



Cognitive impairment prediction in the elderly based on vocal biomarkers

Bea Yu¹, Thomas F. Quatieri¹, James R. Williamson¹, James C. Mundt²

¹ MIT Lincoln Laboratory, Lexington, Massachusetts, USA

² Center for Psych Consulting, USA

[bea.yu,quatieri,jrw]@ll.mit.edu, jmundt@telepsychology.net

Abstract

Remote, automated cognitive impairment (CI) diagnosis has the potential to facilitate care for the elderly. Speech is easily collected over the phone and already some common cognitive tests are administered remotely, resulting in regular audio data collections. Speech-based CI diagnosis leveraging existing audio is therefore an attractive approach for remote elderly cognitive health monitoring. In this paper, we demonstrate the predictive power of several speech features derived from remotely collected audio used for common clinical cognitive testing. Specifically, using phoneme-based measures, pseudo-syllable rate, pitch variance, and articulatory coordination derived from formant cross-correlation measures, we investigate the capability of speech features, estimated from paragraph-recall and animal fluency test speech, to predict clinical CI assessment. Using a database consisting of audio from elderly subjects collected over a 4 year period, we develop support vector machine classification models of the CI clinical assessments. The best performing models result in an average equal error rate (EER) of 13.5%.

Index Terms: mild cognitive impairment, motor coordination, vocal biomarkers, formant frequencies

1. 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 existing 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 patient to be either manually or automatically scored. While the linguistic content of these audio samples is leveraged for cognitive assessment, we hypothesize that non-linguistic features in the speech signal itself can also provide important information about cognitive functioning in the speaker. Leveraging speech features from audio already collected for cognitive tests is desirable for two reasons. First, this approach does not increase the burden of testing on the patient with additional tasks. Second, features from the speech signal potentially provide complimentary and novel information for clinical assessment, when paired with scores based on linguistic content of cognitive testing audio.

* This work was supported by the National Institute on Aging under grants 2U19AG010483 and 1R41AG044218. This work is sponsored by the National Institute of Health under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government

The Home Based Assessment (HBA) study of the Alzheimer’s Disease Cooperative Study (ADCS) was a 4-year longitudinal study evaluating multiple technology platforms for administering home-based assessment measures outside of clinic visits [1]. A speech-enabled, computer-automated telephone system using interactive voice response (IVR) technology was one of the in-home platforms deployed in the HBA study, and was the source of data for the analysis reported below. The audio data was collected for linguistic-content based cognitive testing. We demonstrate here that *non-linguistic content* extracted from audio during speech provides useful information for cognitive assessment.

Certain non-linguistic vocal features have been shown to change with a subject’s mental condition and emotional state, such as depression. 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 timing jitter, amplitude shimmer, and aspiration [2–7]. Discovering the coupling between speech, language, and cognitive functioning status entails determining correlations between vocal features, reflecting prosody, voice quality, and linguistic content, and varying degrees of cognitive impairment.

Several recent papers explore using speech features for predicting dementia and mild cognitive impairment (MCI). Satt et al., for example, achieve an 18% +/- 6% equal error rate (EER) for MCI/dementia prediction using speech features from Greek speech [8]. In [9], Satt et al. achieve a 20% +/- 6% EER for MCI prediction using a different dataset in French. In both cases, their audio data consist of at least three different speech tasks designed for the express purpose of extracting non-linguistic speech features to detect MCI and dementia of the Alzheimer’s type (AD). In [8], the features are selected by filtering on single tailed p-values. In [9], the authors use another filter-based feature selection method called the Mann-Whitney test. The approach we describe here, on the other hand, examines features extracted from English language audio collected for remote, linguistic cognitive testing and uses a wrapper [10] form of feature selection similar to that described in [11].

Our paper is organized as follows. In Section 2, we describe the database, data topology and our method of addressing dependencies and noise in longitudinal data. In Section 3, we describe our signal-processing methodologies for phoneme-based and pseudo-syllable-based speaking rates, measuring pitch variance and formant-coordination extraction. Section 4 describes our support vector machine classification results. Section 4 also explores the extent to which age influences cognitive status and speech features in our

database. Section 5 provides conclusions and projections toward future work.

