Vocal biomarkers to discriminate cognitive load in a working memory task

Thomas F. Quatieri [1], James R. Williamson[1], Christopher J. Smalt[1], Tejash Patel[1], Joseph Perricone[1], Daryush D. Mehta[1], Brian S. Helfer[1], Greg Ciccarelli[1], Darrell Ricke[1], Nicolas Malyska[1], Jeff Palmer[1], Kristin Heaton[2], Marianna Eddy[3], Joseph Moran[3]

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

Abstract

Early, accurate detection of cognitive load can help reduce risk of accidents and injuries, and inform intervention and rehabilitation in recovery. Thus, simple noninvasive biomarkers are desired for determining cognitive load under cognitively complex tasks. In this study, a novel set of vocal biomarkers are introduced for detecting different cognitive load conditions. Our vocal biomarkers use phoneme- and pseudosyllable-based measures, and articulatory and source coordination derived from cross-correlation and temporal coherence of formant and creakiness measures. A ~2-hour protocol was designed to induce cognitive load by stressing auditory working memory. This was done by repeatedly requiring the subject to recall a sentence while holding a number of digits in memory. We demonstrate the power of our speech features to discriminate between high and low load conditions. Using a database consisting of audio from 13 subjects, we apply classification models of cognitive load, showing a ~7% detection equal-error rate from features derived from 40 sentence utterances (~4 minutes of audio).

Index Terms: cognitive load, vocal biomarkers, phoneme and pause duration, articulatory coordination

1. Introduction

Cognitive load is defined loosely as the mental demand experienced for a particular task [1][2]. More efficient and effective methods are needed to monitor cognitive load under cognitively and physically stressful conditions. Such conditions include environmental and occupational stressors that can result in dangerous scenarios when cognitively overloaded. Examples of mental stressors are repetitive and/or intense cognitive tasks, psychological stress, and lack of sleep. Physical stressors include intense long-duration operations and/or heavy loads. Both stressors can cause cognitive load, and often contribute simultaneously to load. Applications for cognitive load assessment include individualized detection of cognitive load in an ambulatory, field, or clinical setting. In clinical applications, the objective is often to find and measure the specific causes of load. In operational settings, the objective is often to quickly assess cognitive ability and readiness under loaded conditions, regardless of their etiology.

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

Biomarkers for monitoring and detecting cognitive load comprise behavioral, physiologic, and cognitive modalities. A potential class of biomarkers that has recently gained popularity is based on speech characteristics. Vocal features are desirable as biomarkers of cognitive status because they can be obtained easily (e.g., via telephone), greatly increasing global accessibility to an automated method for cognitive assessment. Certain vocal features have been shown to change with a subject's mental and emotional state, under numerous conditions including cognitive load. These features include characterizations of prosody (e.g., fundamental frequency and speaking rate), spectral representations (e.g., mel-cepstra), and glottal excitation flow patterns, such as flow shape, timing jitter, amplitude shimmer, and aspiration [1]-[8].

A motivation for the vocal features developed is the hypothesis that cognitive load can be assessed by measures of speech-segment-based prosodic dynamics and articulatory and source coordination. Specifically, we employ phoneme- and pseudosyllable-based measures that include rate, duration and pitch dynamics, as well as pause information, and articulatory and source coordination measures from cross correlations and cross coherences among extracted signals such as formant tracks, delta mel-cepstra coefficients, and creakiness signals. A subset of these vocal features have been used effectively in other neuro-cognitive contexts such as in detection of depression, traumatic brain injury, and dementia [8]-[11], thus perhaps forming a common vocal feature basis for neurocognitive change.

Our paper is organized as follows. In Section 2, we describe our data collection using a novel cognitive load protocol that taxes auditory working memory by eliciting sentence recall under varying levels of cognitive load. In Section 3, we describe our signal processing methodologies for vocal feature extraction. Section 4 reports cognitive load detection results using a Gaussian classifier. Section 5 provides conclusions and projections toward future work.

