Classification of depression state based on articulatory precision*

Brian S. Helfer[1], *Thomas F. Quatieri*[1], *James R. Williamson*[1],
Daryush D. Mehta[2], *Rachelle Horwitz*[1], *Bea Yu*[1]

[1] MIT Lincoln Laboratory, Lexington, Massachusetts, USA
[2] Ctr. for Laryngeal Surgery & Voice Rehab., Mass. General Hospital, Boston, Massachusetts, USA
[brian.helfer, quatieri, jrw, rachelle.horwitz, bea.yu, daryush.mehta]@ll.mit.edu

Abstract

Neurophysiological changes in the brain associated with major depression disorder can disrupt articulatory precision in speech production. Motivated by this observation, we address the hypothesis that articulatory features, as manifested through formant frequency tracks, can help in automatically classifying depression state. Specifically, we investigate the relative importance of vocal tract formant frequencies and their dynamic features from sustained vowels and conversational speech. Using a database consisting of audio from 35 subjects with clinical measures of depression severity, we explore the performance of Gaussian mixture model (GMM) and support vector machine (SVM) classifiers. With only formant frequencies and their dynamics given by velocity and acceleration, we show that depression state can be classified with an optimal sensitivity/specificity/area under the ROC curve of 0.86/0.64/0.70 and 0.77/0.77/0.73 for GMMs and SVMs, respectively. Future work will involve merging our formant-based characterization with vocal source and prosodic features.

Index Terms: major depressive disorder, motor coordination, articulatory control, vocal biomarkers, formant frequencies

1. Introduction

Major depressive disorder (MDD) has shown itself to be a major public health concern. The World Health Organization found that depressive disorders increased from the fourth-leading cause to the third-leading cause of global burden affecting an individual's quality of life [1]. While the prevalence and burden of MDD has risen, however, diagnosis and monitoring of treatment efficacy continue to be constrained by the limited number of clinically trained physicians globally. This constraint on clinical resources has increased the urgency and necessity of a reliable automated method of depression evaluation. Such a method would aid in determining the efficacy of new medications in clinical trials.

Vocal features are desirable as potential biomarkers of depression as they can be obtained easily (e.g., via telephone), greatly increasing the global accessibility of an automated method toward depression diagnosis and assessment. Certain vocal features have been shown to change with a subject's mental condition and emotional state in depression, including those based on prosody (e.g., fundamental frequency and speaking rate), spectral representations (e.g., mel cepstra), and glottal excitation flow patterns, timing jitter, amplitude shimmer, and aspiration [2–9]. Such features from the voices of depressed subjects relate to auditory perceptions of monotony, hoarseness, breathiness, glottalization, and slurring.

A motivation of the present paper is that neurophysiological changes associated with depression affect motor coordination and therefore the disruption of articulatory control and kinematics [7, 10, 11]. We propose to characterize these disruptions in articulation through 1) modifications of the average formant space (e.g., slurring may compress this space) and 2) modifications of the dynamics of the formants (e.g., agitation may introduce an erratic behavior in a formant track and monotony may reduce the rate of frequency transitions). We anticipate that formant features relating to the precision of motor control in articulation will be noticeably influenced in patients with depression.

The purpose of this study is to assess the performance of Gaussian mixture model (GMM) and support vector machine (SVM) classifiers to determine depression state from features derived from formant frequency trajectories. Our paper is organized as follows. In Section 2, we describe the database, including ratings of depression, and present previous results using this database. In Section 3, we describe our signal-processing methodologies for formant-feature extraction. Section 4 describes our classification approaches, and Section 5 reports results from these approaches. In Section 6, we provide conclusions and projections toward future work.

2. Depression database

2.1. Description

We use a 35-subject database collected by Mundt et al. [2] of subjects with depression treated over a 6-week duration in an open-label, naturalistic study design. The speech data were collected in a depression-severity study that included telephone-based speech recordings from subjects during office visits. Thirty-five physician-referred subjects (20 women and 15 men, mean age 41.8 years) participated in the study. All of the subjects had recently started on pharmacotherapy and/or psychotherapy for depression and continued in treatment over the 6-week assessment period of the study.

