SPECTRO-TEMPORAL ANALYSIS OF SPEECH AFFECTED BY DEPRESSION AND PSYCHOMOTOR RETARDATION

Nicholas Cummins[1,2], Julien Epps[1,2], Eliathamby Ambikairajah[1,2]

[1]School of Elec. Eng. and Telecomm., The University of New South Wales, Sydney Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia

ABSTRACT

To enhance current diagnostic methods used when assessing a depressed individual, an objective screening mechanism, ideally based on non-intrusive behavioral signals, is needed. Given the clinical description of depression speech as 'dull, monotonous and flat' and promising previous results from spectral features, we hypothesize that the effects of depression on speech are embedded in spectro-temporal events. To test this hypothesis we explore different methodologies, based on the modulation spectrum, for extracting long-term spectro-temporal information from speech and assess their suitability as a clinical marker of depression. Results indicate that: depressive speech information is captured in the modulation spectrum, long-term spectro-temporal information is important in depressed speech identification and there are potential differences in the effects that depression and psychomotor retardation have on speech production mechanisms.

Index Terms **–** Depression, Psychomotor Retardation, Spectro-Temporal, Modulation Spectrum, Energy variability

1. INTRODUCTION

Depression is an affective disorder that has a wide ranging clinical profile, symptoms include; cognitive impairments, feelings of worthlessness, diminished interest and a sustained depressed mood lasting for weeks [1]. Another key symptom of depression is psychomotor retardation (PMR) which is the slowing of thought and reduction of physical movements. Speech as a complex cognitive and muscular action is considered a key objective measure of both depression and PMR [2]. Clinically depressed and PMR-affected speech has been described as sounding dull, monotonous and flat [2].

Clinicians often use rating scales such as the Hamilton Rating Scale for Depression (HAMD) [3], to diagnose depression. These tests require clinical training, practice, and certification to produce acceptable results [4], however they are subjective and often require face-to-face interaction with a psychiatrist. To enhance current diagnostic methods an objective screening mechanism, based on physiological and behavioral signals such as speech and PMR, is needed.

Several papers have found significant correlations between a person's clinical level of depression and prosodic speech features. Whilst inconsistent results have been reported for pitch based measures [4-7], more consistent results have been reported for speech timing measures [4, 5, 8]. Results in these papers indicate an increase in both speech timing and pause duration measures with depression severity. Significant correlations have also been reported between clinical PMR scores and articulation rate, phoneme rate and total vocalization time [9].

Spectral and energy based features have been shown to have strong discriminatory properties when automatically classifying depressed speech [10-12]. The default standard in speech recognition systems, when using spectral based measures, is to incorporate temporal based information through the use of either short-term 1st and 2nd order time derivatives (Δ, $\Delta\Delta$) or the medium-term Shifted Delta Coefficients (SDC). Several papers show that the addition of these features offers little improvement when classifying depressed speech [10, 13, 14]. Time derivatives and SDC are designed to capture rapid temporal information, but depression is a more long term condition whose effects potentially vary across longer time scales than those used when extracting these delta features. SDC's are defined by four parameters; N-d-P-k, where N denotes the number of coefficients and d the number of frames the SDC's are calculated over, whilst P denotes the frame shift between blocks and k the number of coefficients used to form final SDC representation. Using the default SDC setting of 7-1-3-7 with a frame shift of 10ms, 190ms of temporal information is incorporated into the overall SDC feature vector, but the individual contributing delta coefficients are computed over just 30ms. Using such a short time window makes it impossible to differentiate between slow and fast rates of spectral change [15].

One method proposed to capture long-term information in a speech signal is the *Modulation Spectrogram*. The modulation spectrogram comprises the frequency components of sub-band frequencies of a spectrogram representation of speech, and is extracted using temporal frames up to 300ms in length. The modulation spectrogram offers an approach for characterizing both slow and fast rates of spectral change, capturing information relating to speech intelligibility by quantifying the power of temporal events relating to articulatory movements in the speech signal [16].

Motivated by recent results published in [7], where significant correlations ($p<0.05$) between an increase in variability associated with energy dynamics and increasing levels of either depression or PMR were reported, this paper explores the benefits of incorporating longer term spectro-temporal information in the identification of both PMR and depressive speech. This is achieved by testing SDC's and two different methodologies based on the modulation spectrum for extracting long-term spectro-temporal information from speech.

