Modeling Spectral Variability for the Classification of Depressed Speech

Nicholas Cummins[1, 2], Julien Epps[1, 2], Vidhyasaharan Sethu[1], Michael Breakspear [3, 4] and Roland Goecke[5]

[1]School of Elec. Eng. and Telecomm., The University of New South Wales, Sydney Australia

[2]ATP Research Laboratory, National ICT Australia (NICTA), Australia

[3]Black Dog Institute and School of Psychiatry, The University of New South Wales, Australia

[4]Division of Mental Health Research, Queensland Institute of Medical Research, Australia

[5]Faculty of Education, Science, Technology and Mathematics, University of Canberra, Australia

n.p.cummins@unsw.edu.au, j.epps@unsw.edu.au, vidhyasaharan@gmail.com

mjbreaks@gmail.com, roland.goecke@ieee.org

10.21437/Interspeech.2013-242

Abstract

Quantifying how the spectral content of speech relates to changes in mental state may be crucial in building an objective speech-based depression classification system with clinical utility. This paper investigates the hypothesis that important depression based information can be captured within the covariance structure of a Gaussian Mixture Model (GMM) of recorded speech. Significant negative correlations found between a speaker's average weighted variance - a GMM-based indicator of speaker variability - and their level of depression support this hypothesis. Further evidence is provided by the comparison of classification accuracies from seven different GMM-UBM systems, each formed by varying different parameter combinations during MAP adaption. This analysis shows that variance-only adaptation either outperforms or matches the de facto standard mean-only adaptation when classifying both the presence and severity of depression. This result is perhaps the first of its kind seen in GMM-UBM speech classification.

Index Terms: Depression, spectral variability, MFCC, GMM-UBM, MAP adaptation, average weighted variance

1. Introduction

In recent years, research into the automatic and objective classification of mood disorders, such as depression, using behavioural signals has gained popularity. Clinical depression has a wide range of potential symptoms including low mood, observable psychomotor retardation and cognitive impairments leaving sufferers prone to feelings of worthlessness and strong negative conceptualizations. Clinical diagnosis indeed rests upon these features during a clinical assessment, with speech playing an explicit role in this regards. However, the broad clinical profile means there is no single clinical characterization of a depressed individual making diagnosis subjective in nature and time consuming.

Current state-of-the-art diagnostic methods include interview style assessments such as the Hamilton Rating Scale for Depression (HAMD) [1] or self-assessments such as the Quick Inventory of Depressive Symptomatology (QIDS) [2]. However, the accuracy of these tests depends upon the frankness of the patient's responses and their ability and desire to communicate their symptoms [3], occurring at a time when, by definition, their outlook and motivation are impaired. It is believed that finding a set of objective behavioural indicators will help improve the diagnosis accuracy of depression [3].

Speech has the potential to be used as part of a behavioural classification system; speech in patients with depression is often described as being monotonous and "lifeless", with a diminished prosody [4]. As the outcome of a complex cognitive and motoric muscular act, speech can be affected by both physiological symptoms and changes in cognitive ability relating to the effects of depression. Paralinguistic analysis of depressed speech has shown that physiological symptoms relating to depression affects vocal tract properties [5-8], whilst changes to cognitive ability can affect measures relating to speech rate [9-11].

Spectral and energy based features have consistently been observed to change with a speaker's mental state, although there is some disagreement as to the nature of the effect. Some studies have reported a relative shift in energy from lower to higher frequency bands with increasing depression severity due to increasing vocal tract tension [5, 6]. Energy variability has also been shown to decrease with increasing levels of depression, due in part to a decrease in the motor action associated with speech production [7, 8].

Changes in speech rate and motor actions associated with depression have been shown to be encoded at the phoneme level of speech production. Work done in [12] shows speaker recognition rates for depressed speakers were 25% lower than for non-depressed speakers when using a Mel Frequency Cepstral Coefficient (MFCC) / Gaussian Mixture Model (GMM) speaker recognition system. These results suggest fundamental changes in the phoneme spectral structure of depressed speech. This result has been verified in part by results published in [13]. By grouping phonemes together by manner of articulation, the authors found consistencies in the correlations within the articulatory groups with the level of depression associated with each phone and phoneme specific energy and timing measures [13]. Amongst a wide range of acoustic features and spectral features tested in [14], - MFCC's specifically - displayed the strongest discriminatory characteristics when classifying the presence/absence of depression. Given these results, and the stability of MFCC/GMM in partitioning the acoustic space [15], it is not surprising that this classifier set-up has performed well when classifying between either low/high levels of depression [16], or the presence/absence of depression [17, 18].