Speech recordings (sampled at 8 kHz) were collected over a standard office telephone on weeks 0, 2, 4, and 6. In the current work, we analyze speech samples of conversational (free-response) speech and sustained vowels (/a/, /i/, /u/, /ae/) recorded over telephone while the subjects were present in the clinic. Approximately 3–6 minutes of speech per session (i.e., per day) were available for analysis.

The subjects were interviewed in-office by a clinician to obtain ratings of depression on the 17-item Hamilton depression rating scale (HAM-D) [12]. Total HAM-D scores are the sum of the individual 17 symptoms ratings, which

*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. The work of R. Horwitz was supported in part by The National Institutes of Deafness and Other Communicative Disorders under Grant T32 DC00038.

reflect items such as depressed mood and psychomotor retardation, with each symptom scored on a scale from 0 to 2 or 0 to 4.

2.2. Previous results on the Mundt database

Mundt et al. [2] investigated correlations of variance of pitch and numerous parameters that relate to average speaking and pause rates with HAM-D scores and an alternative depression scale, the quick inventory of depressive symptomatology [13]. In that study, average speaking rate and pause parameters were found to have correlations in the approximate range $-0.20 < r < 0.20$ with p-values less than 0.05. Mundt et al. also found a correlation between the variance of the second formant with the total HAM-D score ($r{\sim}0.17$, $p < 0.05$), but no corresponding correlation for the first formant (the two formants studied).

In other work [8], average measures of speaking rate were categorized into phone-specific characteristics. Combined phone-duration measures uncovered stronger relationships between speaking rate and depression severity than the global measures previously reported as speech-rate biomarkers. More importantly, a general framework for analyzing individual symptom categories through phonological units was developed and showed that a biomarker (e.g., speaking rate) can correlate with sub-symptoms (e.g., psychomotor retardation) more highly than with a total assessment. Other correlation studies on the Mundt database have involved shimmer, jitter, and harmonics-to-noise ratio [9]. One classifier study involved a GMM-based classifier using mel-cepstral features obtaining approximately 80% sensitivity and 20% specificity for the binary classification problem (depressed or not depressed) [14].

3. Formant tracking and analysis

In this section, we describe our formant-feature extraction methodology.

3.1. Formant tracking algorithm

There are a variety of approaches to the on-going challenge of formant estimation and tracking. We have selected an algorithm based on the principle that formants are correlated with one another in both frequency and time [15, 16]. Figure 1 shows an example of formant-track outputs, where the shaded areas around each track represent the 3-dB formant bandwidth. Embedded in the algorithm is a speech-activity detector that enables a Kalman predictor to smoothly coast consistently through non-speech regions. All measurements are made during frames with speech activity.

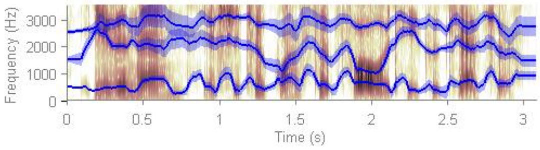

Figure 1: *Example spectrogram of the first three formant tracks. Shaded blue region around each track depicts 3-dB formant bandwidth.*

3.2. Formant Features

In characterizing formants and their dynamics, we extract nine features per formant per 20-ms segment at a 10-ms frame interval, which is a basis for feature selection for our classifiers. Figure 2 defines these features as samples of the raw formant frequency track, its lowpass component, its highpass component, and the velocity and acceleration of those trajectories. Derivatives are computed by averaging first forward and backward differences across frames. This provides a form of dynamics of each formant, their low-frequency target trajectories, as well as their variability around these targets.

The high-frequency component captures the transitional formant properties that are more high-pass in nature, whereas the low-frequency component captures formant properties that are perhaps more semantic-based, representing a smoothed rendition of underlying planned targets. The two formant components may therefore reflect different effects of neurophysiological change in depression. This lowpass/highpass paradigm was inspired by a prosodic modeling approach of Fujisaki and Hirose [17]. Figure 3 displays an example output of the lowpass/highpass track decomposition.

