A Vocal Modulation Model with Application to Predicting Depression Severity *

Rachelle L. Horwitz-Martin, Thomas F. Quatieri, Elizabeth Godoy, and James R. Williamson

***Abstract*— Speech provides a potential simple and noninvasive "on-body" means to identify and monitor neurological diseases. Here we develop a model for a class of vocal biomarkers exploiting modulations in speech, focusing on Major Depressive Disorder (MDD) as an application area. Two model components contribute to the envelope of the speech waveform: amplitude modulation (AM) from respiratory muscles, and AM from interaction between vocal tract resonances (formants) and frequency modulation in vocal fold harmonics. Based on the model framework, we test three methods to extract envelopes capturing these modulations of the third formant for synthesized sustained vowels. Using subsequent modulation features derived from the model, we predict MDD severity scores with a Gaussian Mixture Model. Performing global optimization over classifier parameters and number of principal components, we evaluate performance of the features by examining the root-mean-squared error (RMSE), mean absolute error (MAE), and Spearman correlation between the actual and predicted MDD scores. We achieved RMSE and MAE values 10.32 and 8.46, respectively (Spearman correlation=0.487, p<0.001), relative to a baseline RMSE of 11.86 and MAE of 10.05, obtained by predicting the mean MDD severity score. Ultimately, our model provides a framework for detecting and monitoring vocal modulations that could also be applied to other neurological diseases.**

I. Introduction

Speech is an easily obtainable physiological signal that results from complex neurological and motor production processes. With the brain driving the process, physiological production of speech involves air being pushed from the lungs, through vocal folds vibrating at the larynx, into a resonating vocal tract, with the acoustic output ultimately radiating from the lips. Thus, embedded in the speech signal is information about control and functioning of underlying production mechanisms. Consequently, speech represents an important biomarker that can be used to classify and monitor neurological disorders that affect speech production, such as Parkinson's disease [1]–[3], Traumatic Brain Injury [4], [5], and the more common Major Depressive Disorder (MDD) [6]–[8].

* 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. The work of R. L. Horwitz-Martin was supported in part by the National Institute of Deafness and Other Communication Disorder Grant T32 DC00038.

R. L. Horwitz-Martin and T. F. Quatieri are with MIT Lincoln Laboratory, Lexington, MA 02421 USA and the Speech and Hearing Bioscience and Technology program, Harvard-MIT, Cambridge, MA 02139 USA (emails: rachelle.horwitz@ll.mit.edu, quatieri@ll.mit.edu).

E. Godoy and J. R. Williamson are with MIT Lincoln Laboratory, Lexington, MA 02421 USA (emails: elizabeth.godoy@ll.mit.edu, jrw@ll.mit.edu).

In particular, for depression severity prediction, a modulation spectrum of speech has been shown to be a salient feature [7]. The modulation spectrum captures the frequency content of temporal envelopes of speech in different frequency bands across time. Underlying this representation is a premise that certain oscillations of speech production mechanisms are manifested as modulations (amplitude and/or frequency) of the acoustic signal. The current work aims to provide a principled framework for modeling and ultimately exploiting vocal modulations observable in the speech signal, with a direct link to specific underlying production mechanisms.

Our proposed model addresses two mechanisms that lead to amplitude modulation (AM) in the temporal envelope of the speech signal. The first component of AM captures oscillations of the respiratory muscles used to push air through the vocal folds and into the vocal tract. The second AM component results from a more complex acoustic interaction between the source signal of the vocal folds and the resonance pattern of the vocal tract. This phenomenon is known as resonance-harmonics interaction (RHI). One primary cause of RHI is modulation of the frequency at which the vocal folds vibrate, where the frequency at which the vocal folds vibrate is known as the fundamental frequency (f_0). This results in the shifting of f_0 harmonics back and forth through vocal tract resonances, or formants (peaks), in the speech spectrum [9]. In this work, we introduce a modeling and synthesis framework explicitly capturing both sources of AM in the temporal envelope of the speech signal.

To distinguish these two AM components in speech, several signal envelope estimation techniques are examined, including a novel iterative nonlinear approach we designed to capture modulations that are slowly varying in time, without fitting the speech signal pattern with excessive detail. In analyses of the simulated modulated speech, the resulting envelope estimation technique is shown to better resolve the different AM components than two standard methods. Finally, using this envelope estimation technique, vocal modulation features are generated for real speech and used to predict MDD severity.

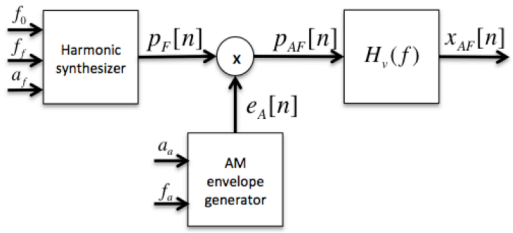

Figure 1. AM-FM model components.

