Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers

Bea Yu1, Thomas F. Quatieri1, James R. Williamson1, James C. Mundt2,
1 MIT Lincoln Laboratory, Lexington, Massachusetts, USA
2 Center for Telepsychology
[bea.yu,quatieri,jrw]@ll.mit.edu
jmundt@telepsychology.net

Abstract
Neurophysiological changes in the brain associated with early dementia can disrupt articulatory timing and precision in speech production. Motivated by this observation, we address the hypothesis that speaking rate and articulatory coordination, as manifested through formant frequency tracks, can predict performance on an animal fluency task administered to the elderly. Specifically, using phoneme-based measures of speaking rate and articulatory coordination derived from formant cross-correlation measures, we investigate the capability of speech features, estimated from paragraph-recall and naturalistic free speech, to predict animal fluency assessment scores. Using a database consisting of audio from elderly subjects over a 4-year period, we develop least-squares regression models of our cognitive performance measures. The best performing model 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 provide a reduction by about 30% in MSE from a baseline (mean score) in predicting cognitive performance derived from the animal fluency assessment.

Index Terms: early dementia, motor coordination, vocal biomarkers, formant frequencies, animal fluency task

1. Introduction
More efficient and effective methods are needed to screen and monitor the health status of the elderly in clinical practice and for research. In-person visits by trained clinicians are time consuming and costly, and many older persons have mobility restrictions and/or limited access to transportation. Data-mining of existing resources to inform new health care practices in geriatric medicine and community programs for older adults is desirable. Such research efforts leverage existing information from large-scale community studies and can promote development of innovative new technologies. The current effort relies on the Home Based Assessment (HBA) study of the Alzheimer’s Disease Cooperative Study (ADCS), a 4-year longitudinal data collection, entitled “Multi-Center Trial to Evaluate Home-Based Assessment (HBA) Methods for Alzheimer Disease Prevention Research in People over 75 Years Old,” to evaluate different technology platforms for administering home-based assessment measures outside of clinic visits [15]. 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 analyses reported below.

Vocal features are desirable as potential biomarkers of early dementia because they can be obtained easily (e.g., via telephone), greatly increasing global accessibility to an automated method for dementia assessment. Certain 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 [1–8]. Discovering the coupling between speech, language, and dementia [9,10] entails determining correlations between vocal features reflecting prosody, voice quality, and linguistic complexity with varying degrees of cognitive impairment [11,12].

A motivation of the present paper is the affect of neurophysiological changes associated with dementia on motor timing and coordination and therefore the disruption of articulatory control and kinematics [9,10]. More specifically, our approach is based on the hypothesis that general psychomotor slowing due to dementia affects speaking rate and articulatory coordination. We propose to characterize these disruptions in articulation and speaking rate with two recent methods: a phoneme-based rate measure (including pause information) [7] and articulatory coordination using formant-track cross correlations [14].

Our paper is organized as follows. In Section 2, we describe the database, including performance rating on the HBA animal fluency task. In Section 3, we describe our signal-processing methodologies for speaking rate and formant-coordination extraction. Section 4 motivates our regression approach with two exploratory studies: multi-level modeling of animal fluency assessment scores across patients and time, and correlations of speaking rate and articulatory coordination measures with the animal fluency scores. Section 5 reports predictions of mean animal fluency scores using regression models motivated by the exploratory studies. Section 6 provides conclusions and projections toward future work.

2. Early dementia 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 Disease Prevention Research in People over 75 Years Old,” 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...
3. Vocal feature extraction

Below, we describe our vocal features based on speaking rate and articulatory measures. We selected these features under the hypothesis that psychomotor and cognitive differences among the elderly are reflected in speech production timing and articulatory precision.

3.1 Speaking rate and average pause duration

We investigated 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 approach, reported with a phoneme-recognition accuracy of about 80% [17]. 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 [23]. 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 proceeding consonants grouped with it. For example, “V,” “CV,” and “CCV” are all valid pseudo-syllables. We found highest correlations with speaking rate, in contrast to articulation rate.

We also examined the silence or “pause” region within free-response speech recordings. Our automatic phoneme recognition algorithm categorizes pauses as distinct speech units, with lengths determined by estimated boundaries. We used the average pause duration in our correlation and regression analyses.

3.2 Articulatory coordination features

Articulatory coordination measures are based on the dynamics of vocal resonances, or “formants,” over time. Computing these measures requires the estimation of smooth, continuous, formant tracks. We use a formant tracking algorithm based on the principle that formants are correlated with one another in both frequency and time [18, 19]. Embedded in the algorithm is a speech-activity detector that enables a Kalman predictor to smoothly coast through non-speech regions. Kalman updates are made at 10-ms intervals.