2. DEPRESSION DATABASE

The database used in this paper contains voice samples from 35 patients undergoing depression treatment over a 6 week period, originally collected for a depression severity study by Mundt et al. [4]. At weeks 0, 2, 4, and 6 of the study, the participants undertook clinical sessions in which their depression severity was measured using the HAMD assessment. The HAMD assessment rates the severity of symptoms observed in depression, to give a patient a score which relates to their level of depression (HAMDtotal). The scores are arranged into 5 categories; 'Normal' (0-7), 'Mild' (8-13), 'Moderate' (14-18), 'Severe' (19-22) and 'Very Severe' ($\geq$23).

Due to differences in how individuals responded to treatment over the course of the trial no individual speaker has data in all five classes, but all speakers have data in two or more classes. One of the subtopics in the HAMD is PMR; individuals are given a ranking between 0-4 (HAMDpmr), based on their level of observable PMR.

As part of these clinical sessions, repetitions of the tri-syllabic sequence 'PATAKA', used to test diadochokinetic rate [17], were recorded as well as samples of four held vowels sounds; /a/, /i/, /o/ and /u/ (sampling rate 8kHz). To minimise phonetic content whilst still allowing analysis of articulatory movements and speech intelligibility all results reported and figures plotted unless otherwise stated are taken from the 'PATAKA' sound.

3. SPECTRO-TEMPORAL CHARACTERIZATION OF DEPRESSIVE SPEECH

Given that depression and PMR potentially cause prosodic, articulatory and phonetic errors in speech [4, 8, 9] as well as altering spectral properties [7, 10], to objectively assess depression and PMR affected speech, a feature or group of features that capture these properties is needed. This information is present in the spectrogram (Figure 1), but this is a high dimensional representation of a speech signal with substantial redundancy, unsuitable for statistical learning algorithms.

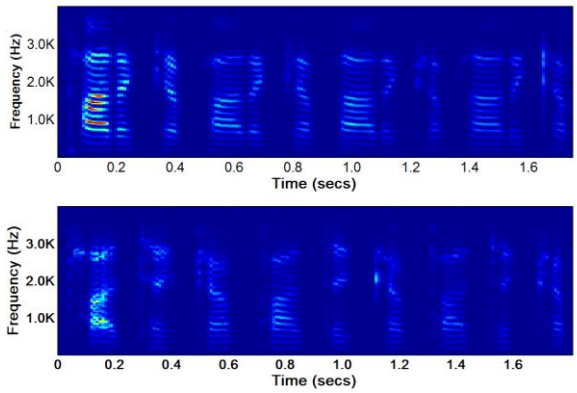

Figure 1: Spectrograms of four repetitions of the 'PATAKA' sound for an individual in a normal state (top) and very severe depressed state (bottom). Low energy regions are darker, high energy regions are lighter. Notice the lack of sustained energy in the 1 kHz and 3 kHz regions in the very severe example.

3.1 Modulation Spectrum

The Modulation Spectrum (MS) is defined as the frequency composition of the temporal-trajectory of each acoustic frequency channel in a spectrogram [18]. The MS provides information on the dynamic characteristics of a speech signal extracted over longer time scales (typically 200-300ms) [16]. The MS is given by calculating the Fourier Transform of each frequency bin along the time axis of the spectrogram;

$$X(\theta,\omega)=\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}x[n,m]e^{-j(\theta n+\omega m)} \qquad (1)$$

where $x[n,m]$ is a short term speech segment, n is the frame index and m is the time index, ω is the acoustic frequency and θ is the modulation frequency. By transforming the spectrogram into the MS, using longer-term temporal windows, 200-500ms in length, we can reveal the slowly varying temporal envelope components at different acoustic frequencies in the speech signal [16].

In Figure 2 the two dimensional representation of the MS is shown; the y-axis represents the acoustic frequencies (AF) present in the signal whilst the x-axis is the modulation frequency (MF), so that the intensity of each pixel represents the amount of temporal change present in each acoustic frequency. The MS is low-pass in shape, with most of the modulation energy located below 25Hz: these low modulation frequencies relate to speech rhythm and are the relevant range of modulation frequencies for speech intelligibility [16].

