with the AF test score improves performance for all combinations at the 5% significance level, with the exception of fset3 from the AF test. Gray boxes are feature sets that perform better than random at a 5% significance level. AUC values below 0.50 are common when using small, imbalanced datasets for classification because stratification bias in such data causes AUC estimates to be lower than expected (see Parker *et al.* (2007)). AUC values presented here are therefore likely conservative estimates of the true AUC that would result from a balanced and larger dataset sampled from the same population. (b). Performance results for the SVM classifier on the various audio/feature-set permutations. Pairing speech features with the animal fluency test score impacts performance in unpredictable ways with the SVM. Both the GC and SVM show best performance with EBi audio and the full feature set and AF audio and fset3. Gray boxes are feature sets that perform better than random at a 5% significance level. Note, compared to the GC, fewer feature sets perform significantly better than random at the 5% significance level.**

GC AUC/ SE Table	fset2	fset2 + fset1	fset3	fset3+ fset1	fset4	fset4 + fset1	fset5	fset5+ fset1	fset2+fset3+ fset4+fset5	fset2+fset3+ fset4+fset5+ fset1	fset1
EBi	**0.49**/ 0.07	**0.52**/ 0.07	**0.32**/ 0.07	**0.44**/ 0.07	**0.56**/ 0.07	**0.69**/ 0.06	**0.59**/ 0.07	**0.68**/ 0.06	**0.74**/ 0.05	**0.77**/ 0.05	**0.54**/ 0.07
EBd	**0.52**/ 0.07	**0.55**/ 0.07	**0.38**/ 0.07	**0.54**/ 0.07	**0.58**/ 0.07	**0.63**/ 0.06	**0.37**/ 0.07	**0.52**/ 0.07	**0.46**/ 0.07	**0.50**/ 0.07	
AF	**0.53**/ 0.07	**0.62**/ 0.07	**0.74**/ 0.05	**0.73**/ 0.05	**0.43**/ 0.07	**0.53**/ 0.07	**0.49**/ 0.07	**0.60**/ 0.07	**0.59**/ 0.07	**0.67**/ 0.06	
All Audio	**0.64**/ 0.06	**0.67**/ 0.06	**0.53**/ 0.07	**0.65**/ 0.06	**0.61**/ 0.06	**0.66**/ 0.06	**0.51**/ 0.07	**0.59**/ 0.07	**0.57**/ 0.07	**0.62**/ 0.07	

(a)

SVM AUC/ SE Table	fset2	fset2 + fset1	fset3	fset3+ fset1	fset4	fset4 + fset1	fset5	fset5+ fset1	fset2+fset3+ fset4+fset5	fset2+fset3+ fset4+fset5+ fset1	fset1
EBi	**0.48**/ 0.07	**0.55**/ 0.07	**0.32**/ 0.07	**0.41**/ 0.07	**0.66**/ 0.06	**0.57**/ 0.07	**0.64**/ 0.06	**0.52**/ 0.07	**0.75**/ 0.05	**0.72**/ 0.05	**0.58**/ 0.07
EBd	**0.52**/ 0.07	**0.55**/ 0.07	**0.38**/ 0.07	**0.53**/ 0.07	**0.51**/ 0.07	**0.59**/ 0.07	**0.50**/ 0.07	**0.55**/ 0.07	**0.46**/ 0.07	**0.50**/ 0.07	
AF	**0.53**/ 0.07	**0.62**/ 0.06	**0.74**/ 0.05	**0.75**/ 0.05	**0.49**/ 0.07	**0.38**/ 0.07	**0.48**/ 0.07	**0.54**/ 0.07	**0.58**/ 0.07	**0.60**/ 0.07	
All Audio	**0.45**/ 0.07	**0.58**/ 0.07	**0.54**/ 0.07	**0.59**/ 0.07	**0.54**/ 0.07	**0.50**/ 0.07	**0.53**/ 0.07	**0.52**/ 0.07	**0.64**/ 0.06	**0.62**/ 0.06	

(b)

few examples of CI and such little data would not be possible. In follow-up confirmatory studies with larger datasets collected under more controlled conditions, these important issues should be studied in more detail.

VI. CROSS-VALIDATION AND EVALUATION PIPELINE COMPARISONS

The LPOCV results in Section V were obtained using an automated, filter method of feature selection [34] *within* each cross-validation fold based on the Cohen's d measure of effect size computed on training data *only*. One goal of this work is to understand the degree to which feature selection and cross-validation methodologies result in biased estimates of CI detection performance, a topic considered in some depth in [38]. Most prior work in speech-based CI detection selects features *outside* the cross-validation loop, using *both* training and testing data. Even though cross-validation is subsequently done in the statistical modeling stage, the selected features still possibly engender an optimistic bias in the resulting performance estimates because the selection was already done using test data. In addition, in many cases, results from cross-validation studies were used to further refine and select the sets of features and the statistical modeling parameters that optimize performance. In [21], the features computed using the entire dataset are selected by filtering on single tailed p-values.

