ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity

Thomas F. Quatieri and Nicolas Malyska

MIT Lincoln Laboratory, Lexington MA

[quatieri,nmalyska]@ll.mit.edu

Abstract[1]

A hypothesis in characterizing human depression is that change in the brain‟s basal ganglia results in a decline of motor coordination [6][8][14]. Such a neuro-physiological change may therefore affect laryngeal control and dynamics. Under this hypothesis, toward the goal of objective monitoring of depression severity, we investigate vocal-source biomarkers for depression; specifically, source features that may relate to precision in motor control, including vocal-fold shimmer and jitter, degree of aspiration, fundamental frequency dynamics, and frequency-dependence of variability and velocity of energy. We use a 35-subject database collected by Mundt et al. [1] in which subjects were treated over a six-week period, and investigate correlation of our features with clinical (HAMD), as well as self-reported (QIDS) Total subject assessment scores. To explicitly address the motor aspect of depression, we compute correlations with the *Psychomotor Retardation* component of clinical and self-reported Total assessments. For our longitudinal database, most correlations point to statistical relationships of our vocal-source biomarkers with psychomotor activity, as well as with depression severity.

Index Terms: major depressive disorder, motor coordination, laryngeal control, vocal biomarkers

1.0 Introduction

MAJOR DEPRESSIVE DISORDER (MDD) is the most widely affecting of the mood disorders; the lifetime risk has been observed to fall between 10-20% for women and 5-12% for men [2]. Accurate diagnosis of MDD requires intensive training and experience. Thus the growing global burden of depression suggests that an automatic means to monitor depression severity would be a beneficial tool for patients, clinicians, and healthcare providers. For example, such a tool would be useful in monitoring the effects of new treatments. Reliable classifiers could also be used as a tool to aid in the standardization of depression ratings. One such approach relies on the extraction of biomarkers to provide reliable indicators of depression.

A class of biomarkers of growing interest is the group of vocal features observed to change with a patient‟s mental condition and emotional state, motivated by perception of monotony, hoarseness, breathiness, glottalization, and slur in the voice of a depressed subject. Vocal characteristics studied include prosody (e.g., fundamental frequency and speaking rate), spectral features, and glottal (vocal fold) excitation flow patterns, timing jitter, amplitude shimmer, and „spirantization,‟ a measure that reflects aspirated leakage at the vocal folds [1][3-7]. Although not always consistent across studies, vocal features have been shown to bear statistical relationships with the presence of depression, and in some cases have been applied towards developing automatic classifiers. Discrepancies across studies are due to differences in patient population (no two studies have used the same population), small data sets, different forms of depression (“agitated” and “slowed”), and differences in signal processing methods for vocal feature extraction [1][3-7].

A hypothesis in the voice quality study of this paper is that neuro-physiological change in depression generally affects motor coordination including laryngeal control and dynamics [6][8][14]. Under this hypothesis, source features are selected that may relate to precision in motor control in source generation, including shimmer and jitter of vocal-fold vibration, degree of aspiration, dynamics of the fundamental frequency (henceforth termed „pitch‟), and frequency-dependence of variability and velocity of energy. We use a 35-subject database collected by Mundt et al. [1] of subjects treated for depression over a 6-week duration, and investigate correlation of our features with clinical HAMD, as well as self-reported QIDS Total subject assessment scores. To more explicitly address the motor aspect of depression, we also compute correlations with the Psychomotor Retardation [2] component of each Total assessment.