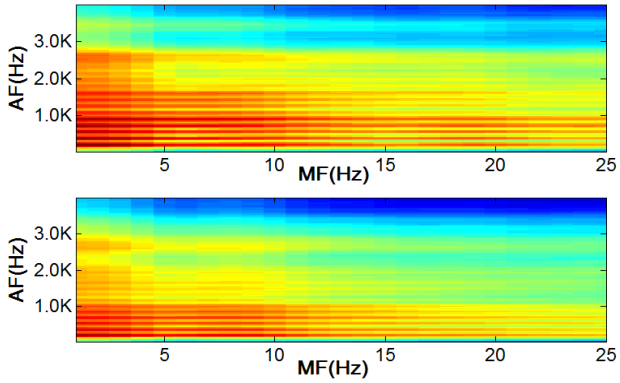

Figure 2: Modulation Spectra extracted over 250ms, for the same normal state (top) and very severe state (bottom) utterances as Fig. 1. Low energy regions are blue, high energy regions are red. Notice the difference in magnitudes between the two figures across all MF's between approximately 1 kHz to 3 kHz AF.

3.2. Modulation Spectrum and Log Mean Subtraction

In [19], log mean subtraction (LMS) was proposed to help minimize the effects of source excitation in the MS. By first computing the log of the spectrogram, it is possible to emphasize the dynamic properties of the speech production apparatus, then perform mean subtraction along each acoustic frequency of the log-spectrogram, minimizing the effect of the excitation source component of the speech signal [19]. In the MS-LMS representation, most of the energy is located below 12 Hz (Figure 3); this modulation frequency range relates to the articulator movement rate [20].

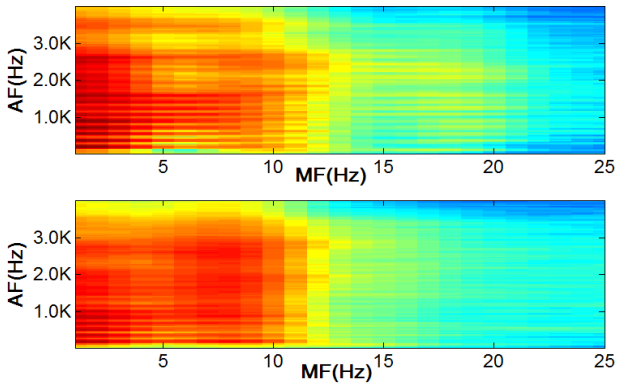

Figure 3: Modulation Spectra extracted over 250ms, after LMS for the same normal state (top) and very severe state (bottom) utterances as Fig. 1. Low energy regions are blue, high energy regions are red. Note that the difference in magnitudes between the two figures, in the 1 kHz (AF) / 5 Hz (MF) region, has been retained and the steeper low-pass shape of the transformed MS when compared with Figure 2.

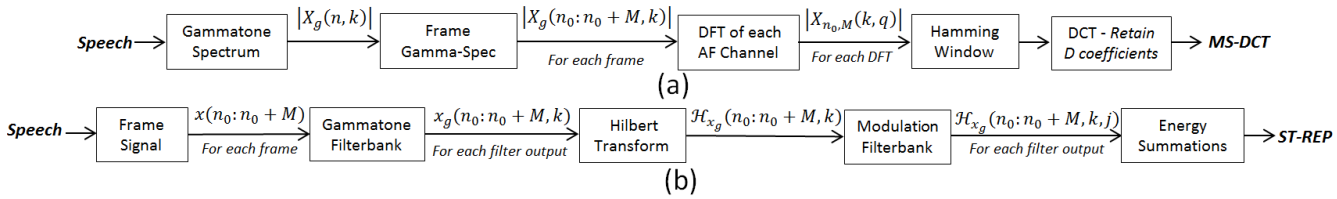

Figure 4: Outline of extraction methods used to reduce the dimensionality of the Modulation Spectrum: (a) Steps in extracting MS-DCT [16] (b) Steps in extracting ST-REP [20]

3.3. Dimensionality Reduction of Modulation Spectrum

The MS transform results in a three dimensional feature space; acoustic frequency, modulation frequency and time. In Figures 2 and 3 the modulation spectra were extracted using an acoustic frequency FFT size of 256 and a modulation frequency FFT size of 128, resulting in 8192 features per frame. We now consider possible techniques to reduce the dimensionality of the extract Modulation Spectrum.

The first technique, MS-DCT, was used in [16] to extract MS for text independent speaker recognition (Figure 4(a)). It uses a filterbank to reduce the dimensionality of the acoustic frequency dimension; for this paper we used an auditory-motivated Gammatone filterbank. The dimensionality of the modulation frequency bank is reduced by applying a DCT to each of the temporal trajectories and retaining the first *D* coefficients.