In [22], the Mann-Whitney test is used for filter-based feature selection using features computed from all data [35]. Feature selection was further tuned using classification accuracy estimates from cross-validation tests in [22]. Roark *et al.* [17] selected features based on *t-values* computed from the full dataset. In [15], we use a wrapper-based feature selection method that selects feature by maximizing AUC over CV trials in our SVM model [36]. This approach uses

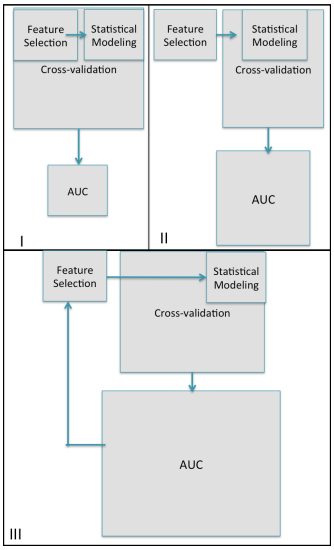

FIGURE 3. Schematic of three possible pipelines including feature selection (FS), supervised learning (SL) and cross-validation (CV). Pipeline I is the recommended method in which both FS and SL occur on training data only. No information from CV performance estimates, such as AUC, are used in either FS or SL. In pipeline II, FS is done over all data and not subjected to CV. Pipeline III also does FS over all data but additionally uses CV performance to influence FS. Pipelines II and III are likely to produce optimistically biased CV performance estimates due to over-fitting that do not generalize well.

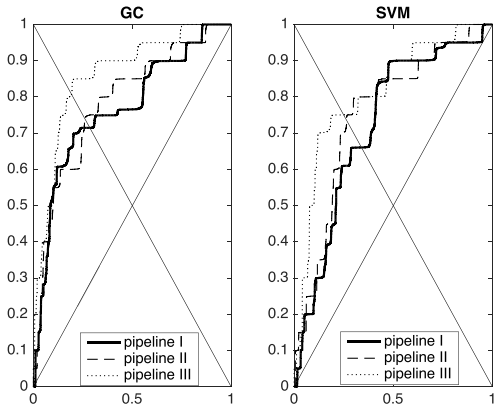

FIGURE 4. ROC curves for various cross-validation methodologies on the EBi audio using all features in the feature selection process. Clear performance decreases result from progressively removing in-sample information from the feature selection process. Pipeline II includes all data in the feature selection process. Pipeline III includes all data in feature selection and additionally tunes feature selection using the AUC performance metric. All known prior art in speech-based CI detection uses pipelines II and III, indicating decreased generalizability of performance estimates therein relative to results derived from pipeline I.

information from the entire dataset contained in the AUC performance metric. In Figure 3, a schematic is shown of three standard approaches for doing feature selection, statistical modeling and performance evaluation using cross-validation. Processing pipelines based on these different approaches are described in the next section.

A. GENERAL PIPELINE COMPARISONS

In Figure 3., pipeline I is depicted as the approach yielding the most accurate assessment of performance generalizability. This pipeline uses no information from the performance metric, e.g. area under the ROC curve (AUC), or from out-of-sample data in the feature selection step. To get a more realistic assessment of the discrimination power of fset3 from AF audio and fset1+fset2+fset3+fset4+fset5 from EBi audio and how they will generalize, the feature-selection/classification/cross-validation approach generating the results in Table 5 was based on pipeline I. This process is repeated multiple times for different testing and training partitions of the data using exhaustive leave-pair-out cross validation, to derive an estimate of CI detection accuracy on unseen test data. To our knowledge, no previous work in speech-based CI detection adheres to this rigorous methodology. Pipeline II, on the other hand, removes the feature selection process entirely from the cross-validation loop. Feature selection in this scenario is done on the entire data set (including test data) and only the model parameter-learning step is subjected to cross-validation. Pipeline II has the potential to produce optimistically biased performance estimates because feature selection is not subjected to cross-validation. Pipeline III further compounds this risk by using the cross-validation performance estimates to further influence feature selection in a spiral development cycle. All known prior art in speech-based CI prediction uses pipeline II or III to some extent. In the next section, experimental results are shown that quantify the biasing of performance estimates based on using pipeline II and III.