2. Elderly speech database

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,” in order to evaluate 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 one of three study arms, one of which was a speech-enabled, computer-automated telephone system using interactive voice response (IVR) technology [1]. From this, a 214-subject database of audio was compiled from elderly subjects enrolled in the HBA study and randomized to the IVR assessment arm. The sample comprises 72 male and 142 female participants. Speech recordings (sampled at 8 kHz) were collected over standard home telephones either quarterly (50%) or annually (50%).

2.1. Previous results with animal fluency data

In the first stage of our research, we investigated the capability of speech features, estimated from paragraph-recall (the East Boston Recall task), to predict Animal Fluency assessment scores. In the Animal Fluency memory task (AF), participants list as many animals as possible during a one-minute interval. This task has demonstrated good sensitivity and specificity as a dementia screening instrument [13]. In 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.

Because most of the variability in animal fluency scores was explained by the differences in average performance between subjects, our initial objective was to determine how well we could predict, from EB vocal signals, subjects’ mean animal fluency scores. We tested multiple regression models constructed with a cross-validation procedure, in which training data is obtained purely out-of-sample (i.e., from other subjects). For each test subject, the training set consists of all sessions from the 1315-session data set belonging to different subjects. The features that proved most predictive were phoneme-based measures of speaking rate [7] and articulatory coordination indicators derived from formant cross correlation measures [8].

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 thus provided a reduction by about 30% in RMSE from a baseline (mean score) in predicting cognitive performance derived from the animal fluency assessment. A more comprehensive summary of our methodology, features, and results in this first research stage is given in [2].

2.2. Data subset selection

In this work, we explore and quantify the power of speech features from different speech tasks for clinical assessment prediction. Cognitive functioning in HBA participants was assessed remotely using a variety of tasks including AF and

EB (described above). We use audio from these tasks to predict an on-site, clinical cognitive impairment diagnosis. We select for analysis a subset of observations 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. The 183 audio samples we use from 132 patients were collected within three months of a clinical evaluation. 21 samples are from study participants diagnosed with some form of cognitive impairment: dementia (2 observations), Amnesic MCI single domain (14 observations), and Amnesic MCI multiple domain (5 observations). Dementia criteria for this study are for Alzheimer’s Disease-based dementia. 162 samples are from patients diagnosed with normal cognitive functioning status. We assign our observations 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 amnesic MCI or with dementia.

Our data was collected in a longitudinal study with multiple observations of many patients. There are four observations of two patients, three observations of three patients, two observations of thirty-nine patients, and one observation of eighty-eight patients. Five of the forty-four patients with multiple observations transition from normal to MCI within the study, with one patient transitioning further to dementia. Longitudinal data can have additional structure relative to cross-sectional data due to correlations induced by patient-level covariates, 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 feature information, can influence classifier performance. Under these conditions it is challenging to isolate the predictive power of a feature set [13]. 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 patient [14] [15]. These factors are a source of noise that can degrade CI classification accuracy.

To mitigate these effects, we do cross-observation averaging of speech features obtained from multiple observations with an unchanging CI diagnosis. For example, for a patient with four normal clinical assessments, we use averages of speech features obtained from the four audio collections. We do not average features across sessions where a transition from normal to MCI or from MCI to dementia takes place, under the assumption that these speech samples were generated from different cognitive states. This results in 5 patients in the dataset with multiple observations across different CI diagnoses. The remaining 134 patients have either one observation or one averaged observation. The resultant dataset includes 20 CI observations and 119 normal observations for testing and training.

3. Vocal feature extraction and selection

3.1. Feature extraction

We compute a suite of speech features based on phonetic, pseudo-syllable, and articulatory measures. For each observation, 106 speech features are obtained from both the AF and the EB tests, for a total of 212 features per observation.

Pseudo-syllable rate and phoneme duration metrics: We investigate measures of speech rate derived from the durations of individual phonemes. For the phoneme-based rate measurements, we use a phone recognition algorithm based on a Hidden Markov Model (HMM) approach, reported with a phoneme-recognition accuracy of about 80% [16]. This model was trained with English speech but not elderly speech in particular. We therefore observed lower phoneme classification accuracy on our audio. Accurate phoneme classification, however, did not appear to be as important as consistency/precision in this study. Interestingly, this has been true also for other studies using phoneme-like features from this HMM model for depressed state classification in German audio [7] [19].

