Audiovisual Behavior Descriptors for Depression Assessment

Stefan Scherer
USC Institute for Creative Technologies
12015 Waterfront Dr.
Playa Vista, CA
scherer@ict.usc.edu

Giota Stratou
USC Institute for Creative Technologies
12015 Waterfront Dr.
Playa Vista, CA
stratou@ict.usc.edu

Louis-Philippe Morency
USC Institute for Creative Technologies
12015 Waterfront Dr.
Playa Vista, CA
morency@ict.usc.edu

ABSTRACT

We investigate audiovisual indicators, in particular measures of reduced emotional expressivity and psycho-motor retardation, for depression within semi-structured virtual human interviews. Based on a standard self-assessment depression scale we investigate the statistical discriminative strength of the audiovisual features on a depression/no-depression basis. Within subject-independent unimodal and multimodal classification experiments we find that early feature-level fusion yields promising results and confirms the statistical findings. We further correlate the behavior descriptors with the assessed depression severity and find considerable correlation. Lastly, a joint multimodal factor analysis reveals two prominent factors within the data that show both statistical discriminative power as well as strong linear correlation with the depression severity score. These preliminary results based on a standard factor analysis are promising and motivate us to investigate this approach further in the future, while incorporating additional modalities.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*dimensionality reduction, correlation analysis*; I.5.4 [**Computing Methodologies**]: Pattern Recognition—*applications*; J.3 [**Computer Applications**]: Life and Medical Sciences—*health*

General Terms

Human Factors, Experimentation, Algorithms

Keywords

Audiovisual Analysis, Depression, Nonverbal Indicators, Factor Analysis

ICMI'13, December 9–12, 2013, Sydney, Australia
Copyright 2013 ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522886.

1. INTRODUCTION

Recent developments in audiovisual behavior assessment and tracking technologies, enable us to characterize a person's nonverbal behavior in a quantitative and dynamic way. Such quantitative assessments of nonverbal behavior have great potential for manifold applications, including training applications [4] and healthcare support systems [29].

Within a large body of research several nonverbal behaviors were identified to clinical conditions, such as movement disorders (e.g. Parkinson's disease or hyperkinesia) and depression. Within this work we focus our attention on nonverbal audiovisual behavior descriptors of depression. In particular, we investigate audiovisual behaviors related to reduced emotional expressivity [10, 6, 24] and psychomotor retardation [14, 16, 5, 8].

Emotional expressivity, such as the frequency or duration of smiles, was found to be diagnostic of clinical state. For example, depressed patients frequently display flattened or negative affect including less emotional expressivity [24, 6], fewer mouth movements [13, 25], more frowns [13, 24] and fewer gestures [16, 24, 5].

Also acoustic indicators for depression were investigated in [12]. The analysis involved glottal flow features as well as prosodic features for the discrimination of depressed read speech of 15 male and 18 female speakers. The extracted glottal flow features comprised instances such as the minimal point in glottal derivative, maximum glottal opening, start point of glottal opening, and start point of glottal closing. The prosodic features extracted consist of fundamental frequency (f_0), energy, and speaking rate. The authors identified glottal flow features to be chosen by the feature selection algorithm for the majority of the classifiers as well as energy-based features for female speakers. In [8], several spectral and energy based features were investigated for their discriminative capabilities of read speech using Gaussian mixture models, with Mel frequency cepstral coefficients and the first three formants yielding promising results.

Little multimodal studies are found in the literature with [7] being one of the exceptions. In [7], facial action units and variability of fundamental frequency (f_0) as well as latency to respond to questions were investigated. Both approaches, yield promising discriminative power with about 80% accuracy for each modality. Unfortunately, no multimodal classification experiments are presented. Within this work we aim to investigate multimodal fusion for the analysis of depression using two approaches: we concatenate acoustic and visual measures automatically extracted from our interview data to form a joint multimodal feature vector in an early fusion classification experiment. Secondly, we conduct a standard factor analysis on the joint multimodal feature vectors to see if we can find multimodal behavior descriptors that are indicative of depression as measured with the self-assessment scale PHQ-9.