B. PIPELINE COMPARISONS ON ADCS DATA

To place the speech-based CI detection results in [15], [17], [21], and [22] in a broader context, this section explores the hypothesis that using a version of pipeline I, in which feature selection is done inside the cross-validation loop, would affect performance relative to pipeline II, in which feature selection is done outside of the cross-validation using all data, and pipeline III, in which all data as well as CV performance estimates inform feature selection, is tested. The results of this test done on the data used in this study are shown in Figure 4. For both classifiers, performance quantified by AUC, decreased for pipeline I compared to pipelines II and III at a 5% significance level. These results highlight the importance of placing feature selection fully in the cross-validation loop to obtain accurate performance estimates that generalize. Furthermore, using CV performance estimates, such as AUC or accuracy, to tune feature selection inflates performance estimates further, as would be expected.

Hastie *et al.* [34] demonstrate that the best way to do cross-validation places the feature selection step in the cross-validation loop and is done only using information in the training data. The work presented here demonstrates a particular instance of why this approach is important.

Performance estimates are inflated when information from the entire dataset is used in the feature selection step. Prior work in [15], [17], [21], and [22] benchmarks performance with different metrics, EER and AUC, and uses speech from different databases, often from different languages. These factors make comparisons among results on par impossible. However, based on results in Figure 4 and because all known publications in the area of CI detection use a version of pipeline II or III, it is reasonable to wonder if performance estimates quoted therein will generalize to unseen data. Yu *et al.* [15] and Satt *et al.* [21], [22] clearly use pipeline III to generate results. The approach in [17] uses pipeline II but it is not clear whether presented performance estimates are selected from a larger set based on AUC or whether they constitute exhaustive, unfiltered results.

Sometimes, generalizable performance estimates are not the goal in a particular study. For example, in exploratory analyses, producing generalizable results is not the objective. In these cases, pipelines II and III are reasonable approaches provided they are framed in their proper context, e.g. as methods for exploring the data rather than producing accurate estimates of performance generalizability. This would have helped the reader interpret results more clearly in [15], [17], [21], and [22]. It also is noteworthy that the best performing models from Tables 5a and 5b are not only characterized using a rigorous cross-validation methodology but that they are also parsimonious. The EBi audio model, using all feature types, has five features and the AF audio model, using the feature from fset 3 with the largest effect size, has just one feature. All best performing models from [15], [17], and [21] include more features (eighteen in [17], twenty in [21] and six in [15]), which is an additional indication of over-fitting.

VII. CONCLUSION

This paper details several results on CI detection from remotely collected audio recordings. In the first, on an expanded dataset with more candidate features relative to [15], speech features from audio remotely collected for cognitive testing are demonstrated to provide information for CI detection. The best feature set achieves an AUC of 0.77, which is different from random at the 5% significance level. Furthermore, the detection pipeline, a version of pipeline I (strict cross-validation), is automated, reproducible and does feature selection within CV. The second is that cognitive test scores can be combined with speech features from the cognitive testing audio to provide improved CI detection over either one in isolation at a 5% significance level with a Gaussian classifier. The performance animal fluency score alone is significantly better than random with both classifiers but does not have a strong effect size (Cohen's $d = 0.38$).

In this dataset, the animal fluency score does not provide as much detection power as the best speech features, however. This finding is noteworthy because DAT and MCI diagnosis protocols commonly leverage fluency tests. Future, completely speech-based approaches to CI detection using the phoneme rate, including pauses, from AF cognitive test audio rather than the AF score, appear promising. This feature demonstrates a stronger effect size for CI detection than the AF score and also shares information with the score, quantified by correlation in [16].

The final narrative is a proscriptive one. Performance estimates from CV need to be framed clearly and properly in terms of 1) what pipeline was used to generate them and 2) size and stratification characteristics of the dataset, in order to ensure proper interpretation. If estimates derive from pipelines II or III, this should be clearly stated along with caveats on the expected generalizability of performance estimates derived from these pipelines. Although it is natural to tune performance to some extent using AUC or EER, the potential effect this has on generalizability must be recognized and stated clearly, allowing results to be interpreted in their proper context. If performance estimates derive from small, imbalanced and low-signal datasets, it is possible they are negatively biased and would be more favorable if larger, balanced datasets sampled from the same population were used.

Future work will include vetting this system on unseen data to determine performance generalization. In particular, larger, balanced data sets with more CI examples will be needed to fully understand the robustness of the method and features explored in this work. If rigorous, randomized experimental design techniques are used to guide further data collection, insights into causal relationships between CI and speech characteristics will be possible, in addition to the associational relationships determined in the predictive modeling described in this work. We leveraged simple classifiers to avoid overfitting limited training data in this study. With larger data sets in future studies, more sophisticated classifiers, such as deep neural networks, could be used and would likely provide improvements in predictive performance. Further work investigating other cognitive tests with stronger effect sizes than the animal fluency scores is also of interest because the influence of the AF scores on CI detection was small in this study.