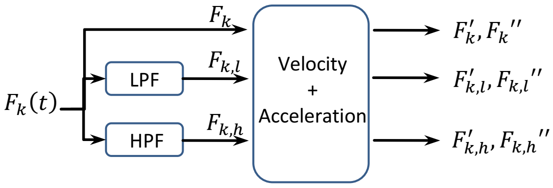

Figure 2: *Derivation of nine features per formant (k) per frame (shown here for* F_k *over time* t; $k \in \{1, 2, 3\}$*). After applying a lowpass filter (LPF) and highpass filter (HPF) to the formant trajectories, velocities (') and accelerations ('') are also computed. The common 3-dB cutoff frequency of the filters is 55 Hz.*

4. Classification of depression state

Recent studies have investigated correlations of vocal features with depression and have begun to incorporate them into the development of classifiers [4, 5, 7, 10, 14, 18]. Formants and their dynamics, however, have not been explicitly analyzed. To investigate the effectiveness of classifying depression state with a formant characterization, we make use of both a generative and discriminative classifier, using truth labels specifying each patient-session as depressed or not depressed. Subject sessions were labeled as depressed if their HAM-D score was 17 or greater, and as non-depressed if their score was 7 or lower, with scores of 8 to 16 excluded because their depression status is ambiguous. Four patients whose scores all ranged between 8 and 16 were excluded. The generative classifier is a Gaussian mixture model (GMM), and the discriminative model is a support vector machine (SVM).

We evaluate classifier performance through the analysis of two sections of the clinical interview: vowels and free response. Results are analyzed in terms of sensitivity (true positive prediction rate), and specificity (true negative prediction rate). Our performance is also summarized using a ROC curve, which allows us to balance the sensitivity and specificity of our classifiers, as well as report the area under the curve (AUC). AUC is an informative statistic describing the average tradeoff between sensitivity and specificity.

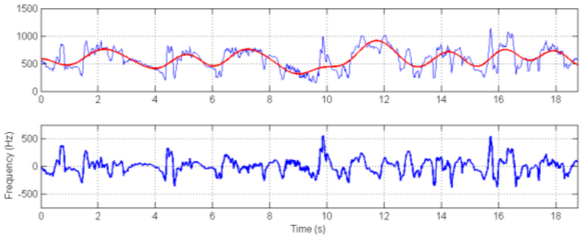

Figure 3: *Example of the lowpass/highpass filtering process on the first formant track estimated from a speech segment. Upper panel: raw formant track (blue) and low-passed component (red); Lower panel: high-passed component. Original formant tracks coast through non-speech regions enabled by a Kalman predictor.*

4.1. Feature selection

Formant features are categorized into three different groups: 1) *Vowels*—the average F_k extracted from sustained vowels /a/, /i/, /u/, /ae/; 2) *Free Formant*—the frame-sampled formant estimates F_k, along with their low-passed $(F_{k,l})$ and high-passed $(F_{k,h})$ components extracted from the free response section; 3) *Free Dynamic*—the velocity (F_k') and acceleration (F_k'') versions of the frame-sampled formant trajectories, and their respective filtered components $(F_{k,l}', F_{k,h}'$ and $F_{k,l}'', F_{k,h}'')$ extracted from the free response section. Within these three groups, we explore combinations of features for classification.

In order to reduce dimensionality and prevent over-fitting, we select group-specific formant features based on joint mutual information and classifier performance metrics of sensitivity, specificity, and AUC. For both classifiers, these criteria resulted in selecting average F_1 estimates within the *Vowels* group and F_3, $F_{1,l}$, and $F_{3,l}$ estimates within the *Free Formant* group. Finally, for the *Free Dynamic* group, $F_{k,l}'$ and $F_{k,h}'$ $(k \in \{1,2,3\})$ are selected for the GMM, and $F_{k,l}''$ and $F_{k,h}''$ $(k \in \{1,3\})$ are selected for the SVM classifier.

The group-specific formant features are then Z-normalized and passed into a principle component analysis (PCA) algorithm. A subset of the principle components are used to train the classifiers. The *Vowels* group is represented by the first two PCA dimensions in both classifiers. The *Free Formant* group is represented by the first PCA dimension in both classifiers. Finally, the *Free Dynamic* group is represented by the first three PCA dimensions in the GMM and the first four PCA dimensions in the SVM classifier.

Finally, as a comparative reference, we study three acoustic features developed by Sapir and colleagues that characterize "vowel centralization" for differentiating dysarthric speech in subjects with and without Parkinson's disease [19–21]. These features are the Formant Centralization Ratio: $FCR = (F_{2u} + F_{2a} + F_{1i} + F_{1u})/(F_{2i} + F_{1a})$, its inverse the Vowel Articulation Index: $VAI = (F_{2i} + F_{1a})/(F_{2u} + F_{2a} + F_{1i} + F_{1u})$, and the ratio of the second formants of /i/ and /u/: F_{2i}/F_{2u}.