II. Framework

In this section, we develop a framework for our envelope modulation model of the speech waveform. We begin with a brief review of the source-filter model of speech production and then describe a baseline model for the two AM contributions to the envelope: oscillations of the respiratory muscles, and interaction between the source signal of the vocal folds and the resonance pattern of the vocal tract.

A. Source-Filter Model

During the voicing state of speech, vibration of the vocal folds generates periodic, quasi-periodic, or irregularly spaced puffs of air that excite the vocal tract. Assuming a linear, time-invariant model of the vocal tract, the output at the lips, $x[n]$, is given by the convolution of the source $s[n]$ and the vocal tract impulse response $h_v[n]$ and is expressed as $x[n] = s[n] * h_v[n]$. In the frequency domain, with periodic or quasi-periodic fold motion, the source spectrum, denoted by $S(f)$, contains harmonics (or approximate harmonics) of f_0. The vocal tract transfer function (VTTF), denoted by $H_v(f)$, has resonances (formants) associated with peaks in $|H_v(f)|$, and its output $X(f)$ is expressed by

$$X(f) = S(f)H_v(f).$$

B. Modulation Components

According to Stockham [10], an approach to model an acoustic signal, $x[n]$, is to model it as the product of an envelope $e[n]$, where $e[n] > 0$, that contains low frequency components, and a high frequency signal, $v[n]$:

$$x[n] = e[n]v[n]. \quad (1)$$

Estimates of envelope AM frequency ranges vary. Some estimates are 4.5 Hz to 12.6 Hz [11] and 1.1 Hz to 24 Hz [12]. In our model, we interpret $e[n]$ to contain AM components due to oscillations of the respiratory muscles and interaction between the source signal of the vocal folds and the resonance pattern of the vocal tract.

Contribution from respiratory muscles: The first source of AM originates from oscillations in subglottal pressure, causing the voice source to vary in amplitude. Assuming the fundamental frequency f_0 (and thus frequencies of the harmonics of f_0) and the vocal tract are fixed, the amplitudes of the harmonics change together over time, resulting in AM [9]. As the subglottal pressure increases, the amplitude of the vibration of the vocal folds also increases [13], resulting in increased sound intensity [14]. In a study investigating the relationship between induced respiratory tremor and the acoustic characteristics of vocal tremor, Lester and Story set the induced respiratory modulation frequency to 5 Hz [15].

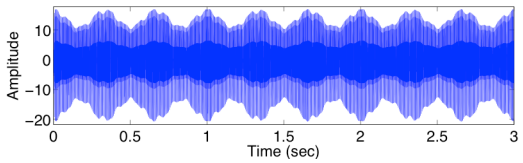

Figure 2. $x_{AF}[n]$ in the time domain, with f_0=200 Hz, f_f=11 cycles/sec^2, a_f=5 Hz, a_a=0.1, and f_a=3 Hz. The formants are 820 Hz, 1220 Hz, and 2810 Hz, with bandwidths of 125 Hz, 125 Hz, and 250 Hz.

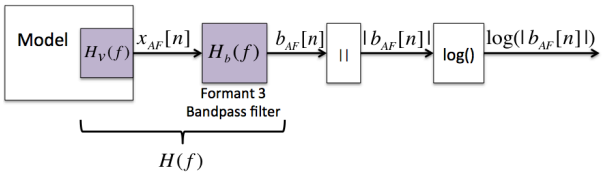

Figure 3. Block diagram of Stockham's method.

Contribution from resonance-harmonics interaction: One cause of resonance-harmonic interaction (RHI) is a change in the shape of the vocal tract. For example, the tongue and pharynx can move, causing a change in the formant frequencies. If f_0 remains fixed, AM can occur as the formant frequencies of $H_v(f)$ move through the harmonics of f_0. Likewise, AM can arise from changes to the rate and extent of the f_0 via changes in vocal fold motion, i.e., harmonics move through the formants. The AM component contributed by the RHI tends to be faster than that of the respiratory muscles and often is on the order of f_0 or its multiples depending on how harmonics and formants interact [9], [16]. Numerous rates of f_0 change have been reported, varying from 1.1 Hz to 25 Hz in both normal and pathological phonation [11], [12], although recent literature cites frequencies up to 12 Hz [17]. In this work, we will assume only the latter RHI, where frequency-modulated (FM) harmonics move through a fixed vocal tract. We will denote its envelope component as $e_F[n]$.