ACKNOWLEDGMENT

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

REFERENCES

[1] M. S. Albert *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dementia*, vol. 7, no. 3, pp. 270–279, 2011.

[2] Z. Shao, E. Janse, K. Visser, and A. S. Meyer, "What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults," *Frontiers Psychol.*, vol. 5, p. 772, Jul. 2014.

[3] A. U. Monsch, M. W. Bondi, N. Butters, D. P. Salmon, R. Katzman, and L. J. Thal, "Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type," *Arch. Neurol.*, vol. 49, no. 12, pp. 1253–1258, 1992.

[4] I. Midi, M. Doğan, Y. S. Pata, I. Kocak, A. Mollahasanoglu, and N. Tuncer, "The effects of verbal reaction time in Alzheimer's disease," *Laryngoscope*, vol. 121, no. 7, pp. 1495–1503, 2011.
[5] A. S. Chan, N. Butters, J. S. Paulsen, D. P. Salmon, M. R. Swenson, and L. T. Maloney, "An assessment of the semantic network in patients with Alzheimer's disease," *J. Cogn. Neurosci.*, vol. 5, no. 2, pp. 254–261, 1993.
[6] R. G. Gomez and D. A. White, "Using verbal fluency to detect very mild dementia of the Alzheimer type," *Arch. Clin. Neuropsychol.*, vol. 21, no. 8, pp. 771–775, 2006.
[7] J. D. Henry, J. R. Crawford, and L. H. Phillips, "Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis," *Neuropsychologia*, vol. 42, no. 9, pp. 1212–1222, 2004.
[8] K. E. Nutter-Upham *et al.*, "Verbal fluency performance in amnestic MCI and older adults with cognitive complaints," *Arch. Clin. Neuropsychol.*, vol. 23, no. 3, pp. 229–241, 2008.
[9] A. L. R. Adlam, S. Bozeat, R. Arnold, P. Watson, and J. R. Hodges, "Semantic knowledge in mild cognitive impairment and mild Alzheimer's disease," *Cortex*, vol. 42, no. 5, pp. 675–684, 2006.
[10] K. J. Murphy, J. B. Rich, and A. K. Troyer, "Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia," *J. Int. Neuropsychol. Soc.*, vol. 12, no. 4, pp. 570–574, 2006.
[11] J. A. Lonie *et al.*, "Lexical and semantic fluency discrepancy scores in aMCI and early Alzheimer's disease," *J. Neuropsychol.*, vol. 3, no. 1, pp. 79–92, 2009.
[12] L. J. Clark, M. Gatz, L. Zheng, Y. L. Chen, C. McCleary, and W. J. Mack, "Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimer's disease," *Amer. J. Alzheimer's Disease Dementias*, vol. 24, no. 6, pp. 461–468, 2009.
[13] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in Alzheimer patients," *Brain Lang.*, vol. 17, no. 1, pp. 73–91, 1982.
[14] J. Reilly, A. D. Rodriguez, M. Lamy, and J. Neils-Strunjas, "Cognition, language, and clinical pathological features of non-Alzheimer's dementias: An overview," *J. Commun. Disorders*, vol. 43, no. 5, pp. 438–452, 2010.
[15] B. Yu, T. F. Quatieri, J. R. Williamson, and J. C. Mundt, "Cognitive impairment prediction in the elderly based on vocal biomarkers," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3734–3738.
[16] B. Yu, T. F. Quatieri, J. R. Williamson, and J. Mundt, "Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1038–1042.
[17] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011.
[18] A. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, p. 42, 2011.
[19] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proc.3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2013, pp. 41–48.
[20] J. R. Williamson *et al.*, "Segment-dependent dynamics in predicting Parkinson's disease," in Proc. *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 518–522.
[21] A. Satt *et al.*, "Evaluation of speech-based protocol for detection of early-stage dementia," in *Proc. INTERSPEECH*, 2013, pp. 1692–1696.
[22] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2538–2542.
[23] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, Nov. 2014, pp. 65–72.
[24] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[25] M. Sano *et al.*, "Developing dementia prevention trials: Baseline report of the home-based assessment study," *Alzheimer Disease Associated Disorders*, vol. 27, no. 4, pp. 356–362, 2013.
[26] J. D. Singer and J. B. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. London, U.K.: Oxford Univ. Press, 2003.
[27] S. Ramakrishnan *et al.*, "Can a mathematical model predict an individual's trait-like response to both total and partial sleep loss?" *J. Sleep Res.*, vol. 24, no. 3, pp. 262–269, 2015.
[28] E. B. Larson, K. Yaffe, and K. M. Langa, "New insights into the dementia epidemic," *New England J. Med.*, vol. 369, no. 24, pp. 2275–2277, 2013.
[29] J. R. Williamson, D. W. Bliss, D. W. Browne, and J. T. Narayanan, "Seizure prediction using EEG spatiotemporal correlation structure," *Epilepsy Behav.*, vol. 25, no. 2, pp. 230–238, 2012.
[30] M. E. Rice and G. T. Harris, "Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*," *Law Hum. Behav.*, vol. 29, no. 5, pp. 615–620, 2005.
[31] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Proc. Interspeech. ICSA*, Portland, OR, USA, 2012, pp. 1059–1062.
[32] N. F. Chen, W. Shen, and J. P. Campbell, "A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 5014–5017.
[33] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. London, U.K.: Pearson, 2002.
[34] J. Franklin, "The elements of statistical learning: Data mining, inference and prediction," *Math. Intell.*, vol. 27, no. 2, pp. 83–85, 2005.
[35] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[36] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Inf. Sci.*, vol. 179, no. 13, pp. 2208–2217, 2009.
[37] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[38] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," *ACM SIGKDD Explorations Newslett.*, vol. 12, no. 1, pp. 49–57, 2010.