In the ST-REP technique, used in [21] for emotion recognition (Figure 4(b)), all filtering is done in the time domain to allow higher modulation frequencies to be extracted than possible via the MS-DCT. A Gammatone filterbank and Hilbert Transform are applied to extract a temporal envelope for each frequency channel in a long-term window. Each temporal envelope is then filtered with a modulation filterbank and an energy coefficient of each acoustic / modulation frequency pair is calculated.

4. EXPERIMENTAL SETTINGS

The experimental settings (unless otherwise stated) were as follows: short-term frames were 25ms in length and extracted every 10ms, long-term frames were 250ms in length and extracted every 10ms. Energy coefficients (STFT magnitudes) were extracted using a 24-channel Gammatone filterbank; linearly spaced at Equivalent Rectangular Bandwidth between 100 Hz - 4 kHz, operating on short-term frames and the SDC's were extracted with parameters *N-d-p-k* equal to 24-1-3-7.

In the MS-DCT representation the coefficients were extracted using the long-term frames, all acoustic frequencies were extracted using a 256 point FFT, all modulation frequencies were extracted using a 128 point FFT and 10 DCT coefficients retained. For the ST-REP the coefficients were extracted using the long-term frames, and the modulation filterbank was constructed from five second-order band-pass filters each with a quality factor of two. The centre frequency spacing was evenly spread on a log2 scale from 4 to 64Hz.

Features were tested for their suitability as an objective measure of clinical depression severity in a SVM classifier, both in a 2-class system (HAMDtotal < 17, HAMDtotal ≥ 17) and a 5-class system, as explained in Section 2. As further information reduction, the following statistical functionals were extracted per coefficient, per utterance and used as a single, per-utterance input to the SVM: variance, standard deviation, skewness, kurtosis, mean, max, min, median, 1st quartile, 3rd quartile, median-1st quartile, 3rd quartile – median, 3rd-1st quartile, 1st percentile, 99th percentile and 99th-1st percentile. To avoid variability due to accents, the evaluation dataset was composed of the PAKATA recordings taken from patients of Caucasian ethnicity; this resulted in 132 recordings spread over 32 speakers. The average time of each recording is 7.8s. All classification accuracies were obtained using WEKA's inbuilt SVM with a RGF kernel [22] with ten-fold cross validation used to verify all results.

5. RESULTS

5.1 Classifying Clinical Level of Depression

The results in Table 1 show the benefit of including spectro-temporal information when using speech to classify clinical depression. In the 2-class system there is a further benefit gained when applying LMS, with the best feature being the MS-DCT with LMS. In the 5-class system LMS does not appear to aid classification as consistently. The best 5-class classification accuracy is given again by the MS-DCT feature, this time without LMS. Interestingly given the results in [10, 13, 14], the SDCs give reasonable classification accuracy in both systems, indicating the importance of including longer-term temporal information. Due to the small range and uneven distribution of the HAMDpmr scores we did not test the features in a PMR-based classifier.

Table 1: Classification accuracies for identifying HAMDtotal

Feature	Average Weighted Accuracy (%)			
	2-Class	2-Class (LMS)	5-Class	5-Class (LMS)
Energy	55.1	55.9	28.8	27.1
SDC	**61.9**	63.6	30.5	**33.9**
MS-DCT	55.1	**66.9**	**36.4**	31.4
ST-REP	57.6	N.A	33.1	N.A

5.2 Effect of Applying LMS

To test the effect of LMS on the features used in section 5.1, we investigated the correlations between the variance of the extracted features with HAMDtotal and HAMDpmr scores. We use feature variance, calculated over an entire utterance, to represent feature variability in order to compare with results in [7, 11, 12]. Correlations in this section are reported in terms of no significance ($p \geq 0.05$), mild significance ($p < 0.05$) and strong significance ($p < 0.01$). Due to the quantized nature of the clinical scores, Spearman's correlation coefficient was used [7, 9].

Figure 5(a) shows the correlations between the variance of the 24 short-term energy coefficients and HAMDtotal. The strongly significant negative correlations seen in channels 13-14 (Approx. 900 Hz – 1.1 kHz) indicate a decrease in the variability of these coefficients with increasing depression severity. This roughly agrees with what we see in Figure 1. For the spectro-temporal features, no significant correlations were found for SDC whilst for both the MS-DCT and ST-REP features a mix of mild and strongly significant negative correlations were found between the variability of the extracted feature coefficients in channels 10-14 (Approx. 600 Hz – 1.1 kHz). No strong significant correlations were found between HAMDpmr and any of the extracted features (Figure 5b).