2. Working memory protocol

Subjects gave informed consent to a working memory–based protocol approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). Audio data are collected with a DPA acoustic lapel microphone (with a Roland Octa-Capture audio interface. Following setup and training, each subject engages in the primary task of verbally recalling sentences with varying levels of cognitive load, as determined by the number of digits being held in working memory [18]. Specifically, a single trial of the auditory working memory task comprises: the subject hearing a string

of digits, then hearing a sentence, then waiting for a tone eliciting spoken recall of the sentence, followed by another tone eliciting recall of the digits. This task is administered with three difficulty levels, involving 108 trials per level. The same set of 108 sentences is used in each difficulty level. The order of trials (sentences and difficulty level) is randomized. The entire protocol, approximately two hours in duration, is illustrated in Figure 1. The multi-talker PRESTO sentence database is used for sentence stimuli [15].

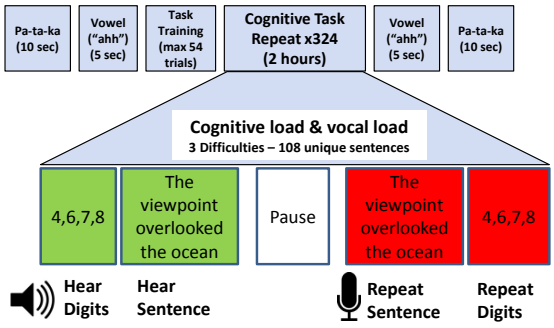

Figure 1: *Auditory working memory protocol.*

The working memory task is split into a training and a testing phase. During training, the length of presented digit sequence is alternately increased and decreased via an adaptive tracking algorithm [16] to determine the geometric mean of the number of digits a subject can recall. This number, n_c, is used to determine the maximum number of digits presented during testing, d_n, used in the three difficulty levels for the test phase. The test phase consists of 324 (= 108x3) consecutive trials randomized and balanced across the three difficulty levels.

We used d_n = {floor(n_c), floor(n_c)-2, floor(n_c)-4} for the first four subjects. Thus, a subject with n_c=5.32 would have d_n={5, 3, 1}. One way these difficulty levels might manifest behaviorally is via subject accuracy, with poorer accuracy at higher loads (more digits). The performance differences between easy and difficult conditions were small for our subjects, so we increased the difficulty by using d_n = {ceil(n_c), ceil(n_c)-2, ceil(n_c)-4}. This modification was used for one subject, after which we adopted d_n = {ceil(n_c), ceil(n_c)-1, ceil(n_c)-2} for the remaining eight subjects. As a result of these changes, digit-span accuracy was consistently lower for the hardest difficulty level compared to the easiest among the final nine subjects.

Despite the minor protocol changes between early and late subjects, a common load assessment test for all 13 subjects is possible due to the fact that all subjects had both a *max number* condition and a *max number minus two* condition. The range of digit spans across all subjects was 2–5 for low load and 4–7 for high load.

3. Feature extraction

Feature vectors are extracted only from the single spoken sentence component of each trial in the test phase of the auditory memory task. *Low-level* vocal features comprise measures of phoneme durations, pseudosyllable rate, pitch dynamics, articulation, spectral dynamics, and creak. We construct *high-level* features that capture inter-relationships among the low-level features. The feature sets are derived under the hypothesis that differences in cognitive load produce detectable changes in speech production timing and articulatory and source coordination. Low-level features, produced every 10 ms, are approximately immune to slowly-varying linear channel effects due to not being directly dependent on spectral magnitude.

3.1. Low-level vocal feature extraction

Phonemes: Using an automatic phoneme recognition algorithm [12], phonetic boundaries are detected, with each segment labeled with one of 40 phonetic speech classes.

Pseudo-syllables: Vocal syllable-like patterns are detected based on the concept of a pseudo-syllable (PS) [19]. The automatic phoneme recognition system detects individual speech sounds, which are combined into PS segments. For example, "v," "cv," and "ccv" are all valid PSs.

Pitch slopes: The fundamental frequency (pitch) is estimated using an autocorrelation method over a 40-ms Hanning window every 1 ms [20]. Within each phone or PS segment, a linear fit is made to the log of the pitch, yielding a pitch slope (Δlog(Hz)/s) for each phonetic or PS speech unit.