4.2. Gaussian mixture model classifier

Our generative modeling approach, adapted from [22], is to form a background statistical model of all the subjects in the database and then to form specific models for depressed and non-depressed subjects. The statistical models are Gaussian mixture models (GMMs), weighted combinations of multiple Gaussian densities. The Gaussian densities are defined using diagonal covariance matrices, where ten mixture components are used. These parameters are obtained using one iteration of the expectation-maximization algorithm on the training data set. The depressed and non-depressed GMMs are obtained from the background GMM using a Bayesian adaptation procedure [22], which is applied to the depressed and non-depressed training data subsets, respectively. The GMMs are evaluated on each subject using cross-validation that leaves all sessions from the test subject out of the training set.

To obtain robustness to small data sets, the likelihoods from ten independently trained GMMs are combined. (Use of 5 and 20 GMMs reduced performance.) Specifically, ten different background GMMs are obtained using independent random initializations, resulting in the adaptation of ten depressed and non-depressed GMMs. The single-frame prediction score is the 2-class log-likelihood ratio obtained from the log of the sum of the ten depressed GMM likelihoods minus the log of the sum of the ten non-depressed GMM likelihoods. Multi-frame classification scores are obtained by averaging these log-likelihood ratios across frames. Predictions from separate feature groups are fused using a weighted combination of the log-likelihood ratios. Log-likelihood ratios obtained from the groups extracted from the free response section are weighted ten times more than ratios taken from the group extracted from sustained vowels.

4.3. Support vector machine classifier

Our discriminative modeling approach makes use of a two-class support vector machine (SVM). SVM classification is done in a supervised manner through a training phase where the model creates a classification boundary to distinguish between the two classes. An SVM classifies depressed and non-depressed subjects by forming a decision boundary that creates the maximum possible separation between classes in the feature space. The SVM is implemented using the models developed by Chang et al. [23] and Fan et al. [24].

As with our GMM classifier, training and testing for the SVM are done using cross-validation, with a single subject being held out for testing at each iteration of model training. The classifier is penalized more heavily for misclassifying non-depressed sessions (versus misclassifying depressed sessions) to account for differences in available data within each group.

The subject responses from the free response section are classified on a frame-by-frame basis. For each frame, the distance from the SVM hyperplane (the decision value) is determined. The decision values are averaged across frames to allow for a prediction to be made for each session. In order to fuse the separate feature groups, decision values are normalized by the maximum value occurring within a group. The normalized decision values are then averaged across the groups.

5. Results

Performance is reported using both sensitivity and specificity, along with the area under the ROC curve (AUC). Figure 4 illustrates the relative performance within each feature group. The GMM shows a marginally higher AUC for the *Vowels* group, whereas the SVM shows a marginally higher AUC on

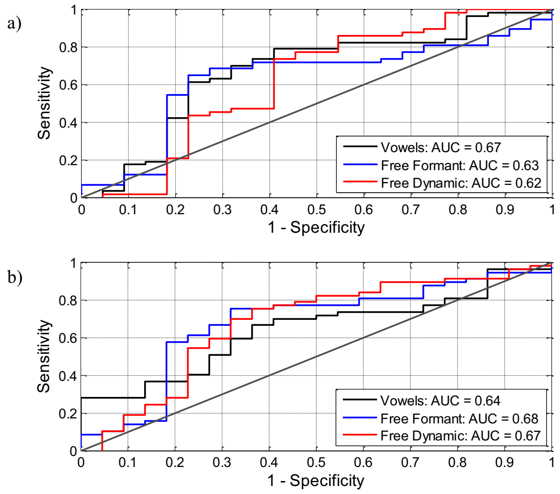

Figure 4: *ROC curves comparing distinct feature groups using a) GMM and b) SVM classifiers.*