The remainder of this work is organized as follows: Section 1.1 states our research goals. Section 2 introduces our dataset and recording setup. In Section 3, we introduce the audiovisual behavior descriptors investigated within this work. These descriptors are statistically evaluated in Section 4 and the discriminative power of the descriptors is assessed in speaker-independent uni- and multimodal classification experiments. Additionally, Section 4.3 introduces a factor analysis in which we investigate if it is possible to create joint multimodal nonverbal descriptors in an unsupervised way that are indicative of depression. Finally, sections 5 and 6 discuss the findings and conclude the paper.

1.1 Research Goals

The research goals of this work are the following:

1 We seek to investigate if audiovisual nonverbal behavior descriptors indicative of depression are observable within semi-structured virtual human interview recordings. Additionally, we assess their correlation with the assessed depression severity.

2 Within uni- and multimodal speaker-independent classification experiments we investigate the discriminative power of the observed behavior descriptors.

3 Lastly, we seek to identify joint multimodal indicators of depression and its severity as assessed using a commonly used self-assessment depression scale with a novel approach utilizing a standard factor analysis.

2. DATASET

In this section, we introduce the analyzed dataset which is similarly structured as the large human-human Distress Assessment Interview Corpus (DAIC), that was described in [29]. The corpus, investigated in the present work, is recorded in a wizard-of-Oz controlled scenario where a virtual human interacts verbally and nonverbally in a semi-structured manner with a participant. [1]

The participants were recruited via Craigslist and were recorded at the University of Southern California Institute for Creative Technologies. In total 45 participants interacted with the virtual human. Unfortunately, only 39 interactions could be used in this study as for two of the interactions the logging of the virtual human's behavior failed to record. All participants who met the requirements (i.e. age greater than 18, and adequate eyesight) were accepted. Their mean age was 41.2 years ($\sigma = 11.6$; 27 male and 16 female).

For the recording of the dataset we adhered to the following procedure: After a short explanation of the study and giving consent, participants complete a series of questionnaires. These questionnaires included amongst others the Patient Health Questionnaire, depression module (PHQ-9) [29].

Upon completion of the questionnaires, the participants were asked to sit down in a chair facing a large screen equipped with a webcam (Logitech 720p) and a Microsoft Kinect recording the upper body of the participant. The screen and the participant were about 1 meter apart. The confederate helped the participant setup the head mounted microphone (Sennheiser HSP 4-EW-3) and then the virtual human would appear and proactively start the semi-structured conversation.

Two wizards in a separate room control the nonverbal behavior (e.g. head nods, smiles, back-channels) and the verbal utterances including questions and verbal back-channels of the virtual human by selecting pre-recorded behaviors from a menu interface. This wizard-of-Oz setup is the first step towards a fully automatic interaction. The interaction between the participants and the wizard-of-Oz controlled virtual human was designed as follows: The virtual human explains the purpose of the interaction and that it will ask a series of questions. It further, tries to build rapport with the participant in the beginning of the interaction with a series of shallow questions about Los Angeles. Then a series of more personal questions with varying polarity follow.

The questions were pre-recorded and animated using the SmartBody architecture [30]. Participants are then debriefed (i.e. the wizard is revealed), paid $25 to $35, and escorted out.

The PHQ-9 scale provides researchers with guidelines on how to assess the participant's condition based on the responses. Our participant-pool got split into 14 participants that scored positive on the PHQ-9 and 25 participants scored negative. The positive scoring participants correspond to moderate depression strength and above (cf. [20]). Additionally, the PHQ-9 provides us with a depression severity assessment ranging from 0 no depressive signs at all to a maximum score of 27.

3. AUTOMATIC AUDIO-VISUAL DESCRIPTORS

In the following, we introduce the audiovisual behavior descriptors extracted from a standard webcam signal and a head mounted microphone as introduced in Section 2. For the analysis of the participant behaviors we apply the multimodal sensing framework MultiSense that integrates several tracking technologies. The benefit of such a system is that the multiple technologies can run in parallel in a synchronized manner allowing for inter-module cooperation for performance improvement and information fusion. Our sensing system provides 3D head position-orientation, facial tracking based on GAVAM HeadTracker [23] and CLM-Z FaceTracker [3] and basic emotion analysis based on SHORE Face Detector [21]. In this analysis we also added results from the Computer Expression Recognition Toolbox (CERT) [22] for expression recognition. The acoustic measurements are currently not integrated in the sensing framework, but we plan to incorporate them in the near future. When available, we used our system's confidence report on the output to automatically screen out corrupt or noisy assessments when analyzing the signals.