C. Signal Processing Model

In this section, we develop a speech signal-processing model with the above two modulation components. To begin, we model the source signal from the vocal folds as a frequency-modulated (FM) set of harmonics of f_0 and denote it by $p_F[n]$. This source signal is then shaped by the envelope $e_A[n]$, due to the respiratory component, and the resulting AM source $p_{AF}[n] = e_A[n]p_F[n]$ drives the vocal tract transfer function (VTTF). This general framework involving an AM envelope $e_A[n]$, multiplying the FM pulse train $p_F[n]$, and convolving with vocal tract impulse response $h_v[n]$ to generate an output $x_{AF}[n]$, is illustrated in Fig. 1.

For illustration, both $e_A[n]$ and $p_F[n]$ are chosen to be sinusoidal. For $e_A[n]$, the depth of modulation (a_a) and rate of modulation (f_a) are constants; for the FM pulse train from the vocal folds $p_F[n]$, the rate of FM (f_f) and extent of FM (a_f), and center fundamental frequency (f_0) are also constants. f_s is the sampling frequency. These parameter specifications are summarized in Fig. 1.

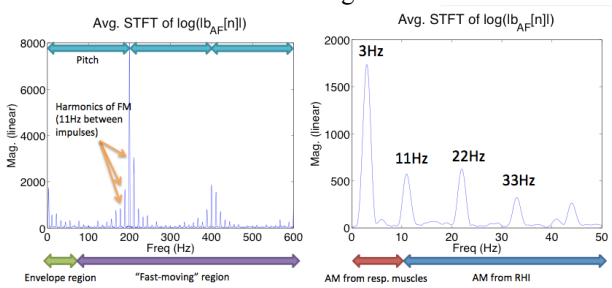

Figure 4. Frequency components in the log of the magnitude of the envelope, plotted over two frequency ranges. Left panel: the frequency range is 0 to 600Hz. Right panel: the frequency range is 0 to 50Hz.

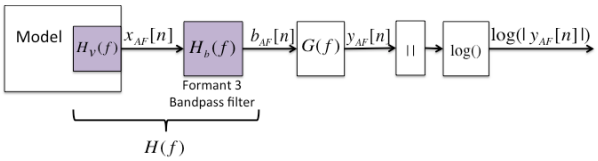

Figure 5. Block diagram the Hilbert-Stockham method.

The equation for the AM signal from the respiratory muscles is given by

$$e_A[n] = \frac{1}{2} + a_a \cos\left(2\pi f_a \frac{n}{f_s}\right) \quad (2)$$

and letting K be the total number of synthesized harmonics, the FM pulse train, $p_F[n]$, is given by

$$p_F[n] = \sum_{k=1}^{K} \cos\left(2\pi k f_0 \frac{n}{f_s} + \frac{a_f k}{f_f} \sin\left(2\pi f_f \frac{n}{f_s}\right)\right) \quad (3)$$

where the vocal tract input $p_{AF}[n]$ is $p_F[n]$ shaped by the envelope $e_A[n]$.

Fig. 2 displays an example simulation. The 3-Hz component from $e_A[n]$ is clearly present in the envelope of $x_{AF}[n]$, while the RHI contributes the higher frequency components to the envelope due to the 11 cycles/s^2 FM frequency component of the source. Envelope extraction must be performed to distinguish the AM components in the envelope.

III. Envelope Extraction

To distinguish the AM component due to the respiratory muscles from the component due to RHI, we isolate a single formant and evaluate its envelope using three methods: (1) Stockham's method, (2) the Hilbert-Stockham method, in which the Hilbert envelope extraction method is performed, followed by Stockham's method, and (3) a novel, nonlinear iterative (NLI) envelope algorithm followed by Stockham's method.

We analyze the envelope of a single formant under the assumption that the limited bandwidth leads to less complex RHI over a limited bandwidth. We select the third formant (F3) in particular because harmonic FM is greater at higher frequencies. This is due to harmonics of the modulated f_0 having a greater depth of FM than the harmonics at lower frequencies, i.e., lower formants. With these two properties, we assume there to be greater discriminability of the RHI and respiratory envelope components.

A. Stockham's Method

Stockham's method is based on his waveform model, provided in (1), and the property that the logarithm (log) of a product of two variables is the sum of the log of each variable. A block diagram of Stockham's method is illustrated in Fig. 3, where a bandpass filter $H_b(f)$ is applied. $H_b(f)$ has a bandwidth of 250 Hz and is centered at the F3 frequency.