Although our significant ($p < 0.05$) correlations reported are low in magnitude[2], they point to statistical relationships of our vocal-source biomarkers with psychomotor activity, as well as with depression severity: With increasing depression severity *and* Psychomotor Retardation, there is tendency for an increase in shimmer, jitter, and aspiration, as well as for more variable pitch and energy dynamics. For pitch, we investigate its variance and average velocity. While for both full- and multi-band energy, we measure correlations of energy variance and velocity with depression assessments, and also explore their frequency-dependence with inverse-filtered speech to move closer to the source. In vocal-feature extraction, we use a variety of signal processing methodologies. For jitter, we rely on inter-pulse-intervals as estimated by Mehta et al [9]; while for shimmer we use an approach developed by Boersma and Weenink [10]. Our aspiration estimate is obtained from the Jackson/Shade harmonic/noise separator [11]; and pitch and its derivatives are estimated using a sinusoidal-based method [12].

Our paper is organized as follows. In Section 2, we describe the 35-subject depression database collected by Mundt et al. [1]. In Section 3, we describe our signal-processing

[1] 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.

[2] The range of Spearman correlation magnitudes in this paper (~0.10-0.20), derived from average statistics, is consistent with previous measuresments by Mundt et al. for the same database [1].

10.21437/Interspeech.2012-311

methodologies for vocal-feature extraction: shimmer and jitter, aspiration, and pitch and energy dynamics. In Section 4, we give correlation results and finally in Section 5 provide conclusions and projections to future work.

2.0 Depression Database

2.1 Database

The data used in this analysis was originally collected by Mundt et al. [1] for a depression-severity study, involving both in-clinic and telephone-response speech recordings. Thirty-five physician-referred subjects (20 women and 15 men, mean age 41.8 years) participated in this study. The subjects were predominately Caucasian (88.6%), with four subjects of other descent. The subjects had all recently started on pharmacotherapy and/or psychotherapy for depression and continued treatment over a 6-week assessment period. Speech recordings (sampled at 8 kHz) were collected at weeks 0, 2, 4, and 6 during an interview and assessment process that involved clinical Hamilton rating (HAMD) and self-reported (QIDS) scoring. To avoid telephone-channel effects, only the samples of conversational (free-response) speech and distinct vowels (/a/, /e/, i/, /o/) recorded in the clinic are used in our follow-up work, the former for our prosodic (i.e., pitch and energy dynamics) and the later for our shimmer, jitter, and aspiration measurements. We used only data from subjects that completed the entire longitudinal study. This resulted in approximately 3-6 minutes of speech per session (i.e., per day).

Within this database, the standard method of evaluating levels of MDD in patients was invoked using the clinical 17-question HAMD assessment [1]. To determine the Total score, individual ratings are first determined for the 17-symptom sub-topics (such as Mood, Guilt, Psychomotor Retardation, Fatigue, Suicidal Tendency, etc.) with scores for component sub-topics having ranges of (0-2), (0-3), or (0-4). The total score is then the aggregate of the ratings for all sub-topics. A similar assessment methodology is used for self-reported (QIDS) scoring. These scores are used as our truth markings in calculating correlations. Although the HAMD and QIDS assessments are standard depression evaluation methods, there is some concern about their reliability. Nevertheless, addressing this concern is outside our scope.

2.2 Previous results

A study by Mundt et al. [1] investigated correlations of variance of pitch and numerous parameters that relate to average speaking and pause rate with HAMD and QIDS Total assessments (no sub-symptom components). While finding significant ($p < 0.05$) correlations (in the approximate range $-0.20 < r < 0.20$) with average rate and pause parameters, Mundt et al. did not find significant correlations using variance of pitch. In a second related work using the Mundt et al. database [7], average measures of speaking rate were dissected into phone-specific characteristics and, in particular, combined phone-duration measures to uncover stronger relationships between speaking rate and depression severity than global measures previously reported for a speech-rate biomarker. In this work, other speech characteristics were not considered.

3.0 Signal-processing Methodologies

3.1 Shimmer, jitter, and aspiration

Jitter is the period-to-period variation in glottal pulse timing during voicing. To estimate this quantity, we first extract approximate glottal pulse times for voiced regions using the To-Pitch function, followed by the To-PointProcess function, both from the speech-signal processing tool Praat [10]. Jitter values are found by calculating the average absolute difference between consecutive intervals, dividing by the average length, and multiplying by 100 to yield a percent value.