3.1 Acoustic Descriptors

The automatically extracted features were chosen based on previous encouraging results in classifying voice patterns

[1] Sample interaction between the virtual agent and a human actor can be seen here: `http://youtu.be/ejczMs6b1Q4`

Table 1: Statistically significant acoustic measures discerning participants with moderate to severe depression (N = 14) and participants without (N = 25). The mean and standard deviation (in parentheses) values as well as the corresponding p-values derived from independent t-tests and Hedges' g value. Additionally, we provide correlation effects of the descriptors with the assessed depression severity as measured by PHQ-9 and significance estimates of observed linear correlation. The abbreviation *std* indicates that the standard deviation of the observed measure was chosen.

	Depression	No-Depression	p-value	Hedges' g	Pearson's ρ	Corr. p-value
NAQ	0.065 (0.035)	0.098 (0.026)	0.002	-1.070	-0.374	0.019
QOQ	0.275 (0.096)	0.360 (0.067)	0.002	-1.061	-0.335	0.037
OQ$_{NN}$	0.610 (0.017)	0.621 (0.011)	0.018	-0.806	-0.356	0.026
std NAQ	0.014 (0.007)	0.019 (0.006)	0.044	-0.683	-0.314	0.051
std QOQ	0.043 (0.018)	0.055 (0.015)	0.029	-0.745	-0.379	0.017
HeadRotation	0.035 (0.024)	0.056 (0.037)	0.065	-0.622	-0.267	0.100
HeadMovement	0.118 (0.063)	0.180 (0.093)	0.033	-0.725	-0.349	0.030
EmotionNeutral	0.473 (0.196)	0.339 (0.128)	0.014	0.840	0.198	0.228
EmotionVariability	0.516 (0.175)	0.645 (0.162)	0.026	-0.757	-0.054	0.745

Table 2: Statistically significant acoustic measures discerning participants with moderate to severe depression (N = 14) and participants without (N = 25). The mean and standard deviation (in parentheses) values as well as the corresponding p-values derived from independent t-tests and Hedges' g value. Additionally, we provide correlation effects of the descriptors with the assessed depression severity as measured by PHQ-9 and significance estimates of observed linear correlation. The abbreviation *std* indicates that the standard deviation of the observed measure was chosen.

	Dep	No-Dep	p-value	Hedges' g	Pearson's ρ
NAQ	0.065 (0.035)	0.098 (0.026)	0.002	-1.070	-0.374*
QOQ	0.275 (0.096)	0.360 (0.067)	0.002	-1.061	-0.335*
OQ$_{NN}$	0.610 (0.017)	0.621 (0.011)	0.018	-0.806	-0.356*
std NAQ	0.014 (0.007)	0.019 (0.006)	0.044	-0.683	-0.314
std QOQ	0.043 (0.018)	0.055 (0.015)	0.029	-0.745	-0.379*
HeadRotation	0.035 (0.024)	0.056 (0.037)	0.065	-0.622	-0.267
HeadMovement	0.118 (0.063)	0.180 (0.093)	0.033	-0.725	-0.349*
EmotionNeutral	0.473 (0.196)	0.339 (0.128)	0.014	0.840	0.198
EmotionVariability	0.516 (0.175)	0.645 (0.162)	0.026	-0.757	-0.054
Factor 1	-1.812 (3.167)	1.015 (2.352)	0.003	-1.039	-0.390*
Factor 2	-0.955 (1.489)	0.535 (1.987)	0.020	-0.798	-0.308

of suicidal adolescents [27], depression [28] as well as the features' relevance for characterizing voice qualities on a breathy to tense dimension [26, 18].