Formant frequencies: A Kalman filter technique is used to characterize vocal tract resonance dynamics by smoothly tracking the first three formant frequencies, while also smoothly coasting through non-speech regions [13].

Mel-frequency cepstral coefficients (MFCCs): 16 delta MFCCs are used to characterize velocities of vocal tract spectral magnitudes. Delta MFCCs are computed using regression with the two frames before and after a given frame.

Creaky voice quality: A creaky voice quality (vocal fry, irregular pitch periods, etc.), is characterized using acoustic measures of low-frequency/damped glottal pulses [21]. Low-level features include previously-studied metrics of short-term power, intra-frame periodicity, inter-pulse similarity [22], and two measures of the degree of sub-harmonic energy (reflecting the presence of secondary glottal pulses) and the temporal peakiness of glottal pulses with long period [23]. These low-level features are input into an artificial neural network to yield creak posterior probabilities on a frame basis [24].

3.2. High-level features

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

Phoneme-dependent features: Building on previous work [8]-[11], features conditioned on time segments of detected phonemes are constructed based on their discriminative value. For each phoneme, the features considered are: phoneme counts, total phoneme durations, and slopes of log-pitch during phonemes [9][11]. These features were computed in two different conditions: for all detected phonemes and for those phoneme instances where pitch slopes are marked as valid. The slope of log pitch values is marked as valid if its absolute value is less than eight [9], indicating that the slope is likely derived from a continuous pitch contour.

Four phoneme-based features were found useful, each an aggregate derived from a linear combination of 25 phonemes, with weights based on their discriminative value. In [9] these weights were derived from correlations with depression scores. Here, each weight is the signed Mahalanobis distance between the measured distributions (using mean and variance) for high and low loads. Table 1 lists the five most important

phonemes and their weights for each of the aggregate features. It is interesting to observe that the total pause count ("sil") plays an important role, consistent with other findings [4].

Table 1. *Phoneme-based features. The top 5 phonemes are listed for each feature, along with their weights.*

All Phns		Phns with valid pitch slopes					
Phn count		Phn count		Phn dur.		Pitch slope	
Phn	*w*	Phn	*w*	Phn	*w*	Phn	*w*
'sil'	2.2	'v'	1.5	'v'	1.4	'ae'	1.1
'v'	1.4	'ch'	1.1	'ch'	1.2	'ay'	1.1
'hh'	-1.2	'w'	-1.0	'w'	-1.0	'ng'	1.0
'zh'	0.8	'zh'	1.0	'zh'	0.9	'd'	0.8
'sh'	-0.7	'hh'	-1.0	'ao'	-0.8	'k'	0.7

Pseudosyllable-based features: A similar processing approach is applied to pseudosyllable (PS) speech segments. The PS dictionary contains silence ('#') and different combinations of consonants ('c') and vowels ('v'). Three different aggregate PS features were found useful, based on linear combinations of the top 10 PS-based measures of counts and pitch dynamics. As with the phoneme-based features, weights are the signed Mahalanobis distances between the measures for high and low loads (Table 2). Again the total pause count ("sil") plays an important role.

Correlation structure: Measures of the structure of correlations among low-level speech features have been applied in the estimation of depression [9], the estimation of cognitive performance associated with dementia [8], and the detection of changes in cognitive performance associated with mild traumatic brain injury [10]. The details for this approach are in [25], where the method was first introduced for analysis of EEG signals for epileptic seizure prediction.

Channel-delay correlation and covariance matrices are computed from multiple time series of vocal parameters. Each matrix contains correlation or covariance coefficients between the channels at multiple time delays. Changes over time in the coupling strengths among the channel signals cause changes in the eigenvalue spectra of the matrices. The matrices are computed at multiple "time scales" corresponding to separate sub-frame spacings. Features consist of the eigenvalue spectra of channel-delay *correlation* matrices, as well as covariance power and entropy from channel-delay *covariance* matrices.