We compute the number of speech units per second over the entire duration of a single patient’s session. *Speaking rate* refers to the average phoneme rate with pauses included, whereas *articulation rate* refers to phoneme rate with pauses excluded. We also computed each rate type based on pseudo-syllables [17]. Our automatic phoneme recognition system 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.

Phoneme-based features: These include the total duration, the average duration and the total count of 40 phonemes. The phoneme dictionary includes ‘sil’, the so-called silence phoneme, which is used to estimate pauses between speech segments. Because some patients have no detections for specific phonemes, we consider the subset of phonemes for which all participants have detections.

Articulatory coordination features: These measures are based on the dynamics of vocal resonances, or “formants,” over time. This feature extraction approach along with phoneme-based duration measures has been successfully applied to vocal signals to predict symptom severity in major depressive disorder [7] [19]. A detailed description of this feature analysis approach, in the context of epileptic seizure prediction from multichannel EEG, is provided in [20]. In summary, the approach computes channel-delay correlation and covariance matrices from the first three formant tracks. Each matrix contains correlation or covariance coefficients between the formant tracks from the audio samples.

Changes in the coupling strengths among the formant tracks cause changes in the eigenvalue spectra of the channel-delay matrices. For this work, we compute matrices with four different sub-frame intervals (which we refer to as “scales”), with 10 time-delays used per scale. Our features comprise the first 10 principle components of the combined eigenvalue spectra from the multiple time scales.

Pitch variance: We investigate pitch variability through pitch variance. Pitch estimation is performed with a sinusoidal-based algorithm [18], and pitch measurements are made in voiced regions as derived from the same sinusoidal-based algorithm. For each utterance in our database, the pitch variance is estimated as the mean-squared pitch deviation from the mean.

3.2. Feature selection

To find the most discriminating group of features and reduce dimensionality, we use a form of sequential feature selection

benchmarked by equal error rate (EER) and area under the ROC obtained during soft margin SVM classification with a radial basis function kernel. Sequential (wrapper) feature selection methods enable optimization of complimentary discriminatory information among sets of features [10] [11]. In Figure 1, we show 2D scatter plots of the six features providing the best area under the ROC and EER. Feature 1 is the total duration of ‘s’ phoneme (EB task), feature 2, the pseudo-syllable rate (AF task), feature 3, the average pause duration (AF task), feature 4, the total count of ‘m’ phoneme (AF task), feature 5, the pitch variance (EB task), and, feature 6, the eighth out of the 10 cross correlation principal component features (EB task).

Patients with CI (red) are not perfectly separated from the normal patients (blue) for any two features. However, for most feature combinations the CI observations span a subset of the distribution from normal observations. The SVM learns a discrimination boundary in the space defined by all six features to distinguish the normal from the CI class. In comparison, an alternative process of filtering features independently and applying a t-test to select those with p-values less than 0.05, described in [10], would eliminate three of the six features, resulting in lower discrimination performance.

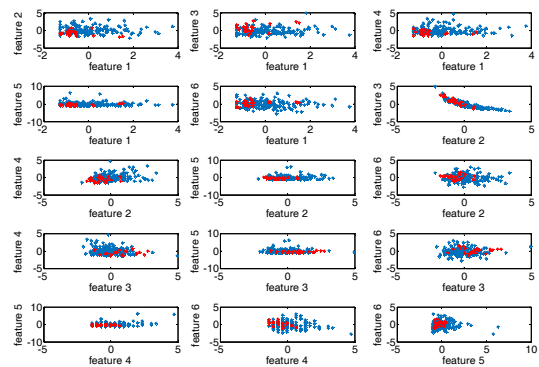


Figure 1. 2D scatter plots of the six features for normal (blue) and CI (red) observations. Feature 1 is total duration of ‘s’ phoneme (EB), feature 2 is pseudo-syllable rate (AF), feature 3 is average pause duration (AF), feature 4 is total count of ‘m’ phoneme, feature 5 is pitch variance (EB), and feature 6 is the eighth cross correlation feature (EB). For all of the pairwise plots, CI observations cluster in localized areas of the distribution for normal observations.

4. Results

4.1. Support vector machine classification

We use 10-fold cross validation to test and train a soft-margin support vector machine (SVM) classifier with a radial basis function kernel. ROC curves are shown in Figure 2 with EER = 13.5 +/- 0.2% for the average ROC derived from 200 iterations of 10-fold cross validation. This EER is lower than Satt et al. obtained in [8] and [9] for MCI prediction. It is also worth noting several other factors that distinguish our work. First, our model is less complex and we have more data for testing and training.