Our envelope separation problem is now posed with respect to the output of the combined filter $H(f) = H_v(f)H_b(f)$ (Fig. 3), which we write as

$$b_{AF}[n] = e[n]v[n] \quad (4)$$

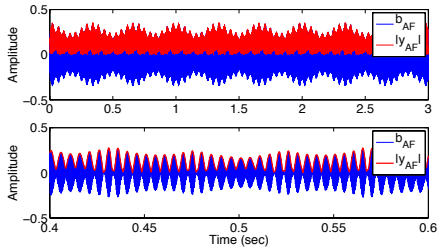

Figure 6. Comparison of $b_{AF}[n]$ and $|y_{AF}[n]|$. Top: $b_{AF}[n]$ and $|y_{AF}[n]|$, plotted between 0 and 3 sec. Bottom: $b_{AF}[n]$ and $y_{AF}[n]$, plotted between 0.4 and 0.6 sec.

where we have used Stockham's waveform model (1). However, $e[n]$ and $v[n]$ now refer to the envelope and the high frequency signal associated with F3, $b_{AF}[n]$.

We assume $e[n]$ is represented by the product of two components: the AM due to the respiratory muscles, $e_A[n]$, and the AM due to the interaction of the source with the vocal tract's third formant, $e_F[n]$:

$$e[n] = e_A[n]e_F[n]. \quad (5)$$

Therefore,

$$|b_{AF}[n]| \approx e[n]|v[n]| \approx e_A[n]e_F[n]|v[n]| \quad (6)$$

where $v[n]$ can be modeled as a series of (constant-height) impulses convolved with the combined vocal tract-bandpass filter impulse response $h[n]$ [18]. Taking the log of (6) and computing the Fourier transform of the logarithm, we obtain

$$\mathcal{F}\{\log(|b_{AF}[n]|)\} = \mathcal{F}\{\log(e_A[n])\} + \mathcal{F}\{\log(e_F[n])\} + \mathcal{F}\{\log(|v[n]|)\}. \quad (7)$$

For the example signal in Fig. 3 (with inputs (2) and (3)), the spectrum of $\log(|b_{AF}[n]|)$ is computed from the short-time Fourier transform (STFT) using a 1-second Hamming window with 50% overlap, and is displayed in the left panel of Fig. 4. The *envelope region* primarily consists of components belonging to the log-envelope, $\log(e[n])$. This contains components from both $\log(e_A[n])$ and $\log(e_F[n])$. The "fast-moving" region corresponds to the frequency components belonging to $\log(|v[n]|)$, showing a periodicity at the 200-Hz fundamental frequency as well as impulses with spacing at the rate of the FM: 11 cycles/s^2.

The right panel of Fig. 4 shows the regions to which $e_A[n]$ and $e_F[n]$ map within the envelope region. The first major peak is at a frequency of 3 Hz, which corresponds to the frequency of the AM due to the respiratory muscles. The next peak is at a frequency of 11 Hz, the FM rate. Subsequent peaks occur at multiples of 11 Hz.

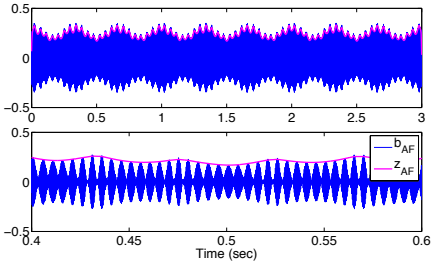

Figure 7. Top: $b_{AF}[n]$ and $z_{AF}[n]$, plotted between times 0 and 3 sec. Bottom: $b_{AF}[n]$ and $z_{AF}[n]$, plotted between 0.4 and 0.6 seconds.

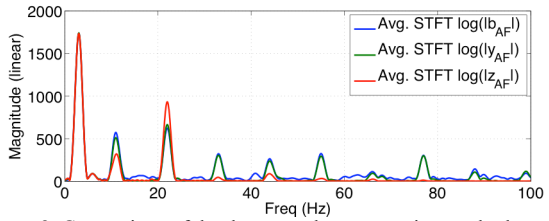

Figure 8. Comparison of the three envelope extraction methods.

The short-time spectrum of log ($|b_{AF}[n]|$) contains peaks at the AM and FM frequencies and their harmonics, but they also contain a high-frequency artifact between the peaks. This artifact is due to computing the magnitude of $v[n]$ in $|b_{AF}[n]|$ (6), which causes the negative time-domain components of $h[n]$ to become positive, and introduces additional high-frequency components, associated with F3, that can leak into the low-frequency component (see Appendix and [18]).