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

Table 2. *The top five pseudosyllable (PS)-based features and their weights (w).*

All PS		PS with valid pitch slopes			
PS count		PS count		PS pitch slope	
PS	*w*	PS	*w*	PS	*w*
'#'	2.2	'c'	1.7	'ccv'	0.7
'c'	1.5	'ccv'	-0.8	'cccv'	-0.6
'ccc'	1.0	'ccc'	0.7	'v'	0.5
'ccccccv'	-0.7	'ccccccv'	-0.7	'cccc'	-0.5
'ccv'	-0.7	'cccccv'	0.7	'ccc'	-0.4

Table 3. *Channel-delay correlation and covariance features.*

Signal type	# channels	Feat. type	Sub-frame spacings	# raw feat.	# PCA feat.
Formant	3	corr.	1,3,7	135	3
Creak	1	corr.	1,3,7,15,31	75	2
Fmt-Crk	4	corr.	1,3,7	180	3
Formant	3	cov.	1,3,7	4	3
dMFCC	16	cov.	1,3,7	4	4

Table 3 shows parameters used to extract correlation structure features from three different low-level speech sources: formant frequency tracks, creak probabilities, and delta MFCCs. Sub-frame spacings of 1, 3, 7, 15, and 21 are used and, due to the 10-ms frame interval of the low-level features, these correspond to time spacings of 10, 30, 70, 150, and 210 ms, respectively. Each matrix (for each scale) is constructed using 15 time delays. The number of correlation-based features is the number of signal channels times the number of scales (i.e., number of sub-frame spacings) times the number of time delays (15) per time scale. The number of covariance-based features is the number of time scales (entropy features) plus one log power feature, as power is invariant across scale. Parameters are similar to those of previous studies [8]-[10], with numbers of principal components chosen based on discrimination performance.

The differences in eigenspectra patterns due to high versus low cognitive loads provide indications about the effect of load on speech. In Figure 2, averages across all subjects of normalized eigenvalues from formant and creak signals at time scale 3 (sub-frame spacing of 7) are shown for low load (blue) and high load (red). The eigenvalues are ordered, left to right, from largest to smallest. So, in both cases, there is greater power in the small eigenvalues during higher cognitive load. This indicates greater dynamical complexity in formant frequencies and creak during higher cognitive load.

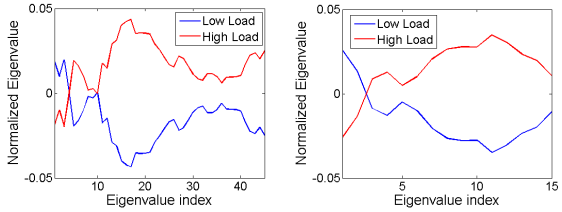

Figure 2: *Correlation structure features: Average normalized eigenvalues from all subjects for low and high cognitive loads, based on formant frequencies (left) and creak (right).*

Coherence structure and power: We have also introduced a feature set that characterizes the structure of signal coherence and power at multiple frequency bands. The coherence between channels, indicating the amount of cross-channel power in a frequency band relative to the amount of within-channel power, provides a measure of how closely related the signals are within a frequency band. The power and cross-power are computed among three formant frequency channels in two different frequency bands, and a 3 × 3 coherence matrix is constructed for each band. The eigenspectra of the coherence matrices indicate the structure of coherence across the channels. PCA is used to project these features into lower dimensional representations. Table 4

indicates the parameters, selected empirically by performance measures, used for the coherence and power features.

Table 4. *Frequency band coherence and power features.*

Signal type	Feature type	Freq. Band (Hz)	# raw features	# PCA features
Formant	Coh.	0.25 – 1.0	3	1
Formant	Coh.	1.0 – 2.0	3	1
Formant	Log Pow.	1.0 – 2.0	3	2

The differences in coherence and power features due to high versus low cognitive load provide indications about the effect of load on speech. In Figure 3 (left), averages across all subjects of normalized coherence eigenvalues from frequency band 1.0–2.0 Hz are shown for low load (blue) and high load (red). The eigenvalues are ordered, left to right, from largest to smallest. Similar to the correlation structure results shown in Figure 2, these results indicate greater power in the smaller eigenvalues for the higher load condition. In Figure 3 (right), it is shown that the higher load condition is also associated with less power for the first two formants.