The assumption when using GMMs to model the acoustic space is that a single mixture component describes a broad acoustic class in terms of its average spectral shape, the mean of the Gaussian, and an estimate of the extent of variations from this mean in the associated covariance matrix [19]. Motivated by recent results published in [7, 8], we presently explore the hypothesis that important depression based information is captured within the covariance matrices of a GMM. This is achieved in two ways; firstly we investigate how the average weighted variance, a simple estimate of the average local variability captured by a GMM [20], is affected by changing levels of depression. Secondly, we explore the effects on classification accuracy updating class specific GMM's from universal background model (UBM) whilst varying which parameters are updated.

2. Depression Corpora

Two depression based databases are used in this paper. The first is the Mundt database originally collected for a depression severity study by Mundt et al. [10]; results using this are also presented in [7, 8, 13, 16]. Participants in the study undertook fortnightly clinical sessions in which their depression severity was measured using both the HAMD and QIDS assessment. Both assessment methodologies rate the severity of symptoms observed in depression, to give a patient a score which relates to their level of depression (Htotal or Qtotal), and have predictive validity when differentiating depressed from non-depressed patients [21]. The major differences between the two assessments are that HAMD is a clinician-rated test, whilst the QIDS is a self-reported measure. Both use different weighting schemes to producing their total score [21].

The scores for both assessments can be split into either two or five classes. For Htotal in the two-class system the categories are 'low' (1-17) and 'high' (17-52, noting a score of 23 and over indicates very severe depression), whilst for Qtotal in the two class system the categories are 'low' (1-13) and 'high' (13-27). As part of these clinical sessions, the Grandfather passage, a standard reading passage used to test speech fluency [22], was recorded as well as responses to three questions on the patient's emotional and physical state.

The second corpus, referred to herein as the Black Dog database, was obtained from audio-video data collected for an ongoing study conducted by the Black Dog Institute, a specialist mood disorder research facility, into measuring the facial activity in depressed patients [23]. The speech database collected so far from the study contains 30 depressed and 30 controls, one recording per subject, with an even gender split, and to avoid variability due to accent all participants speak 'Australian' English. This is the same database used in [18] and a larger set than that used in [14]. All depressed subjects met the Diagnostic and Statistical Manual of Mental Disorders, 4th edition, (DSM-IV) criteria for either moderate or severe depression and subjects were excluded from the control group if they had any personal or family history of mental illness.

As part of the experimental setup, participants were asked to read sentences containing affective content and participated in an interview in which they were asked to describe events that had aroused significant emotions. All subjects gave informed consent and the study was conducted in accordance with local institutional ethics committee approval.

3. Experimental Approach

3.1. GMM-UBM Depression Classification System

The default standard classification system in many speech related classification tasks uses the GMM-UBM paradigm. In this system, class-specific GMMs are adapted, via MAP-adaption [24], from a background model representing the wide distribution of feature vectors in the model domain. A GMM is a convex combination of Gaussian probability density functions:

$$p(\bar{x}) = \sum_{i=1}^{M} \omega_i p_i(\bar{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \quad (1)$$

where $\boldsymbol{\mu_i}$ denotes the mean vector, representing average spectral shape, $\boldsymbol{\Sigma_i}$ denotes the covariance matrix, representing variations in spectral shape, and ω_i the prior probability or mixing weight for the *i*-th Gaussian distribution [19].

In many automatic speaker recognition tasks, it is standard practice to adapt only the means of the Gaussians [24], whilst in many paralinguistic tasks both full (mean, variance and weight) adaptation and mean adaptation are commonly used approaches [25-27]. An advantage of using mean-only adaptation is that the adapted means can be used to form supervectors for use in a Support-Vector-Machine (SVM) classifier. But using mean supervectors clearly only makes sense if all information relevant is contained within the means of the Gaussians. For some applications, such as language recognition [28], this is not the case.

If our hypothesis on the importance of the covariance structure is correct, it may be advantageous for depression classification to form a supervector from the GMM covariance structure. However, before we attempt this task, it is important to understand how depression based information is represented in the GMM domain and specifically, given the importance of spectral variability in analyzing depressed speech [5-8], how variance in the feature domain translates to variance captured within the model domain.