B. Hilbert-Stockham Method

Incoherent envelope detection is one approach to reduce artifacts from the basic Stockham method. This can be performed by bandpass filtering around a carrier frequency, computing the *Hilbert transform* of the filter output, and taking the magnitude [19]. As before, we compute the log of the magnitude in an effort to further separate the fast-moving component from $e[n]$. Fig. 5 shows a block diagram that combines the Hilbert transform and Stockham's method. Given the Hilbert envelope operator $g[n]$, we can show (Appendix and [18]) that the magnitude of output $y_{AF}[n]$ is

$$|y_{AF}[n]| \approx e[n]|\bar{v}[n]| \approx e_A[n]e_F[n]|\bar{v}[n]| \qquad (8)$$

where we define $\bar{v}[n]$ to be the Hilbert transform of $v[n]$. Therefore, to distinguish the two AM components, we compute the Fourier transform of the log of each of the components in (8):

$$\mathcal{F}\{\log(|y_{AF}[n]|)\} = \mathcal{F}\{\log(e_A[n])\} + \\ + \mathcal{F}\{\log(e_F[n])\} + \mathcal{F}\{\log(|\bar{v}[n]|)\}$$

In contrast to $|v[n]|$ in (6), which follows individual zero crossings of the formant and thus the formant frequency, $|\bar{v}[n]|$ provides a smooth formant envelope, reflecting its bandwidth but not frequency.

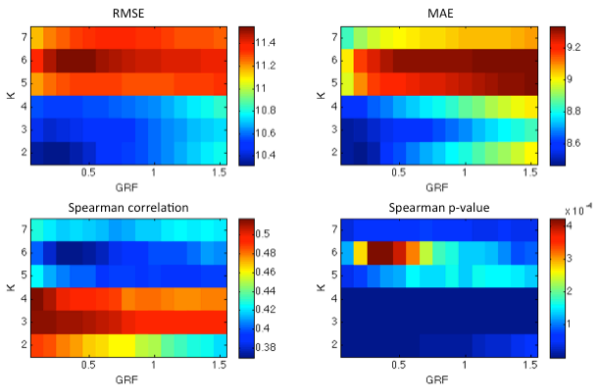

Figure 9. RMSEs, MAEs, and Spearman correlations of the average STFT magnitude of the LM-NLI-Stockham envelope of F3, excluding frequencies above 50 Hz, with feature adaptation. For all plots except the bottom left, a cooler color indicates a lower value, which is desirable.

For the example signal in Fig. 2, Fig. 6 compares $b_{AF}[n]$ to the magnitude of its Hilbert transform, $|y_{AF}[n]|$. In the bottom plot, $|y_{AF}[n]|$ appears on a local scale to follow the periodicity of $b_{AF}[n]$ and not the F3 frequency of $h[n]$.

C. NLI-Stockham Algorithm

To further reduce artifacts from high-frequency leakage with $|v[n]|$, we apply a novel nonlinear, iterative (NLI) algorithm to estimate the envelope of a signal. The result is a temporal envelope that rides over both the high frequency of the 3rd formant as well as pitch periodicity. This approach can be interpreted as a temporal rendition of the iterative spectral envelope estimator given in [20]. As in the other two approaches, we first bandpass filter the vocal tract output to obtain $b_{AF}[n]$. Then, assuming the waveform magnitude representation in (6), we compute the NLI envelope by convolving $|b_{AF}[n]|$ with an equal-weight moving average filter of an empirically determined length of 2.5 ms. For each point k along the length of $|b_{AF}[n]|$, the maximum value between $|b_{AF}[k]|$ and the convolution is retained in a buffer. The subsequent convolutions are between the buffer and the equal-weight moving average filter. After several iterations (150 appeared sufficient to capture a smooth, slowly varying shape), the result is a signal we denote by $z_{AF}[n]$. As before, since it is desirable for the log of the AM from the respiratory muscles and the AM from the RHI to be additive, the log of $|z_{AF}[n]|$ is computed.

Fig. 7 displays $b_{AF}[n]$ and the NLI envelope, $z_{AF}[n]$, for the earlier example. The NLI envelope appears to accurately estimate the envelope in the time domain. Compared to the Hilbert envelope in Fig. 6, it removes additional high frequencies that are present in the periodicity of the Hilbert envelope, thus removing artifacts that are not components of the "true envelope" of the speech waveform.

D. Comparison of Envelope Extraction Techniques

Fig. 8 compares the log-magnitudes of the outputs of the three processing methods. Each method yields peaks at 3 Hz, the rate of the AM, and peaks at integer multiples of 11 Hz, the rate of the FM. However, the 11 Hz component decays more quickly and there is less high-frequency leakage when the NLI -Stockham envelope is extracted due to the ability of the NLI-Stockham envelope to track the "true envelope" of the signal, without capturing information from the pitch periods.

Our analysis indicates that the Stockham, Hilbert-Stockham, and NLI-Stockham methods are effective in distinguishing the two sources of AM in synthetic FM-and-AM signals, but the NLI-Stockham algorithm appears to best reduce high-frequency artifacts.