We hypothesize that an early indicator of dementia may be degradation of the articulatory coordination governing vocal tract trajectories. We use a multivariate feature extraction to assess this degradation, positing changes in coordination are manifest in changes to the correlational structure among formant tracks. This feature extraction approach has been successfully applied to vocal signals to predict symptom severity in major depressive disorder and, along with phoneme-based duration measures [7], was a key component of the winning submission in the AVEC 2013 depression recognition sub-challenge [14]. A detailed description of this feature analysis approach, in the context of epileptic seizure prediction from multichannel EEG, is provided in [20].

This 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 at multiple relative time delays for the tracks sampled at different frame intervals by sub-sampling the original 10-ms frame interval, corresponding to implicit smoothing with different bandwidths. The approach is motivated by the observation that auto- and cross-correlations of measured signals can reveal hidden parameters, at different temporal resolutions and scales, in the underlying stochastic-dynamical systems. Changes over time in the coupling strengths among the formant tracks cause changes in the eigenvalue spectra of the channel-delay matrices. The matrices comprise four different sub-frame intervals (which we refer to as “scales”), with 10 time-delays used per scale. The four scales correspond to 10-, 30-, 70-, and 150-ms frame intervals (or 1, 3, 7, and 15 times the original 10-ms frame interval). At each scale, two summary statistics are also extracted from the channel-delay covariance matrices, characterizing the overall log power and entropy. These are computed as the logarithm of the trace and of the determinant of the covariance matrix, respectively. The resulting number of features is 128: 4 scales × (10 delays/scale × 3 formants + 2). In Section 4, we discuss an approach to reduce this high dimensionality.

4. Exploratory analysis

We used multi-leveling statistical modeling to evaluate within-subject and cross-time correlations in the animal fluency test scores. We then used correlations between these test scores and extracted vocal features as a guide to predictive models of cognitive performance.

4.1 Multi-level modeling on Animal Fluency Scores

In order to properly model correlations between vocal features and cognitive performance, we first examined correlations among the cognitive performance observations. The data in this study were collected longitudinally, with multiple observations from most patients. These observations, taken sequentially through time, are likely to cluster by patient, reflecting overall individual differences in cognitive functioning, and correlate over time due to temporal changes associated with clinical progression. Accounting for ‘multi-level’ structure in longitudinal data controls for underestimation of error in regression coefficients and
overestimation of statistical significance. Modeling this structure also improves accuracy in statistical inference [21].

We conducted two-level, exploratory analyses to understand the influence of time and within-patient score clustering on participants’ performances in the animal fluency task. This identified the sources of variance in the animal fluency scores that could potentially be explained by, or predicted from, speech features. We fit two exploratory models to the cognitive performance scores without including speech features [21,22]. The first model denoted by M, is the unconditional means model, lacking predictors at either level of the model [21]. This model partitions the response variance into between-patient variability, $\sigma^2_{GM}$, and within-patient variability, $\sigma^2_{GM}$. The second model, the unconditional growth model, denoted by G, introduces time into the first-level model [21]. $\sigma^2_{GM}$ represents the remaining variance after time is modeled. 59% of variance in the animal fluency score data, $\frac{\sigma^2_{GM}}{\sigma^2_{GM}+\sigma^2_{G}} = 0.59$, derives from differences between patients. Of the 41% remaining within-patient variation, 6%, $\frac{\sigma^2_{GM}-\sigma^2_{G}}{\sigma^2_{GM}} = 0.06$, is explained by linear regression over time. The hat notation, $\hat{x}$, indicates an estimate of the true parameter value, $x$. Based on these results, we decided to address the largest source of variance in animal fluency scores, between-patient differences, using inference based on speech features.

### 4.2 Correlation analysis of rate and pause duration

We computed Spearman rank correlations between the two speaking rate features (based on phoneme- and pseudo-syllable-rate) and average pause duration features, described in Section 3, and the animal fluency scores. These correlations were computed with features from the animal fluency, EBMT, free speech emotional and free speech functional audio data sets described in Section 2. For comparison, we tested 255 other standard speech features including prosodic (e.g., variance of pitch and its velocity) and multi-channel energy measures.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Speak. Rate (phoneme) (R,p)</th>
<th>Speak. Rate (pseudo syl.) (R,p)</th>
<th>Ave. Duration of Pauses (R,p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ani. Fluency</td>
<td>0.57, $10^{-12}$</td>
<td>0.58, $10^{-12}$</td>
<td>-0.60, $10^{-12}$</td>
</tr>
<tr>
<td>EBMT</td>
<td>0.21, $10^{-13}$</td>
<td>0.22, $10^{-13}$</td>
<td>-0.21, $10^{-13}$</td>
</tr>
<tr>
<td>Emotion FS</td>
<td>0.21, 0.003</td>
<td>0.26, 0.05</td>
<td>-0.25, 0.004</td>
</tr>
<tr>
<td>Function FS</td>
<td>0.27, $10^{-8}$</td>
<td>0.27, $10^{-8}$</td>
<td>-0.21, 0.004</td>
</tr>
</tbody>
</table>