We test our hypothesis by comparing the classification accuracies of a range of two-class low/high level of depression and two-class presence/absence of depression classification systems. By using a GMM-UBM system we can perform mean-only, variance-only and weight-only adaptation where each class-specific GMM pair has a common co-ordinate system, i.e. when doing variance only updates the class specific GMMs means and mixing weights remain unchanged from the UBM's. This analysis will hence allow us to assess the relative importance of each parameter in determining a speaker's level of depression.

3.2. Measuring Covariance

If spectral feature domain variance captured in the model domain is due to a speaker's level of depression; the captured variance should vary significantly with a speaker's level of depression. Therefore to test this hypothesis, we adapted three different GMMs via full, variance-only or variance and weight MAP adaptation, for each speaker/session in the Mundt corpus (118 GMMs per test, see Section 3.3) and performed correlations between the average weighted variance (AWV) and the speaker's Htotal and Qtotal score for that utterance. This analysis was not possible on the Black Dog corpus as the number of individually scored utterances is too low to reliably calculate a correlation coefficient.

AWV is a simple indicator that represents speaker variability captured in a GMM [20]; it is computed using the weights and diagonal covariance matrices of a given GMM:

$$AVW = \frac{1}{K} \sum_{i=1}^{M} \sum_{j=1}^{K} \omega_i \sigma_{i,j}^2 \quad (2)$$

where *M* is the number of mixture, *K* the dimensionality of the feature vector, and $\sigma_{i,j}^2$ is the *j*-th diagonal covariance component taken from the *i*-th mixture [20].

3.3. Experimental Settings

The experimental settings (unless otherwise stated) of the classification system were as follows: a 39-dimensional feature vector was formed from thirteen MFCCs, including C_0, extracted every 10ms using a 25ms window. The 13 coefficients were concatenated with delta (Δ) and delta-delta (ΔΔ) coefficients extracted using the conventional regression equation. The openSMILE toolkit [29] was used to extract all

features. Only voiced frames were used in the modeling, determined using openSMILE's voicing probability function.

To test our hypothesis, we compared classification accuracies of 7 different GMM-UBM classifiers allowing all possible combinations of Gaussian parameters to be updated during MAP adaptation; mean, variance and weight (mvw), mean only (m-only), variance-only (v-only), weight only (w-only), mean and variance (mv), mean and weight (mw) and finally variance and weight (vw).

All tests on the Mundt corpus were performed using two different UBMs; the first formed using the free-response answers (D-ubm), 5hrs, 20min of training data, and the second formed using the 2004 NIST Speaker Recognition Evaluation dataset (N-ubm), 430hrs of training data. Given the relatively small amount of training data available in the Mundt corpus, the 2004 NIST dataset was included to test its suitability as a background model representing a large balanced set of "American English" speech samples; this approach has previously shown to be a suitable in a paralinguistic setting [30]. For the Black Dog corpus, the UBM was formed from the subject's interview data, 5hrs of training data. HTK was used to train the UBMs using 10 EM iterations.

For the Mundt Corpus, the evaluation data was formed from the recorded Grandfather passages. To avoid variability due to accents, we took only the recordings from those patients who are Caucasian, noting also that not every speaker participated in every clinical session; this resulted in 118 recordings, with an average length of 50s, spread over 32 speakers. The evaluation data for the Black Dog were formed using every subject's set of read sentences resulting in 60 recordings with an average length of 40s.

For both datasets, experiments were conducted using 8, 16, 32 or 64 mixtures and 5, 10 or 20 MAP iterations (I) using HTK. To calculate the overall accuracy results reported, the average log-likelihood ratio of each test utterance was calculated with respect to either class model, followed by a maximum-likelihood-ratio decision. Leave-one-out cross validation, performed across the 118 test utterances, was used in all classification tests, with average accuracy reported.

4. Results

4.1. Feature Level

The energy coefficients used in [8] were extracted using a 24-channel Gabor filterbank decomposition of the speech signal; therefore we first checked that we were able to achieve a similar result using MFCC and deltas for the Mundt corpora. Figure 1(a) shows the correlations of the *mean* of each feature dimension, calculated at the utterance level, with Htotal (left) and Qtotal (right). Whilst there are some significant correlations with both scores, there is no consistency in the size and magnitudes of the correlation coefficients.