IV. Depression Prediction

Motivated by the model and envelope extraction techniques discussed in Sections 2 and 3, and also by the work of Cummins [7], we hypothesized that features associated with the log-envelope of F3 would predict MDD. The features that provided the most accurate MDD prediction were computed by performing principal component analysis (PCA) on the average STFT magnitude of the F3 envelope,

extracted using the NLI-Stockham method. This comparative analysis, along with results obtained using other types of features, and features derived from the other two envelope extraction methods can be found in [18]. Although it is difficult to extract physiological meaning from features obtained by performing PCA on spectrum of the F3 envelope, these features might lay the groundwork for features that provide physiological meaning.

A. The AVEC Database

Feature selection and classification/regression were performed on speech samples from the 2013 Audio/Visual Emotion Challenge (AVEC) database. We used 87 sessions of audio recordings from the AVEC database, which contains recordings from males and females with various MDD severities. The 16-bit audio was recorded using a laptop's sound card at a sampling rate of either 32 or 41.1 kHz [21]. For this work, we extracted the middle three seconds of sustained /a/ vowels from the recordings.

The subjects' MDD severity ratings were scored with the self-reported Beck assessment, where a higher Beck score corresponds to higher severity. Among the 87 sessions from the AVEC database, Beck scores vary between 0 and 45.

B. Feature Extraction

After downsampling the waveforms sampled at 41.8 kHz to 32 kHz, the first task in extracting the average STFT magnitude of the log of the magnitude (LM) of the envelope was to extract the NLI-Stockham envelopes from the 87 raw waveforms. Using a Kalman-based autoregressive moving average framework [22], the formants of each of the waveforms were computed, a bandpass filter that isolated F3 ± 250 Hz was applied, and the LM of the NLI-Stockham (LM-NLI-Stockham) envelopes were computed. Finally, to compute the average STFT of the LM-NLI-Stockham envelopes, we applied 1-second Hamming windows with 50% overlap, and computed a very long (262,144-point) Discrete Fourier Transform to each segment in an effort to maximize our ability to visualize the spectral content.

We extracted the average STFT values corresponding to frequencies below 50 Hz, which is half the fundamental frequency of a male, leading to 410 average STFT values. We also computed the average STFT values corresponding to frequencies below 20 Hz, the upper limit of the majority of the spectral content. For results obtained using those features, refer to [18]; here, we focus on the results obtained using frequencies below 50 Hz, as they predicted MDD most accurately.

C. Regression and Prediction Procedure

As a basis for prediction, we used a Gaussian Mixture Model (GMM). To train the GMM, we used Gaussian Staircase Regression, a technique that utilizes multiple data partitions to create a GMM for two classes and has successfully been used to predict MDD on the AVEC data [23]. We added a constant, the Gaussian regularization factor (GRF), to the diagonal of the covariance matrix to prevent over-fitting the data. Since some subjects appeared more than once, the means in the Gaussian model were adapted toward the mean for the subject. This process has been called *feature adaptation* [23]. The purpose of feature adaptation is to mitigate the effects of intersubject feature variability when prior information is available from the training set about a subject's feature values. Feature adaptation is similar to the widely used speaker recognition technique by which speaker models are created from a Universal Background Model [24]. We also performed the regression and prediction procedure without using feature adaptation; for those results, refer to [18].

The training and testing procedure consisted of performing leave-one-session-out cross-validation. We utilized the following three metrics to assess prediction accuracy: mean absolute error (MAE), root mean square error (RMSE), and Spearman correlation (ρ) between the actual Beck scores and predicted Beck scores. To establish a baseline to compare our results against, we predicted the mean Beck score for each session. This led to baseline RMSE and MAE values of 11.86 and 10.05, respectively.

D. LM-NLI-Stockham Envelope Performance

Among the numerous types of modulation features we tested, the average STFT magnitude of the LM-NLI-Stockham envelope yielded the lowest RMSE and MAEs, and the highest Spearman correlation. Without performing PCA, the average STFT magnitudes of the LM-NLI-Stockham envelope comprise 410 features from the STFT. The dimensionality of this feature space needs to be reduced to effectively apply the GMM approach. To reduce dimensionality, we applied Principal Component Analysis (PCA), varying the number of components from 2 to 7. Values for the Gaussian regularization factor (GRF) were simultaneously varied from 0.1 to 1.5 in steps of 0.1. Fig. 9 displays the results from this global optimization procedure.

With the GRF at 0.2 and with 2 PCA components, we achieved RMSE and MAE values 10.32 and 8.46, respectively (Spearman correlation=0.487, $p<0.001$), relative to the baseline RMSE of 11.86 and MAE of 10.05.