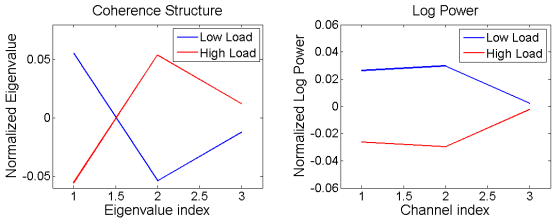

Figure 3: *Left: Average normalized eigenvalues from formant coherence matrix at frequency band 1.0–2.0 Hz for low and high cognitive loads. Right: Normalized log power for the three formant frequencies at frequency band 1.0–2.0 Hz.*

4. Results

Our goal is to detect differences in cognitive load from voice measurements. To evaluate detection performance, for each subject the 108 feature vectors (one vector per spoken sentence and load condition) from the max-digit condition is assigned to the *high* load class, and the 108 vectors from the max-digit-minus-two condition is assigned to the *low* load class. Leave-one-subject-out cross-validation is used with a classifier trained on 12 held-out subjects when discriminating between high and low load on a test subject.

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

Load discrimination is done with a Gaussian classifier (GC), where the Gaussians are centered on the two class means, and a common covariance matrix is used based on the data across both load conditions. In each trial, the GC produces a load score (log-likelihood ratio of high versus low load). A receiver operating characteristic (ROC) curve is obtained by varying a detection threshold to characterize the sensitivity/specificity tradeoff. For each subject, 216 scores are obtained (108 for each load). One ROC curve derived from scores of all 13 subjects characterizes total performance, with the area under the curve (AUC) serving as a summary statistic.

Table 5. *Summary of area under ROC curve (AUC) results for detecting high cognitive load from a single trial (sentence).*

Signal type	# features	AUC
Phoneme-based	4	0.59
PS-based	3	0.55
Corr. structure	15	0.56
Coh. structure	4	0.54
Combined	**26**	**0.61**

Table 5 lists the number of features used by the GC for each feature set, and the AUC results. The feature sets consists of the features described in Tables 1-4. The best overall performance of AUC = 0.61 is obtained by combining (via vector concatenation) all four feature sets.

Although our protocol involves feature processing of single spoken sentences, the ability to detect load across multiple sentences can be assessed by combining the GC scores from different trials, provided that the trials involve the same load condition. This was done by randomly selecting, from the same subject, a number of trials of either high load or low load, and summing their GC scores. For each subject, load condition and combination number, 200 randomly chosen sets of trials were used to determine performance across multiple sentences. Figure 4 (left) contains boxplots summarizing the AUC values for the 13 subjects, given combinations of 1, 5, 10, …, 40 trials. The median AUC value is 0.83 after 10 trials and 0.91 after 20 trials, with AUC for all subjects > 0.9 after 35 trials. In Figure 4 (right) are shown the cross-subject ROC curves from the same multi-trial combinations. For 40 trials (~4 minutes), we obtain an equal error rate of ~7%, corresponding to ~93% detection with ~7% false alarm.

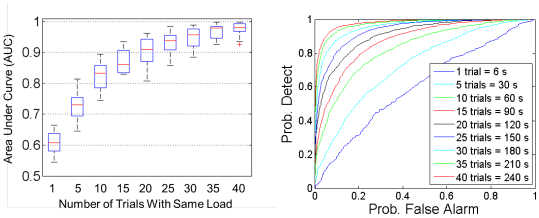

Figure 4: *Results as a function of number of combined trials with same load. Left: AUC values across 13 subjects. Right: Cross-subject ROC curves.*

5. Conclusions and discussion

In this paper, we demonstrated the power of our speech features to discriminate between high and low cognitive load conditions. Our features capture inter-relationships among phoneme durations, pseudosyllable rates, pitch dynamics, articulation, spectral dynamics, and creak. Using a database consisting of audio from 13 subjects and recalled sentences prior to recalling a digit span, we effectively applied classification models of cognitive load. Our approach, uses standard features at a "low-level" from which relational information is derived. Future work will involve a more formal comparison with alternative conventional approaches [1]-[7]. Future work will also involve expansion of our approach to the other modalities that were collected as part of this study (EEG, facial video, and physiology).