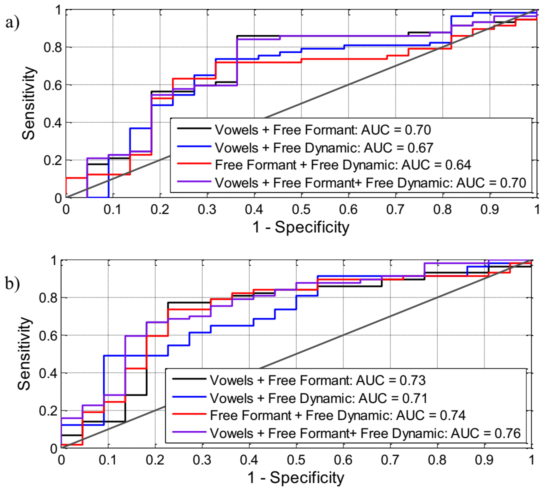

Figure 5: *ROC curves comparing a) GMM and b) SVM classifier performance when feature groups are fused post-classification.*

the features extracted for the *Free Formant* and *Free Dynamic* groups.

To potentially improve performance, the classifier outputs are then fused using multiple combinations of the three feature groups. Along with the AUC, the optimal sensitivity and specificity (Se/Sp)—defined as the point on the ROC curve maximizing the sum of the two values—is reported. We also report the sensitivity at a fixed specificity of 0.70 ($\mathrm{Se}_{\mathrm{Sp}0.70}$). Table 1 reports these metrics, with the corresponding ROC curves shown in Figure 5.

The results presented in Table 1 and Figure 5 demonstrate the differences in performance seen when combining predictions from distinct feature groups for a single classifier. Maximum AUCs of 0.70 and 0.76 are obtained for the GMM and SVM, respectively, when predictions from all three groups are fused. These AUCs indicate an increased range for sensitivity and specificity as the operating point of the ROC is varied. Optimal Se/Sp is 0.86/0.64 for the GMM and 0.77/0.77 for the SVM, obtained through fusing the *Vowels* group with the *Free Formant* group.

Comparing Figure 5 with Figure 4, we see potential differences in classifier performance when incorporating dynamic aspects with the raw formant frequencies from the free response section. For the SVM, the AUC is 0.74 for the *Free Formant + Free Dynamic* fusion versus 0.68 for the *Free Formant* group alone and 0.67 for the *Free Dynamic* group alone. This gain is less definitive for GMM performance.

Groups fused	Classifier	AUC	Se/Sp	$\mathrm{Se}_{\mathrm{Sp}0.70}$
Vowels + Free Formant	GMM	0.70	0.86/0.64	0.60
Vowels + Free Dynamic	GMM	0.67	0.74/0.68	0.65
Free Formant + Free Dynamic	GMM	0.64	0.68/0.72	0.63
All groups	GMM	0.70	0.84/0.64	0.60
Vowels + Free Formant	SVM	0.73	0.77/0.77	0.77
Vowels + Free Dynamic	SVM	0.71	0.65/0.68	0.61
Free Formant + Free Dynamic	SVM	0.74	0.74/0.77	0.74
All groups	SVM	0.76	0.82/0.66	0.70

Table 1: *Summary of classification performance.*

Finally, using the vowel centralization features FCR, VAI, and F_{2i}/F_{2u} individually, depression classification yields AUCs below 0.5, indicating that these features may not be useful for the assessment of depression state.

6. Conclusions and on-going work

Given that neurophysiological changes due to major depressive disorder influence the articulatory precision of speech production [7, 10, 11], we investigated vocal tract formant frequencies and their velocity and acceleration toward automatic classification of depression state. We analyzed formant characteristics of 31 subjects who were rated using clinical measures of depression severity [2]. The relative importance of formant features were explored in the context of both a generative GMM classifier and a discriminative SVM classifier of binary depression state. The GMM and SVM achieved a sensitivity/specificity/AUC of 0.86/0.64/0.70 and 0.77/0.77/0.73, respectively. We have also provided initial evidence for including both formant trajectories and their dynamic features to more accurately classify depression state.

These are important findings given that only the first three formant frequencies and their dynamic characterization were used as a basis for features, leaving the possibility of improved classification accuracy by incorporating other aspects of speech production. Our on-going work involves improving the current feature estimation and selection methods, merging our formant-based features with other formant characteristics and vocal source and prosodic features, and validating and refining the classifiers and feature combinations using a larger similarly recorded database. Such enhancements in this research may provide a quantitative assessment of drug efficacy in a clinical setting.