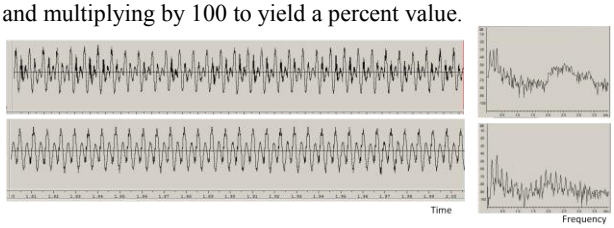

Figure 1. *A vowel /i/ example illustrating a decrease in shimmer, jitter, and aspiration from the first (day 0) (upper panels) to the last (day 42) (lower panels) recording session for one subject from the Mundt et al. database [1].*

Day	Shimmer	Jitter	HNR (dB)
0	16.8	1.2	13.5
42	5.8	0.6	34.2

Table 1. *Decrease in shimmer, jitter, and aspiration from the first (day 0) to the last (day 42) recording session for the vowel /i/ from a subject from the Mundt et al. database. HNR is harmonic-to-noise ratio derived from the Jackson/Shadle separation algorithm.*

Shimmer measures the period-to-period variation in glottal pulse amplitude in voiced regions. To measure this quantity, we first extract approximate glottal pulse times for voiced regions using the Praat function described above. Amplitudes at the pulse times are then calculated using the absolute value of the output of the To-AmplitudeTier function. Shimmer values are found by calculating the average absolute difference between the amplitudes of consecutive periods, dividing by the average amplitude, and multiplying by 100 to yield a percent value.

To measure aspiration, we use a harmonic/noise decomposition technique referred to as *pitch-scaled harmonic filtering* (PSHF) [11]. The PSHF approach uses an analysis window duration equal to four pitch periods and relies on the property that harmonics of the fundamental frequency fall at specific frequency bins of the short-time Fourier transform. Spectral subtraction is subsequently performed to obtain the noise component spectrum. Pitch periods are estimated using the speech-signal processing tool Praat [10]. This decomposition technique approximately isolates the noise component in an aspirated utterance. To minimize leakage of harmonicity into the extracted noise component, our correlation analysis is restricted to the vowel recordings of the Mundt et al. database. The resulting voiced and aspiration components are then used to form a harmonics-to-noise ratio.

A vowel /i/ example is given in Table 1 and Figure 1 illustrating a decrease in shimmer, jitter, and aspiration from the first (day 0) to the last (day 42) recording session for one subject from the Mundt et al. database [1]. The speech waveform and spectral slices show shimmer, jitter, and aspiration consistent with the values in the table.

3.2 Fundamental frequency and energy dynamics

Fundamental frequency: We investigate two different measures of pitch variability: Pitch variance and average pitch velocity. Pitch estimation is performed with a sinusoidal-based algorithm [12], and pitch measurements are made in voiced regions as derived from the same sinusoidal-based algorithm. For each utterance in our database, the pitch variance is estimated as the mean-squared pitch deviation from the mean, while the average pitch velocity is estimatd as the average magnitude of the first-central-pitch difference.

Energy: We investigate four different measures of energy variability: Energy variance and average energy velocity, as well as the same characteristics after inverse filtering. Energy is estimated as the mean-squared signal values, while its

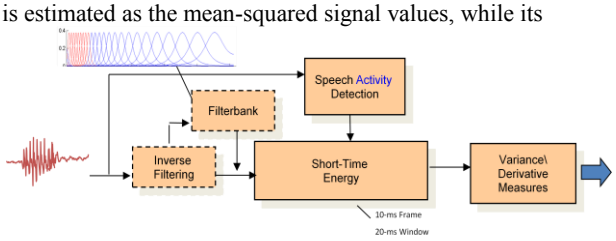

Figure 2. *Short-time energy measurements on speech and inversion.*

velocity estimate is the mean-squared central difference across 10-ms frames with a 20-ms window. Energy measurements were made in active speech frames, determined according to an energy-based speech activity detector. Inverse filtering was performed with the standard autocorrelation method of linear-prediction with 10 poles [10][12]. We also perform a multi-band energy measurement (Figure 2) where the speech signal is decomposed by 24 auditory-like Gammatone bandpass filters [12] and the energy variance and derivative of each filter output is computed.