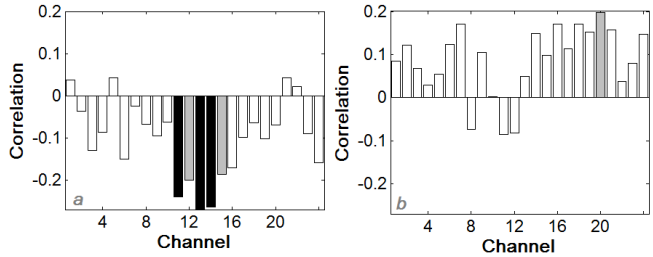

Figure 5: Channel dependent correlations, calculated using all PATAKA samples, of energy variance, extracted using a 24 channel Gammatone filterbank, with HAMDtotal scores (a) and HAMDpmr scores (b). Gray indicates a mild significance (p<0.05) and black indicates a strong significance (p<0.01)

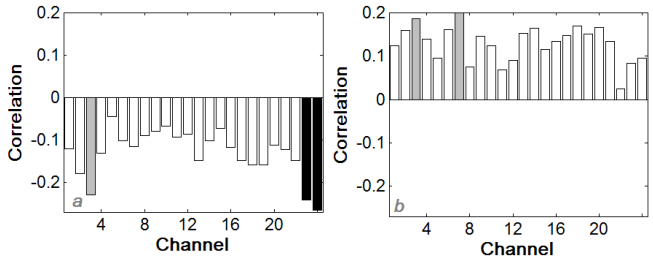

Figure 6: Channel dependent correlations, calculated using all available PATAKA samples, of energy variance, extracted with a 24 channel Gammatone filterbank and *applying LMS*, with HAMDtotal scores (a) and HAMDpmr scores (b). Gray indicates a mild significant (p<0.05) and black indicates a strong significance (p<0.01)

The effect of applying LMS can be seen in Figure 6, where results suggest that there are differences in the effects that depression and PMR have on speech production mechanisms, and that these are consistent across all frequency bands. The negative correlations with HAMDtotal (Figure 6(a)) show a decrease in the variability of energy associated with the dynamical properties of speech production mechanisms as depression severity increases. The opposite effect is seen for PMR (Figure 6(b)), where the positive correlations infer an increase in this variability with increasing PMR. No strongly significant correlations were found between either HAMDtotal or HAMDpmr and SDC's extracted after LMS. For the MS-DCT LMS coefficients a mix of mild and strongly significant negative correlations with HAMDtotal were observed in all modulation frequencies across channels 10-24 (Approx. 600 Hz – 4 kHz).

For HAMDpmr mildly significant positive correlations were observed across all MS-DCT coefficients in channels 1-8 (Approx. 100 Hz –500 Hz). Similar results were also seen for MS-DCT LMS coefficients across the held vowel sounds with the strongest effects being on /a/ and /u/.

The negative correlations seen in channels 23-24 (Approx. 2.3 kHz – 4 kHz) in Figure 6(a) go against the trend of an increase in high frequency energy with increasing depression severity [11], [12]. Further work was done to explore this phenomenon. When dividing the data samples into the 5 classes (Section 2), the general trend seen across the significant energy coefficients was variability increasing in the moderate and mild classes when compared with the normal class. But in the severe and very severe groupings, the energy variability was less than the normal grouping. This result could in part explain the mixed results seen for the 5-class LMS features in Section 5.1.

6. CONCLUSIONS

In this paper we investigated modulation spectrum based speech features as a way of objectively measuring clinical depression severity. MS-based features were found to capture depression based information as well as or better than SDCs. These results show that medium to long term spectro-temporal information has discriminatory properties when identifying an individual's level of clinical depression from their speech. Further research will be done to explore the performance of MS features on a phonetically varied dataset, although it should be noted that it is realistic to use the tri-syllabic 'PAKATA' sequence in a clinical depression test. The sequence is used to test for motor speech disorders in conditions such as ataxic dysarthria and Parkinson's disease [17].