The correlations between the *variance* of each feature dimension and both scores, in Figure 1(b), show a consistent trend of negative correlations, especially in the delta coefficients where several surpass threshold for statistical significance levels. This result is consistent with those presented in [8] showing that in the feature domain, variations in spectral energy or in the case of the delta coefficients changes in spectral energy *decrease* with increasing level depression.

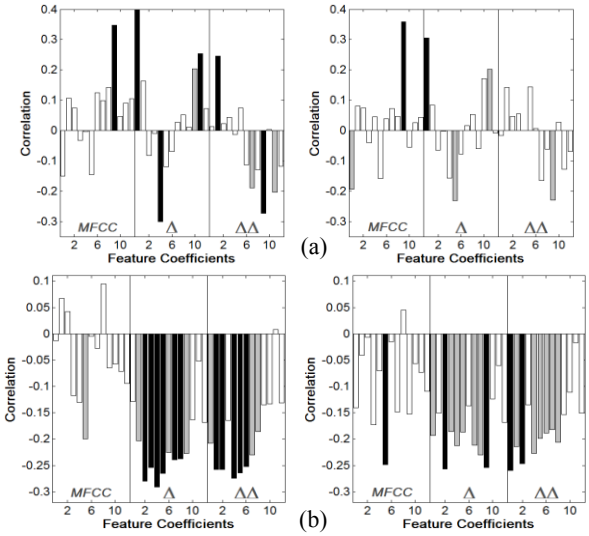

Figure 1: *Correlations per feature dimension calculated using the Grandfather passage, of* ***mean*** *(a) and* ***variance*** *(b) of MFCCs appended with delta coefficients with Htotal (left) and Qtotal scores (right). Gray indicates mild significance (p<0.05) and black indicates strong significance (p<0.01).*

4.2. Average Weighted Variance Results

Significant correlations ($p \leq 0.001$) were found for vw (variance and weight) adaption using the N-ubm with an 8 mixture GMM and 5 MAP iterations for both Htotal and Qtotal. All correlations found were negative, indicating a *decrease* in AWV with increasing levels of depression (Figure 2). Notably, the effects of depression on these correlations match those seen in Figure 1(b), providing support for our hypothesis that important depression information is captured in GMM covariance matrices.

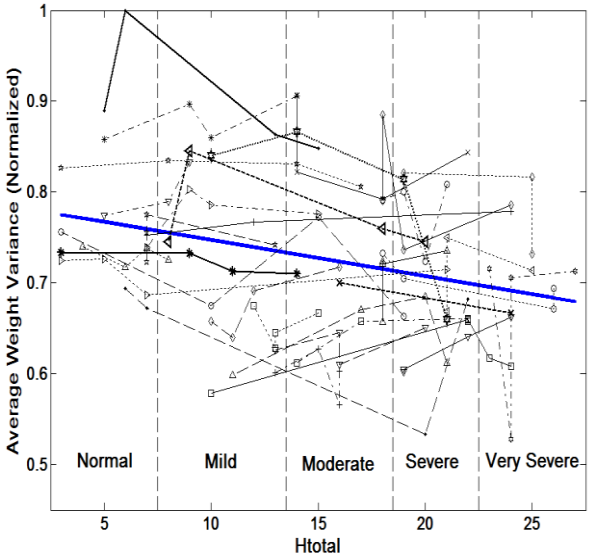

Figure 2: *AWV plots for all patients/sessions in the Mundt database. AWV was extracted using an 8-mixture N-ubm with vw adaption. The thick blue line is the linear regression fit.*

4.3. Classification Results

For the Mundt database, we ran an exhaustive series of preliminary results varying the parameters updated, number of mixtures and MAP iterations. A selection of these results is provided in Table 1. When grouping together the classification

accuracies in terms of UBM, number of mixture components, number of MAP iterations and depression score, a general trend emerged in the classification accuracies; v-only adaptation matched (within 5%) and outperformed m-only adaptation. For some groupings, such as 8-mixtures with N-ubm (Table 1), v-only adaptation gave the highest classification accuracy.