4.0 Correlation Results

Shimmer, jitter, and aspiration measurements tend to be most accurate during steady voicing. Consequently, for this data class, we work with only four vowels recorded in the Mundt database: /a/, /i/, /o/, /u/. Since there is little data with each alone, correlations are made with all four vowels as a group. For pitch and energy dynamics, however, we use speech recorded from conversation in the Mundt el al. database because this option better captures these measurements. Spearman was chosen over Pearson correlation due to the quantized ranking nature of the HAMD and QIDS depression scores and the possible non-linear relationship between score and speech features [13].

4.1 Shimmer and jitter

Figure 3 gives Spearman correlations with our shimmer measurements for the clinical HAMD and self-reported QIDS Total assessments, as well as for the Phychomotor-Retardation sub-symptom component. All correlations in this case are significant, where significance is defined as the p value being less than a threshold of 0.05, i.e., $p < 0.05$.

With positive correlations, our interpretation is that shimmer is increasing with increasing overall depression severity, as well as with increasing Psychomotor Retardation as a sub-symptom. For our jitter measurements (not shown in Figure 3), on the other hand, the only significant correlation occurs with the Total clinical HAMD assessment (r~0.11 with p~0.03). This is perhaps due to the difficulty in measuring jitter in the presence of strong aspiration, typical of depressed subjects.

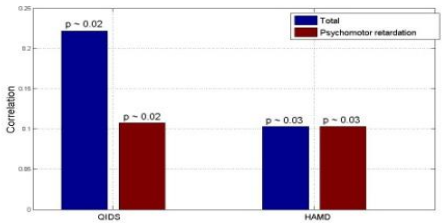

Figure 3. *Spearman correlations with shimmer for the clinical HAMD and self-reported QIDS Total assessments, as well as for the Psychomotor Retardation sub-symptom component. Significance (the p value) is given above each bar.*

4.2 Aspiration

As described in Section 3, aspiration in the acoustic signal is measured using the Jackson/Shadle algorithm [11]. Figure 4 gives Spearman correlations for the clinical

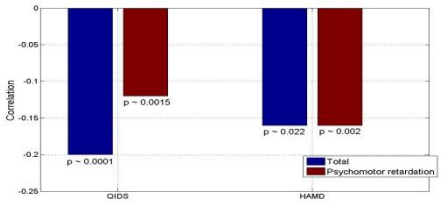

Figure 4. *Spearman correlations with harmonic-to-noise ratio (aspiration) for the clinical HAMD and self-reported QIDS Total, as well as for the Psychomotor Retardation assessment.*

HAMD and self-reported QIDS Total assessments, as well as for the Psychomotor-Retardation sub-symptom component. All correlations in this case are again significant.

With negative correlations, our interpretation is that harmonics-to-noise ratio is decreasing and thus aspiration is increasing with increasing overall depression severity, as well as with increasing Psychomotor Retardation as a sub-symptom. This may be consistent with the presence of motor retardation in depression reducing laryngeal muscle tension thus resulting in a more open, turbulent glottis.

4.3 Pitch and energy

In this section, we investigate the correlation of pitch and energy dynamics with Total and Psychomotor Retardation assessments. Correlations with average pitch and energy are not included because they were not found significant for any condition. For energy, we also perform a multi-band decomposition of the speech and its inverse-filtered rendition, thus providing a more direct measurement of energy variability of the source.