V. Conclusions

We proposed a model of vocal modulation as a basis for developing biomarkers of neurological disease and, in particular, Major Depressive Disorder (MDD). The modulation model was developed in the context of a sustained vowel, assuming that two components contribute to AM: oscillations of respiratory muscles and AM resulting from RHI. The two AM components were represented in the model as multiplicative contributions to the speech signal's envelope. We produced a synthetic speech signal incorporating these components. To explore the separability of the modulation contributions, we bandpass filtered the synthesized speech signal to isolate the third formant, F3, thus accentuating the envelope contribution from the RHI. We applied three envelope extraction techniques to the F3 signal: (1) Stockham's method, where the natural logarithm of the magnitude of the signal is extracted [10], (2) the Hilbert-Stockham method, in which the magnitude of the Hilbert transform of the signal is computed and Stockham's method is subsequently applied, and (3) the NLI method (a

temporal rendition of [20]), combined with the Stockham approach, referred to as the NLI-Stockham method.

We derived features from the third formant of real speech signals from depressed subjects, and predicted depression severity using Gaussian Staircase Regression [23] while globally optimizing parameters. Of all features tested, we found that using 2 PCA components from the average STFT of the LM-NLI-Stockham envelope yielded the most accurate Beck score prediction. This most accurate prediction obtained was a decrease of 1.54 points from baseline (from 11.86 to 10.32) in the RMSE, and a decrease of 1.59 (from 10.05 to 8.46) in the MAE. The corresponding Spearman correlation between the predicted Beck score and actual Beck score was 0.487 ($p<0.001$).

The modeling and prediction methodologies described provide a foundation for the future work, which includes improving the underlying model, implementation of the model, the pre-processing methods, feature extraction methods, and machine learning. Although cross-validation evaluation was performed, since global optimization was also performed, it is important to note that these results are preliminary, and additional analysis must be performed to assess generalization. In particular, the optimal parameters we found must be tested on a new held-out data. Our ultimate goal is for results from future work to be applied toward diagnosing and monitoring depression and other neurological disorders using a simple, noninvasive "on-body" platform.

APPENDIX: DERIVATION OF ENVELOPE REPRESENTATION [18]

The rapidly-varying component of the speech signal in (4), $v[n]$, can be approximated by a flat-amplitude series of impulses, denoted by $\tilde{p}[n]$, convolved with the combined vocal tract-bandpass filter impulse response $h[n]$ [10]. The spacing of each impulse in $\tilde{p}[n]$ is equal to the reciprocal of the fundamental frequency. An expression for $v[n]$ is thus given by

$$v[n] \approx \tilde{p}[n] * h[n]. \quad \text{(A1)}$$

Substituting (A1) into (4), we have

$$b_{AF}[n] \approx e[n](\tilde{p}[n] * h[n]). \quad \text{(A2)}$$

From (A2) and the block diagram in Fig. 6, the Hilbert transform of $b_{AF}[n]$, denoted by $y_{AF}[n]$, is obtained through:

$$y_{AF}[n] \approx e[n](\tilde{p}[n] * h[n]) * g[n]$$

where $g[n]$ is the Hilbert transform operator. Since we are interested in the magnitude and not the phase, we obtain the following approximation [25]:

$$y_{AF}[n] \approx e[n]\tilde{p}[n] * (h[n] * g[n]). \quad \text{(A3)}$$

Letting $\bar{h}[n] = h[n] * g[n]$, (i.e. $\bar{h}[n]$ is the Hilbert transform of $h[n]$):

$$y_{AF}[n] \approx (e[n]\tilde{p}[n]) * \bar{h}[n]. \quad \text{(A4)}$$

Using the same approximation to obtain (A3) from (A2), we obtain:

$$y_{AF}[n] \approx e[n](\tilde{p}[n] * \bar{h}[n]). \quad \text{(A5)}$$

Substituting (A1) into (A5), defining $\bar{v}[n]$ to be the Hilbert transform of $v[n]$, and assuming $e[n] \geq 0$:

$$|y_{AF}[n]| \approx e[n]|\bar{v}[n]| \approx e_A[n]e_F[n]|\bar{v}[n]|. \quad \text{(A6)}$$

In contrast to $|v[n]|$ in (6) which follows individual zero crossings of the formant and thus the formant frequency, $|\bar{v}[n]|$ provides a smooth formant envelope, reflecting its bandwidth but not frequency.

ACKNOWLEDGMENT

The authors thank Daryush Mehta for writing the code to extract the formants.

REFERENCES

[1] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *Biomed. Eng. IEEE Trans. On*, vol. 59, no. 5, pp. 1264–1271, 2012.