6. References

[1] S. E. Lively, D. B. Pisoni, W. Van Summers, and R. H. Bernacki, "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences," The Journal of the Acoustical Society of America, vol. 93, no. 5, pp. 2962–2973, 1993.

[2] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," *Proc. ICASSP*, 2008.

[3] B. Yin and F. Chen, "Towards automatic cognitive load measurement from speech analysis," in *Human-Computer Interaction. Interaction Design and Usability*. Springer Berlin Heidelberg, 2007, pp. 1011–1020.

[4] M. A. Khawaja, N. Ruiz, and F. Cheng, "Think before you talk: An empirical study of relationship between speech pauses and cognitive load," *Proc. OZCHI*, December 8–12, 2008.

[5] P. Le, J. Epps, H. C. Choi, and E. Ambikairajah, "A study of voice source- and vocal tract-based features in cognitive load classification," *Proceedings of the International Conference on Pattern Recognition*, 2010, pp. 4516–4519.

[6] H. Boril, O. Sadjadi, T. Kleinschmidt, and J. Hansen, "Analysis and detection of cognitive load and frustration in drivers' speech," *Proceedings of Interspeech*, 2010, pp. 502–505.

[7] T. F. Yap, "Speech Production Under Cognitive Load: Effects and Classification," PhD Thesis, The University of New South Wales School of Electrical Engineering and Telecommunications Sydney, Australia, 2011.

[8] B. Yu, T. F. Quatieri, J. W. Williamson, and J. Mundt, "Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers." *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[9] 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 *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 65–72.

[10] B. S. Helfer, T. F. Quatieri, J. R. Williamson, L. Keyes, B. Evans, W. N. Greene, J. Palmer, and K. Heaton, "Articulatory dynamics and coordination in classifying cognitive change with preclinical mTBI," in *15th Annual Conference of the International Speech Communication Association, September 9–13, Portland, Oregon, Proceedings*, 2014.

[11] A. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, vol. 42, pp. 1–18, 2011.

[12] W. Shen, C. White, T. J. Hazen, "A comparison of query-by-example methods for spoken term detection," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 2010.

[13] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, 2012.

[14] J. D. Singer and J. B. Willett, "Applied longitudinal data analysis: Modeling change and event occurrence," Oxford University Press, 2003.

[15] H. Park, R. Felty, K. Lormore, D. Pisoni, "PRESTO: Perceptually robust English sentence test: Open set—Design, philosophy, and preliminary findings," *The Journal of the Acoustical Society of America*, vol. 127, p. 1958, 2010.

[16] H. Levitt, "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical Society of America*, vol. 49, pp. 467–477, 1971.

[17] P. N. Le, E. Ambikairajah, H. C. Choi, and J. Epps, "A non-uniform sub-band approach to speech-based cognitive load classification," *Proceedings of ICICS*, 2009, pp. 1–5.

[18] J. D. Harnsberger, R. Wright, and D. B. Pisoni, "A new method for eliciting three speaking styles in the laboratory," *Speech Communication*, vol. 50, no. 4, pp. 323–336, 2008.

[19] J. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 6, pp. 1904–1911, 2007.

[20] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of the Institute of Phonetic Sciences*, vol. 17, 1993, pp. 97–110.

[21] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, no. 4, pp. 365-381, 2001.

[22] C. T. Ishi, K. I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 47–56, 2008.

[23] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," Computer, Speech and Language, vol. 27, no. 4, pp. 1028–1047, 2013.

[24] http://tcts.fpms.ac.be/~drugman/Toolbox/

[25] J. R. Williamson, D. Bliss, D. W. Browne, and J. T. Narayanan, "Seizure prediction using EEG spatiotemporal correlation structure," *Epilepsy and Behavior*, vol. 25, no. 2, pp. 230–238, 2012.