BEA YU received a B.A. degree in philosophy with a minor in linguistics from the University of Texas at Austin, and an M.S. degree in chemistry with an emphasis on theoretical quantum chemistry and an M.S. degree in applied mathematics with an emphasis on statistics from the University of New Mexico, in 2007 and 2010, respectively. Since 2010, she has been with the MIT Lincoln Laboratory, where she is currently an Associate Technical Staff Member of the Bioengineering Systems and Technologies Group. Her research covers computational modeling and simulation in a variety of domains, including statistical modeling and prediction using speech to detect various types of neurocognitive states, such as depression and age-related cognitive impairment.

JAMES R. WILLIAMSON received the B.S. degree in psychology from the University of Massachusetts, Amherst and the Ph.D. degree in cognitive and neural systems from Boston University in 1995. From 1996 to 2000, he was a Post-Doctoral Fellow and a Research Assistant Professor in cognitive and neural systems, investigating self-organizing neural network models of vision and pattern recognition. Since 2001, he has been with the MIT Lincoln Laboratory, where he is currently a Technical Staff Member of the Bioengineering Systems and Technologies Group. He works on the detection and modeling of neurocognitive changes and disorders based on analysis of speech, facial expressions, gait, eye tracking, and EEG.

JAMES C. MUNDT received the M.Sc. and Ph.D. degrees in psychology from the University of Wisconsin–Madison in 1987 and 1991, respectively. Since 1991, he has been involved in a wide range of research and development organizations, including the Vermont Alcohol Research Center, Dean Foundation for Health Research and Education, Healthcare Technology Systems, Center for Telepsychology. He is currently a Research Analyst with the State of Wisconsin Department of Health Services. He has authored over 60 peer-reviewed journal articles and conference papers. His primary research interests focus on the interface and use of technology for collecting and analyzing reliable, valid clinical outcome measures across a range of conditions ranging from Alzheimer's disease to mood, anxiety, and substance use disorders in adult and adolescent patient populations. He was a recipient of the Phase I and II Small Business Innovation Research Grants from the National Institute on Aging, the National Institute of Mental Health, and the National Institute on Alcohol Abuse and Alcoholism. He serves as a reviewer for several international journals.

THOMAS F. QUATIERI (F'99) received the B.S. degree *(summa cum laude)* from Tufts University, and the S.M., E.E., and Sc.D. degrees from the Massachusetts Institute of Technology (MIT). He is currently a Senior Member of the Technical Staff with the MIT Lincoln Laboratory (MIT LL), where he is involved in applying speech, auditory, and neuromotor science to detection and monitoring of neurological disorders and cognitive stress conditions. He holds a faculty appointment at the Harvard Speech and Hearing Bioscience and Technology Program under the Harvard–MIT Division of Health Sciences and Technology. He is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, ICSA, and the Acoustics Society of America. He was a recipient of four IEEE best paper awards in speech and signal processing and the 2010 MIT LL Best Paper Award for an IEEE TASLP article. He led the MIT LL Team that received the 2013 and 2014 AVEC Depression Challenges and the 2014 MIT LL Team Award for vocal and facial biomarkers.