Pitch variability: Figure 5 shows correlations of pitch variance and average velocity with our selected assessments; three correlations do not show significance. Nevertheless, the general trend is a positive correlation of pitch dynamics with both assessments.

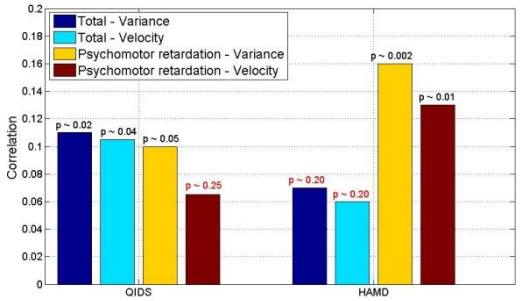

Figure 5. *Correlations of pitch variance and average velocity; three correlations (with p values in red) do not show significance.*

Specifically, positive correlations indicate that pitch variance and average velocity are *decreasing* with decreasing depression severity. This is in contrast to previous work of another study with a different database (variance only for control vs depressed) that shows decreased pitch variability with depression [15], i.e., the depressed voice is more „monotonous" than the control. For the database of our study, Mundt et al. did not find significant correlation of pitch variance with Total assessments [1]; this discrepancy with our results may be due to the use of read speech in the Mundt et al. study, in contrast to our use of the conversational component.

Energy variability: Table 2 shows correlations of energy variance and average velocity with our selected assessments. A number of observations can be made. First, the uniformly negative correlations of energy variance with both Total and

Measure	QIDS Total	HAMD Total	QIDS PR	HAMD PR
Variance (S)	-0.07/0.16	-0.08/0.12	-0.06/0.27	-0.15/0.003
Variance (I)	-0.17/0.15	-0.10/0.05	-0.16/0.25	-0.15/0.005
Velocity (S)	0.18/0.0006	0.17/0.0007	0.11/0.04	0.05/0.40
Velocity (I)	0.18/0.005	0.18/0.006	0.18/0.04	0.19/0.001

Table 2. *Correlations of energy variance and average velocity with Total and Psychomotor Retardation (PR) assessments for speech (S) and its inversion (I); correlations in blue are significant ($p < 0.05$).*

Psychomotor Retardation assessment indicate that the variance is *decreasing* with increasing depression severity. Though generally with $p > 0.05$, the trend is consistent with previous studies with different (control vs depressed) databases that show decreased energy variability with depression [15]. Next, the uniformly positive correlations of energy velocity with both Total and Psychomotor Retardation (primarily significant) indicate that the velocity is *increasing* with increasing depression severity, perhaps an indication that motor coordination improves in the less-depressed state.

We have also explored correlation properties associated with the multi-band (auditory-like) decomposition of speech (described in Section 3.2). As with full-band energy, we computed correlations with multi-band energy variance and average energy velocity in each band. We have also computed these same correlations with our inverse-filtered rendition of the signal, thus providing a closer look at frequency-dependence of source energy. (The gain of the inverse filter was normalized to unity, thus assuming all gain is imparted by the source.) Generally, we have found that for each condition, significant correlations can differ in each band both in magnitude and sign. As examples, Figure 6 shows correlations for the average energy derivative with QIDS Total assessment using speech and its source estimate, and also correlations with the Psychomotor Retardation assessment using the source estimate. Uniformly positive high-frequency correlations with the source estimates imply decreased energy velocity in this region with subject improvement.

6.0 Conclusions and Future Work

In this paper, we investigated vocal source features as biomarkers for depression severity. Specifically, source features were selected that may relate to precision in motor control, including shimmer and jitter of vocal-fold vibration, degree of aspiration, fundamental frequency dynamics, and frequency-dependence of variability and velocity of energy. Correlation of our features with clinical HAMD, as well as QIDS self-reported assessment were presented. To further address our hypothesis of the effect of neuro-physiological change, we also computed correlations with the Psychomotor Retardation component of each assessment. For data over a six-week therapy period, correlation results point to statistical relationships of our vocal-source biomarkers with psychomotor activity, as well as with depression severity.