For any given grouping, the highest classification accuracy was never obtained with mean-only adaptation; this was generally achieved by either mvw or vw adaptation. Interestingly whilst w-only adaptation achieved chance-level classification for lower numbers of mixtures, i.e. 8 (Table 1) or 16, for 32 and 64 mixtures classification accuracy in the range 50% - 67% was achieved, often matching the accuracies achieved with m-only and v-only adaptation.

Table 1: *Comparison of classification accuracies when updating different MAP parameters for two-class low/high depressed speech classification accuracy, evaluated using an 8-mixture GMM on the Mundt 32-speaker database.*

I	P	N-ubm		D-ubm	
		Htotal	**Qtotal**	**Htotal**	**Qtotal**
5	mvw	58.2	62.7	65.6	67.9
	m-only	56.2	60.0	61.6	60.3
	v-only	*65.7*	*65.9*	60.4	65.2
	w-only	46.8	51.1	53.8	55.7
	mv	*58.2*	*62.7*	64.8	67.1
	mw	*56.2*	*60.0*	62.3	58.6
	vw	*65.7*	*65.9*	*59.7*	*64.3*
10	mvw	*59.6*	53.1	*64.0*	*67.8*
	m-only	50.3	52.5	61.6	60.3
	v-only	*65.7*	*65.9*	59.7	66.0
	w-only	55.8	51.0	53.8	55.7
	mv	60.5	*52.1*	*64.0*	*67.8*
	mw	50.4	*52.5*	63.1	58.6
	vw	*65.7*	*65.9*	*58.7*	*64.3*

The best classification accuracy for Htotal was 68.6%, obtained using vw adaptation with the N-ubm, 8 mixtures and 20 MAP iterations. For Qtotal the best result was 69.04%, obtained using mvw adaptation with the N-ubm, 16 mixtures and 10 MAP iterations. These results match the maximum two-class classification accuracy of 66.9% achieved in [8], although a direct comparison is not straightforward as those results were found using a different feature (Modulation Spectrum) and classifier (SVM with RGF kernel).

The results for the Black Dog data show that either w-only or vw adaptation give the best performance (Table 2). Again, v-only adaptation either matches or outperforms m-only adaptation. There was no advantage in using either 32 or 64 mixtures on these data (results not shown). The best result on the Black Dog data was 63%, found using weight-only adaption for any combination of 8 and 16 mixtures and either 5 or 10 iterations. This is well below the MFCC two-class accuracy of 77% presented in [14] and 71% presented in [18]. Again, this is not a straightforward comparison as these papers used non-adapted GMMs as their classification system.

The performance of v-only adaptation offers support for our hypothesis that important depression information is captured in the covariance matrices of a GMM. The classification accuracies obtained in the w-only adaptation, particularly in the Black Dog database, were unexpected. We speculate that as weights control the mixture occupancy count (the amount of data assigned to a mixture during MAP adaptation) they capture spectral variability on a global level as opposed to the localized variance captured in the covariance matrices.

Table 2: *Comparison of classification accuracies when updating different MAP parameters for two-class presence/absence of depression classifier, evaluated on the Black Dog 60-speaker database.*

	8 mixtures		16 mixtures	
P	**5 Iter.**	**10 Iter.**	**5 Iter.**	**10 Iter.**
mvw	48.3	*51.7*	51.7	50.0
m-only	53.3	50.7	51.7	51.7
v-only	50.0	51.7	51.7	53.3
w-only	***63.3***	***63.3***	***63.3***	***63.3***
mv	48.3	53.3	51.7	55.0
mw	55.0	53.3	50.0	53.3
vw	56.7	55.0	***63.3***	***63.3***

To check whether the mean-only, variance-only and weight-only systems capture different aspects of depressed speech, we fused the result of the m-only, v-only and w-only systems, using score level fusion, and compared this with the results gained from full adaptation. The results, not shown, demonstrated that the fused systems either matched or outperformed full adaption system. Hence we speculate that the different GMM parameters capture different aspects of depressed speech.

5. Conclusion

Variations in spectral and energy based features have been previously linked to a speaker's level of depression [5-8]. The work in this paper shows in the feature space that spectral variability decreases with increasing levels of depression agreeing with the results presented in [7, 8]. The results of the average weighted variance analysis shows that this negative correlation can also been seen in the model domain, supporting our initial hypothesis. Intuitively these results match the 'dull' and 'monotonous' clinical descriptions of speech affected by depression.