A trend emerged across these four audio data sets. The three features correlating most strongly with animal fluency performance scores were our two speaking rate and average pause duration features. These findings were consistent across the four types of recorded verbal content, indicating robustness of this approach to speech data for prediction of this cognitive performance measure (Table 1). Speech features derived from the animal fluency recordings should correlate well with the animal fluency scores, since speech rate should be influenced by the cognitive ease with which participants recite the names of different animals in one minute. The generalization to recordings of other speech tasks is particularly important. In previous work, models incorporating multiple features, with moderate but significant correlations with the response, demonstrated strong depression prediction performance [7,14]. For the remainder of the study, we focused on features from the EBMT audio. EBMT speech provides a consistent response structure and is different enough from the animal fluency speech that predictions of the animal fluency score using EBMT speech demonstrate robustness of our speech features to variations in verbal content.

### 4.3 Analysis of articulatory coordination features

Examples of the formant coordination features from EBMT vocal segments are given in Figures 1 and 2. Figure 1 displays two channel-delay correlation matrices at the third scale (sub-frame interval), containing 10 delays spaced at 70 ms, which were obtained from two EBMT passages of 24.6 and 19.5 second duration. The matrices each contain nine 10×10 blocks. Each block consists of the within- or cross-channel correlation coefficients for one pair of formant tracks. These coefficients were computed using all possible pair-wise combinations of the 10 time-delayed versions of each formant track. The 10×10 blocks along the main diagonal contain the correlations within the same formant tracks, and the 10×10 off-diagonal blocks contain the correlations across different formant tracks. The matrix on the left was derived from Participant 126 at month 12 from baseline (animal fluency score = 4) and the matrix on the right was derived from Participant 34 at month 24 from baseline (animal fluency score = 46). The matrix associated with the higher cognitive performance score contains auto- and cross-correlation patterns that appear to exhibit greater level of complexity in the sense of increased energy in high-frequency components.

![Figure 1](image1.png)

**Figure 1.** Channel-delay correlation matrix from formant tracks for low fluency score (left) and high score (right).

![Figure 2](image2.png)

**Figure 2.** Left: Eigenspectra from formant channel-delay matrices for a low (blue) and high (red) fluency score. Right: Average normalized eigenvalues for different score ranges.

Qualitative differences in the appearances of the two matrices can be quantified using the matrix eigenspectra, which are the rank-ordered eigenvalues. These features are invariant to the underlying ordering of the channels (randomly permuting them will produce identical eigenspectra), and capture instead the levels of correlation among all three channels. The eigenspectra from the two matrices are in Figure 2 (left). The eigenspectra from the low-score subject contain more power in the first few and less power in higher eigenvalues, indicating reduced complexity and independent
Features set. The predicted mean scores have an RMSE=5.66, Figure 3 (left) plots the predicted mean scores of the 61 test rate and pause duration features described in Section 4.2. from training. The first feature set consists of the speaking a 2nd-order least-squares regression model (not shown) derived subjects is then correlated with their actual average scores. are reliable. The average predicted fluency score for these subjects with 860 sessions total), for whom the average scores are predicted fluency scores as a function of actual mean fluency scores: speaking rate features (left), articulatory coordination features (middle), and combined features (right).

The prediction task is complicated by the high variability in both the fluency scores and the features derived from the animal fluency test. Our method, using leave-one-patient-out cross validation, relies solely on data from other patients for training and helps account for the largest source of variability in the fluency test. Both feature domains, by reflecting vocal timing and frequency-based coordination measures, may help reduce effects of varying channels as introduced in standard telephone transmission and handsets.

We are currently considering expanding our vocal-only approach to correlations with other forms of cognitive testing in the elderly, as well as to more individualized methods to facilitate diagnosis and treatment. For example, a baseline of cognitive ability from speech samples could be obtained in an initial session. This baseline could then be used to benchmark fluency changes from voice in future sessions. Likewise, our method could be used to learn a patient-level model using speech features to predict future fluency trends for the patient, thus providing early interventions. In our future work, we will explore different approaches toward these objectives.

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