Results of this paper stimulate a number of important areas of future research. Alternative vocal-source features, both existing (e.g., glottal flow characterization [5]) and novel (e.g., measures of extent and style of glottalization) should be explored for consistency and for statistical- and physiologically-based inter-relationships, along with further study of those presented in this paper. Clearly, establishing stronger significance in these relationships requires a larger database of subjects and, for therapeutic studies, separation

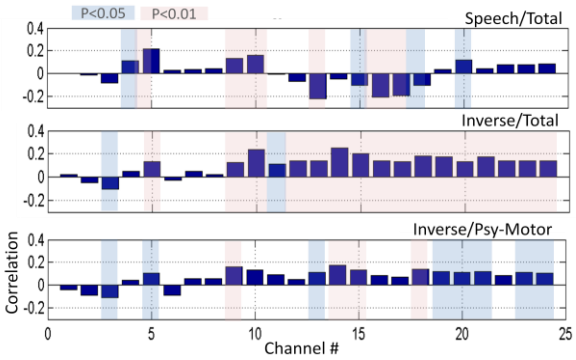

Figure 6. *Frequency-dependent correlations of average energy velocity with QIDS assessments. Top panel: Correlation with Total assessment (from speech); Middle panel: Correlation with Total assessment (from source estimate); Lower panel: Correlation with Psychomotor Retardation assessment (from source estimate). Regions with significance are depicted by blue-shaded ($p < 0.05$) and red-shaded ($p < 0.01$) rectangles.*

of responders from non-responders to treatment, both useful toward our ultimate objective of designing predictors and classifiers of depression state. Finally, our correlations of laryngeal biomarkers with Psychomotor Retardation assessment motivate an important direction in the improved understanding of the neuro-physiologic basis for changes in voice quality with depression, as well with other central nervous system disorders that result in speech degradation.

References

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

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, Text Revision, Washington, DC, American Psychiatric Association, 2000.

[3] D. France, R. Shiavi, S.E. Silverman, M.K. Silverman, D.M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering* 47(7): 829, 2000.

[4] L.A. Low, T. Maddage, M. Lech, L. Sheeber, N. Allen, "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] E. Moore II, M. Clements, J. Peifer, L. Weisser, "Analysis of prosodic variation in speech for clinical depression," *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, 2003.

[6] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, D. Mitchell, "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), 2004.

[7] A.Trevino, T.F. Quatieri, N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing: Special Issue on Emotion and Mental State Recognition from Speech*, August 2011.

[8] M.P. Caligiuri, J. Ellwanger, "Motor and cognitive aspects of motor retardation in depression," *J Affect Disord.*, Jan-Mar, 57(1-3):83-93, 2000.

[9] D.D. Mehta, D.D. Deliyski, S.M. Zeitels, T.F. Quatieri, and R.E. Hillman, "Voice production mechanisms following phonosurgical treatment of early glottic cancer," Ann. Otol. Rhinol. Laryngol. 119(1), 2010.

[10] P. Boersma and D. Weenink, Praat: Doing phonetics by computer (Version 5.1.05). May 1, 2009; http://www.praat.org/

[11] J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 713-726, 2001.

[12] T.F. Quatieri, *Discrete-Time Speech-Signal Processing: Principles and Practice*, Prentice Hall, 2001.

[13] J. Myers, J. and A. Well, *Research design and statistical analysis*, Lawrence Erlbaum, 2003.

[14] C. Sobin, and H. A. Sackeim, "Psychomotor Symptoms of Depression," Am J Psychiatry 154:1, January 1997.

[15] M. Cannizzaro, B. Harel, et al, "Voice acoustical measurement of the severity of major depression." Brain and cognition 56(1): 30-35, 2004.