A key result is the effect of applying LMS; the mostly negative correlations, seen across all sounds tested, between HAMDtotal and energy variability, indicate that as depression severity increases there is a decrease in the energy variability associated with speech production mechanism. This is consistent with results seen in [6, 7]. The positive correlations seen across all sounds tested, between HAMDpmr and energy variability, indicate that as PMR severity increases there is an increase in the energy variability associated with speech production mechanism. We could hypothesize that this indicates that more effort might be required to produce and sustain PMR-affected speech. This could be due to a lack of motor coordination which improves as the effects of PMR decreases [7]. Further work will be done to explore these results and this hypothesis in more detail.

7. ACKNOWLEDGEMENTS

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. The authors would like to thank Dr James Mundt for the use of his database. The collection of these data was made possible by a Small Business Innovation Research grant (R43MH068950: JC. Mundt, PI) supported by the United States National Institute of Mental Health.

8. RELATION TO PRIOR WORK

Work presented here has explored the potential of modulation spectrum based features for extracting longer-term spectro-temporal information from a speech signal for use as an objective measure of depression or psychomotor retardation. Motivated by the recent results published in [7], where significant correlations (p<0.05) between energy dynamics and increasing levels of either depression or PMR were reported, this paper showed that there is benefit in including medium / longer term spectro-temporal information when using speech to classify clinical depression, with both the modulation spectrum based features and shifted delta coefficients being shown to have discriminatory characteristics. Previous results in the literature have shown that including either short-term spectro-temporal information in the form of RDC's and medium term spectro-temporal information in the form of SDC's did not aid classification accuracy [10, 13, 14]. Further results, previously not seen in depression classification literature, show that by using log mean subtraction, as a means to isolate speech production dynamics [19], depression and PMR have potentially conflicting effects on the speech production mechanism.

9. REFERENCES

[1] A. T. Beck and B. A. Alford, *Depression: Causes and Treatment*: University of Pennsylvania Press, 2008.

[2] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depression," *Am J Psychiatry,* vol. 154, pp. 4-17, 1997.

[3] H. Hamilton, "HAMD: A rating scale for depression," *Neurosurg Psychiat,* vol. 23, pp. 56-62, 1960.

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

[5] H. H. Stassen, S. Kuny, and D. Hell, "The speech analysis approach to determining onset of improvement under antidepressants," *European Neuropsychopharmacology,* vol. 8, pp. 303-310, 1998.

[6] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain and Cognition,* vol. 56, pp. 30-35, 2004.

[7] T. F. Quatieri and N. Malyska, "Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity," in *INTERSPEECH-2012*, Portland, USA, 2012, p. NA.

[8] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biological Psychiatry,* p. unavailable, 2012.

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

[10] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in *INTERSPEECH-2011*, Florence, Italy, 2011, pp. 2997-3000.

[11] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *Biomedical Engineering, IEEE Transactions on,* vol. 47, pp. 829-837, 2000.

[12] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *Biomedical Engineering, IEEE Transactions on,* vol. 51, pp. 1530-1540, 2004.

[13] L. S. A. Low, M. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of Clinical Depression in Adolescents; Speech During Family Interactions," *Biomedical Engineering, IEEE Transactions on,* vol. 58, pp. 574-586, 2011.

[14] D. Sturim, P. A. Torres-Carrasquillo, T. F. Quatieri, N. Malyska, and A. McCree, "Automatic Detection of Depression in Speech Using Gaussian Mixture Modeling with Factor Analysis," in *INTERSPEECH-2011*, Florence, Italy, 2011, pp. 2983-2986.

[15] H. Hermansky, "Should recognizers have ears?," *Speech Communication,* vol. 25, pp. 3-27, 1998.

[16] T. Kinnunen, K. A. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," in *The Speaker and Language Recognition Workshop (Odyssey 2008)*, Stellenbosch, South Africa, 2008.

[17] W. Ziegler, "Task-Related Factors in Oral Motor Control: Speech and Oral Diadochokinesis in Dysarthria and Apraxia of Speech," *Brain and Language,* vol. 80, pp. 556-575, 2002.

[18] T. F. Quatieri, *Discrete-Time Speech-Signal Processing: Principles and Practice*. Upper Saddle River, NJ 07458 Prentice Hall, 2001.

[19] A. V. Ivanov and X. Chen, "Modulation Spectrum Analysis for Speaker Personality Trait Recognition," in *INTERSPEECH-2012*, Portland, U.S.A, 2012, p. NA.

[20] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing,* vol. 7, pp. 668-675, 2003.

[21] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication,* vol. 53, pp. 768-785, 2011.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.,* vol. 11, pp. 10-18, 2009.