7. Acknowledgements

The authors thank Dr. Nicolas Malyska from MIT Lincoln Laboratory and Dr. James Mundt from the Center for Psychological Research, Training, and Consultation, Madison, WI, for assistance with data structures and helpful discussions. The authors also acknowledge the National Institute of Mental Health who supported Dr. James Mundt in the data collection used in this article.

8. References

[1] Ferrari, J. Somerville, J., Baxter, J., Norman, R., Patten, S. B., Vos, T. and Whiteford, H. A., "Global variation in the prevalence and incidence of major depressive disorder: A systematic review of the epidemiological literature," Psychological Medicine, 43(3):471–481, 2013.

[2] Mundt, J., Snyder, P., Cannizaro, M. S., Chappie, K. and Geralts, D. S. "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," Journal of Neurolinguistics, 20(1):50–64, 2007.

[3] France, D., Shiavi, R., et al., "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE Transactions on Biomedical Engineering, 47(7):829–837, 2000.

[4] Low, L. A., Maddage, T., Lech, M., Sheeber, L. and Allen, N., "Influence of acoustic low-level descriptors in the detection of clinical depression in adults," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2010.

[5] Low, L. A., Maddage, T., Lech, M., Sheeber, L. and Allen, N. "Mel frequency cepstral feature and Gaussian mixtures for modeling clinical depression in adolescents," Proceedings of the IEEE 8th International Conference on Cognitive Informatics, 2009.

[6] Moore II, E., Clements, M., Peifer, J. and Weisser, L., "Analysis of prosodic variation in speech for clinical depression," Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, 2003.

[7] Ozdas, A., Shiavi, R., Silverman, S., Silverman, M. and Mitchell, D., "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," IEEE Transactions on Biomedical Engineering, 51(9):1530–1540, 2004.

[8] 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.

[9] Quatieri, T. F. and Malyska, N., "Vocal-source biomarkers for depression: A link to psychomotor activity," Proceedings of Interspeech, 2012.

[10] Caligiuri, M. P. and Ellwanger, J., "Motor and cognitive aspects of motor retardation in depression," Journal of Affective Disorders, 57(1–3):83–93, 2000.

[11] Sobin, C. and Sackeim, H. A., "Psychomotor symptoms of depression," American Journal of Psychiatry, 154:4–17, 1997.

[12] Hamilton, M., "A rating scale for depression," Journal of Neurology, Neurosurgery, and Psychiatry, 23:56–62, 1960.

[13] Rush, A. J., et al. "The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression," Biological Psychiatry, 54(5):573–583, 2003.

[14] Sturim, D., Torres-Carrasquillo, P., Quatieri, T. F., Malyska, N. and McCree, A. "Automatic detection of depression in speech using Gaussian mixture modeling with factor analysis," Proceedings of Interspeech, 2011.

[15] Rudoy, D., Spendley, D. N. and Wolfe, P. J., "Conditionally linear Gaussian models for estimating vocal tract resonances," Proceedings of Interspeech, 526–529, 2007.

[16] Mehta, D. D., Rudoy, D. and Wolfe, P. J., "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," The Journal of the Acoustical Society of America, 132(3):1732–1746, 2012.

[17] Fujisaki, H. and Hirose, K., "Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation," Preprints of Papers, Working Group on Intonation, 13th International Congress of Linguists, 57–70, 1982.

[18] Moore II, E., Clements, M. A. and Peifer, J. W., "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," IEEE Transactions on Biomedical Engineering, 55(1):96-107,2008.

[19] Sapir, S., Ramig, L. O., Spielman, J. L. and Fox, C., "Formant Centralization Ratio (FCR): A proposal for a new acoustic measure of dysarthric speech," Journal of Speech, Language, and Hearing Research, 53(1):114–125, 2010.

[20] Sapir, S., Spielman, J. L., Ramig, L. O., Story, B. H. and Fox, C., "Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings," Journal of Speech, Language, and Hearing Research, 50(4), 899–912, 2007.

[21] Hazan, H., Hilu, D., Ramig, L., Sapir, S. "Early diagnosis of Parkinson's disease via machine learning on speech data," Proceedings of the IEEE 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI), 1–4, 2012.

[22] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, 10(1–3):19–41, 2000.

[23] Chang, C. and Lin, C., "LIBSVM: A library for support vector machines," 1–39, 2012.

[24] Fan, R., Wang, X. and Lin, C., "LIBLINEAR: A library for large linear classification," 9:1871–1874, 2008.