[2] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[3] M. Novotny, J. Rusz, R. Cmejla, and E. Ruzicka, "Automatic Evaluation of Articulatory Disorders in Parkinson's Disease," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014.

[4] M. Falcone, N. Yadav, C. Poellabauer, and P. Flynn, "Using isolated vowel sounds for classification of mild traumatic brain injury," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7577–7581.

[5] B. S. Helfer, T. F. Quatieri, J. R. Williamson, L. Keyes, B. Evans, W. N. Greene, T. Vian, J. Lacirignola, T. Shenk, T. Talavage, and others, "Articulatory Dynamics and Coordination in Classifying Cognitive Change with Preclinical mTBI," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] J. C. Mundt, "Voice Acoustic Measures of Depression Severity and Treatment Response Collected Via Interactive Voice Response (IVR) Technology," *J. Neurolinguistics*, vol. 20, no. 1, pp. 50–64, Jan. 2007.

[7] N. Cummins, J. Epps, and E. Ambikairajah, "Spectro-temporal analysis of speech affected by depression and psychomotor retardation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, Vancouver, Canada, 2013, pp. 7542–7546.

[8] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. C. Mundt, "On the Relative Importance of Vocal Source, System, and Prosody in Human Depression," presented at the 2013 IEEE International Conference on Body Sensor Networks, Cambridge, MA, 2013, pp. 1–6.

[9] J. Sundberg, "Acoustic and psychoacoustic aspects of vocal vibrato," *Vibrato*, pp. 35–62, 1995.

[10] T. G. Stockham, "The Application of Generalized Linearity to Automatic Gain Control," *IEEE Trans. Audio Electroacoustics*, vol. AU-16, no. 2, pp. 267–270, Jun. 1968.

[11] W. S. Winholtz and L. O. Ramig, "Vocal tremor analysis with the vocal demodulator," *J. Speech Lang. Hear. Res.*, vol. 35, no. 3, pp. 562–573, 1992.

[12] A. Aronson, W. S. Winholtz, L. O. Ramig, and S. R. Sibler, "Rapid voice tremor, or 'flutter,' in amyotrophic lateral sclerosis," *Ann. Otol. Rhinol. Laryngol.*, vol. 101, pp. 511–518, 1992.

[13] I. R. Titze, "On the relation between subglottal pressure and fundamental frequency in phonation," *J. Acoust. Soc. Am.*, vol. 85, no. 2, Feb. 1989.

[14] R. L. Plant and R. M. Younger, “The interrelationship of subglottic air pressure, fundamental frequency, and vocal intensity during speech,” *J. Voice*, vol. 14, no. 2, pp. 170–177, 2000.

[15] R. A. Lester and B. H. Story, “Acoustic characteristics of simulated respiratory-induced vocal tremor,” *Am. J. Speech Lang. Pathol.*, vol. 22, pp. 205–211, May 2013.

[16] I. R. Titze, “Phonation threshold pressure: A missing link in glottal aerodynamics,” *J. Acoust. Soc. Am.*, vol. 91, no. 5, pp. 2926–2935, May 1992.

[17] J. Kreiman, B. Gabelman, and B. R. Gerratt, “Perception of vocal tremor,” *J. Speech Lang. Hear. Res.*, vol. 46, pp. 203–214, Feb. 2003.

[18] R. L. Horwitz, “Vocal Modulation Features in the Prediction of Major Depressive Disorder Severity,” S.M., MIT, Cambridge, MA, 2014.

[19] N. Malyska, T. F. Quatieri, and D. Sturim, “Automatic Dysphonia Recognition Using Biologically Inspired Amplitude-Modulation Features.,” presented at the Proceedings of the ICASSP, Prague, 2005, pp. 873–876.

[20] A. Robel and X. Rodet, “Efficient Spectral Envelope Estimation and Its Application to Pitch Shifting and Envelope Preservation,” in *Proceedings of the 8th International Conference on Digital Audio Effects*, Madrid, Spain, 2005, pp. DAFX1–DAFX6.

[21] M. Valstar, F. Eyben, S. Schnieder, B. Schuller, B. Jiang, R. Cowie, K. Smith, S. Bilakhia, and M. Pantic, “AVEC 2013 - The Continuous Audio/Visual Emotion and Depression Recognition Challenge,” in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, Barcelona, Spain, 2013, pp. 3–10.

[22] D. D. Mehta, D. Rudoy, and P. J. Wolfe, “Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking,” *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1732–1746, Sep. 2012.

[23] 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 *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 41–48.

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.

[25] T. F. Quatieri, “Phase Estimation with Application to Speech Analysis-Synthesis,” Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1979.