We use six features (compared to ~20 in [8] and [9]) and a dataset consisting of 139 observations (compared to ~90 in [8] and ~20 in [9]). Second, the audio data we used came from the responses collected during a battery of remote cognitive

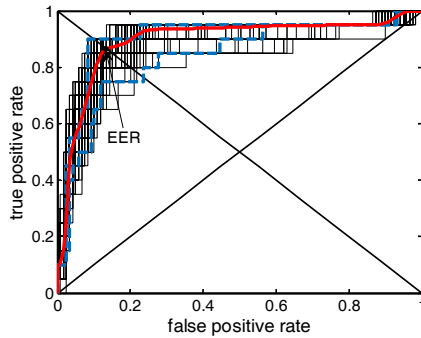


Figure 2. ROC plots showing variability in SVM classifier performance from 200 iterations of 10-fold cross validation. Each iteration uses the six speech features from the entire dataset. EER estimates fall on the $y = 1 - x$ line. The ROCs bounding upper and lower EER for the set are blue dashed lines. The average ROC is in bold red and the average ROC EER is marked with a black dot.

tests. It was not collected specifically for speech analysis. Our results are therefore obtained at a lower cost, in the sense that participants spent no extra time generating data for our non-linguistic speech analysis. Third, it is not clear if possible confounders such as age, alcoholism, depression, and other diseases common in the elderly, such as Parkinson's, were properly accounted for in [7] and [8]. We eliminated observations with known confounders due to illness or injury (17 different possible confounding factors were tested for during the clinical evaluation). We show in the next section that age was not a strong influence in our dataset on either CI diagnosis or speech feature metrics.

4.2. Age distribution study

Age is a potential confounding factor in this study: CI and dementia are known to increase with age in the elderly [21], and our speech features could also change with age. To be of clinical use, speech features need to provide additional predictive capability for CI beyond that predicted with aging. We examined the age distributions of the patients in our dataset in order to determine the influences of age on both CI diagnosis and speech features. In Figure 3, we show histograms of age for normal and CI patients. The distributions do not look drastically different in shape and there is no strong bias toward higher ages in the CI population.

While Figure 3 indicates that age is likely not a confounding factor, it is still desirable to approximately quantify the influence of age on classification results. In order to do so, we used the full dataset including repeated observations from the same participants at different ages. Using 10-fold cross validation on a reduced observation set of 19 CI observations and 156 normal observations (8 observations in the original set had no age data), we obtain hard-margin SVM classification results using: 1) speech features & age, 2) speech features without age, and 3) age alone (see Table 1.).

Age does not have strong predictive utility alone, with accuracy only slightly above chance. Including age in our feature set does not have a conclusive effect on performance. Specificity increases using age, but sensitivity decreases. We computed Pearson correlations between our z-scored speech features and age, and found no correlations larger than 0.15 in magnitude, with high associated p values (0.04-0.9).

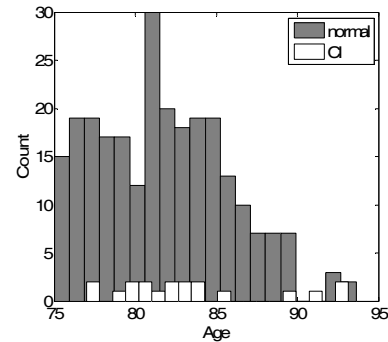


Figure 3. Histogram of participant ages in our sample. CI patients do not cluster in higher age ranges for this sample and are distributed similarly to the normal patients.

Table 1. We tested if age is a significant covariate by including it as a feature alone and with the set of six speech features described above. Results show age has little predictive utility for CI in this dataset, lending confidence to the predictive capability of our features.

	Accuracy	Sensitivity	Specificity
Reduced Dataset with Age	93%	63%	97%
Reduced Dataset without Age	90%	68%	92%
Reduced Dataset Only Age	56%	63%	55%

5. Conclusions

We demonstrate that non-linguistic features from speech samples already collected for remote cognitive testing can be used to identify cognitive impairment. Remote testing can therefore efficiently generate speech features without putting patients through extra speech tasks designed specifically for speech analysis. Our small set of six vocal features show promise for cognitive impairment classification. A soft margin SVM with 10-fold cross validation obtained an average EER of 13.5% using: durations and rates of several phonemes, an articulatory correlation measure, pitch variability, and pseudo-syllable rate. This performance compares favorably with recent results [8] [9], in which the authors design speech tasks for explicit speech analysis. We eliminate age as well as many common illnesses and injuries, known to affect cognitive functioning in the elderly, as confounding factors in the performance of our speech features.