The classification results show that variance-only adaptation either outperforms or matches the de facto standard mean-only adaptation. The performance of the weight-only classification system, especially on the Black Dog data, was unexpected, although given performance of the UBM weight posterior probability (UWPP) supervector in other paralinguistic tasks [31, 32] it is perhaps not too surprising.

The analysis undertaken in this paper has shown it may be advantageous in depression classification to form a supervector from the GMM covariance structure. Future work will also be undertaken to explore benefits in weight-only adaptation and gain further insights into the fusion results.

6. Acknowledgements

This research was funded in part by ARC Discovery Project DP130101094. 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 this 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. The authors also thank Sharifa Alghowinem for her work annotating the Black Dog database.

7. References

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

[2] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber, M. E. Thase, J. H. Kocsis, and M. B. Keller, "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,* vol. 54, pp. 573-583, 2003.

[3] M. J. H. Balsters, E. J. Krahmer, M. G. J. Swerts, and A. J. J. M. Vingerhoets, "Verbal and Nonverbal Correlates for Depression: A Review," *Current Psychiatry Reviews,* vol. 8, pp. 227-234, 2012.

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

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

[6] 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," *Bio-Eng, IEEE Transactions on,* vol. 51, pp. 1530-1540, 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] 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*, 2013, p. NA.

[9] M. Alpert, E. R. Pouget, and R. R. Silva, "Reflections of depression in acoustic measures of the patient's speech," *Journal of Affective Disorders,* vol. 66, pp. 59-69, 2001.

[10] 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.

[11] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biological Psychiatry,* vol. 72, pp. 580 –587, 2012.

[12] S. Memon, N. Maddage, M. Lech, and N. Allen, "Effect of Clinical Depression on Automatic Speaker Identification," in *Bioinformatics and Biomedical Engineering, 2009. (ICBBE '09). 3rd International Conference on*, 2009, pp. 1-4.

[13] 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.

[14] 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.

[15] J. M. K. Kua, J. Epps, M. Nosratighods, E. Ambikairajah, and E. Choi, "Using clustering comparison measures for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 5452-5455.

[16] 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," *Interspeech 2011,* pp. 2983 - 2986, 2011.

[17] L. S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 5154-5157.

[18] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech," in *Twenty-Fifth International FLAIRS Conference*, 2012, pp. 141-146.

[19] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on,* vol. 3, pp. 72-83, 1995.

[20] T. Hasan and J. H. Hansen, "A study on universal background model training in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1890-1899, 2011.

[21] D. Maust, M. Cristancho, L. Gray, S. Rushing, C. Tjoa, and M. E. Thase, "Chapter 13 - Psychiatric rating scales," in *Handbook of Clinical Neurology*. vol. Volume 106, F. B. Michael J. Aminoff and F. S. Dick, Eds., ed: Elsevier, 2012, pp. 227-237.

[22] R. T. Wertz, L. L. LaPointe, and J. C. Rosenbek, *Apraxia of speech in adults: The disorder and its management*: Grune & Stratton Orlando, FL, 1984.

[23] G. McIntyre, R. Goecke, M. Hyett, M. Green, and M. Breakspear, "An approach for automatically measuring facial activity in depressed subjects," in *Affective Computing and Intelligent Interaction and Workshops, 2009. (ACII '09). 3rd International Conference on*, 2009, pp. 1-8.

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[25] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication,* vol. 53, pp. 1062-1087, 2011.

[26] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language; State-of-the-art and the challenge," *Computer Speech & Language,* p. NA, 2012.

[27] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states - A review on intoxication, sleepiness and the first challenge," *Computer Speech & Language,* p. NA, 2012.

[28] W. Campbell, "A covariance kernel for SVM language recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4141-4144.

[29] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, Firenze, Italy, 2010, pp. 1459-1462.

[30] N. Cummins, J. Epps, and J. M. K. Kua, "A Comparison of Classification Paradigms for Speaker Likeability Determination," in *INTERSPEECH-2012*, Portland, USA, 2012, p. NA.

[31] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language,* p. NA, 2012.

[32] D. Bone, M. Li, M. P. Black, and S. S. Narayanan, "Intoxicated Speech Detection: A Fusion Framework with Speaker-Normalized Hierarchical Functionals and GMM Supervectors," *Computer Speech & Language,* p. NA, 2012.