We envision an automated system that uses remotely collected cognitive testing audio for diagnosis of cognitive impairment. This system will fuse linguistic and non-linguistic features derived from the same audio source. We will next investigate combining test scores, non-linguistic features and linguistic features to obtain high confidence classifications of cognitive function.

6. Acknowledgements

The authors acknowledge the National Institute on Aging who supported this work.

7. References

- [1] Sano, M., S. Egelko, M. Donohue, S. Ferris, J. Kaye, T. L. Hayes, J. C. Mundt, et al. "Developing Dementia Prevention Trials: Baseline Report of the Home-Based Assessment Study." *Alzheimer Dis Assoc Disord* 27, no. 4 (Oct-Dec 2013): 356-62.
- [2] Yu, B., Quatieri, T. F., Williamson, J. R., & Mundt, J. C. (2014). Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers. In *Fifteenth Annual Conference of the International Speech Communication Association*.
- [3] Appell, J., et al, "A study of language functioning in Alzheimer patients," *Brain Language* 17: 73-91, 1982.
- [4] Reilly, J., et al, "Cognition, language and clinical features of non-Alzheimer's dementias: an overview," *Journal of Communication Disorders* 43(5): 438-452, 2010.
- [5] Roark, B., Mitchell, M., Hosom, J.P., Hollingshead, K., and Kaye, J., "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 19, No. 7, September 2011.
- [6] 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
- [7] Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B., Mehta, D.D., "Vocal biomarkers of depression based on motor incoordination," *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM*, 2013.
- [8] Satt, A., Sorin, A., & Toledo-Ronen, O. (2012). Vocal biomarkers for dementia patient monitoring. *Proceedings of Interspeech 2012*.
- [9] Satt, A., Hoory, R., König, A., Aalten, P., & Robert, P. H. (2014). Speech-Based Automatic and Robust Detection of Very Early Dementia. *studies*, 3, 6.
- [10] Saeys, Y., Inza, I., & Larrañaga, P. (2007), "A review of feature selection techniques in bioinformatics," *Bioinformatics*, 23(19), 2507-2517.
- [11] Maldonado, S., & Weber, R. (2009), "A wrapper method for feature selection using support vector machines," *Information Sciences*, 179(13), 2208-2217.
- [12] Sager, M. A., Hermann, B.P., La Rue, A., and Woodard, J.L., "Screening for Dementia in Community-Based Memory Clinics," *Wisconsin Medical Journal* 105, no. 7 (2006): 25-29.
- [13] Singer, J.D. and Willett, J.B., "Applied longitudinal data analysis: Modeling change and event occurrence," Oxford university press, 2003.
- [14] Rantala, L., Vilkman, E., & Bloigu, R. (2002), "Voice changes during work: subjective complaints and objective measurements for female primary and secondary schoolteachers." *Journal of voice*, 16(3), 344-355.
- [15] Zhou, G., Hansen, J. H., & Kaiser, J. F. (2001). "Nonlinear feature based classification of speech under stress," *Speech and Audio Processing, IEEE Transactions on*, 9(3), 201-216.
- [16] 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)*.
- [17] Rouas J., "Automatic Prosodic Variations Modeling for Language and Dialect Discrimination," *IEEE Trans. Audio, Speech, and Language Proc.*, Vol. 15, No. 6, August 2007.
- [18] Quatieri, T. F. (2002). *Discrete-time speech signal processing: principles and practice*. Pearson Education.
- [19] 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* (pp. 65-72). ACM.
- [20] Williamson, J.R., Bliss, D.W., Browne, D.W., and Narayanan, J.T., "Seizure prediction using EEG spatiotemporal correlation structure," *Epilepsy & Behavior*, 25(2), 230-238, 2012.
- [21] Larson, E. B., Yaffe, K., and Langa, K. M. (2013). "New insights into the dementia epidemic," *New England Journal of Medicine*, 369(24), 2275-2277.