The first three features are derived from the glottal source signal estimated by iterative adaptive inverse filtering (IAIF, [1]). The output is the differentiated glottal flow. The normalized amplitude quotient (**NAQ**, [2]) is calculated using:

$$\text{NAQ} = \frac{f_{ac}}{d_{peak} \cdot T_0} \tag{1}$$

where d_{peak} is the negative amplitude of the main excitation in the differentiated glottal flow pulse, while f_{ac} is the peak amplitude of the glottal flow pulse and T_0 the length of the glottal pulse period. The quasi-open quotient (**QOQ**, [15]) is also derived from amplitude measurements of the glottal flow pulse. The quasi-open period is measured by detecting the peak in the glottal flow and finding the time points previous to and following this point that descend below 50% of the peak amplitude. The duration between these two time-points is divided by the local glottal period to get the QOQ parameter. Further, we extract **OQ**$_{NN}$ a novel parameter estimating the open quotient using standard Mel frequency cepstral coefficients and a trained neural network for open quotient approximation [19].

The final feature involves a dyadic wavelet transform using $g(t)$, a cosine-modulated Gaussian pulse similar to that used in [9] as the mother wavelet:

$$g(t) = -cos(2\pi f_n t) \cdot exp\left(-\frac{t^2}{2\tau^2}\right), \tag{2}$$

where the sampling frequency $f_s = 16$ kHz, $f_n = \frac{f_s}{2}$, $\tau =$

$\frac{1}{2f_n}$ and t is time. The wavelet transform, $y_i(t)$, of the input signal, $x(t)$, at the i^{th} scale, s_i, is calculated by:

$$y_i(t) = x(t) * g\left(\frac{t}{s_i}\right), \tag{3}$$

where $*$ denotes the convolution operator and $s_i = 2^i$. This functions essentially as an octave band zero-phase filter bank.

3.2 Face Descriptors

We further investigate nonverbal indicators of depression using visual cues extracted from the web-camera video aimed at the participant's face. In particular, we are interested in the participant's variability in emotional expressivity and motor retardation.

Emotional Variability: Reduced facial behavior, also mentioned as lack of emotional variability, is considered a valid indicator for depression; and in clinical studies a 'flat, mask like face' has also been reported as indicator of depression [11]. This serves as good motivation to examine *emotional variability* as a feature, and also the intensity of a *neutral* face that can be another measure of 'emotional flatness'. For this descriptor, we will use the automatic measurements of basic expressions of emotion plus 'Neutral' face which measures lack of emotions: {*Anger*, *Disgust*, *Contempt*, *Fear*, *Joy*, *Surprise*, *Sadness*, *Neutral*}. Looking at the variance of these expressions all together, is a good measure of emotion variability as discussed above. In the same category, the intensity of the 'Neutral' expression is a good measure of emotional flatness, or lack of emotion.

Motor Variability or motor retardation has also been observed in depressed population [11] including reduced hand gesturing and/or head movements. As a measure of motor variability we will look at the *head movement variance*. We will extract signals of head rotation in all three directions {*HeadRX* (Head rotation-Up/Down), *HeadRY* (Head rotation-side), *HeadRZ* (Head tilt)}. From these signals we can extract information about the head gaze and the head rotation variability.

4. EVALUATION

In the following, we provide the evaluation results of our statistical descriptor evaluation for both unimodal as well as multimodal behavior descriptors. Further, we provide results using third degree polynomial support vector machines for the unimodal and early fusion classification. All experiments are conducted using a leave-one-speaker-out strategy: for the training of the classifiers in each fold, we leave out the observed features of one speaker entirely from the training and test the classifiers on the speech of the left-out speaker.

As a measure of effect size we will use Hedge's g [17], a descriptive statistic that conveys the estimated strength of an effect by estimating how many standard deviations separate the two distribution means. We consider a Hedge's g$\leq$-0.4 to show existence of at least moderate effect with a negative trend. In our case this means that the depressed population shows lower scores on that indicator than the non-depressed. Symmetrically, an indicator with Hedge's g$\geq$0.4 is considered to have an effect with a positive trend. We also report the t-test statistical significance p-value of the difference of the distributions between depressed and non-depressed participants, to complement the Hedge's g effect size.

Table 3: Uni- and multimodal classification results. Mean accuracy (in %), $\mathbf{F_1}$ measure, with associated precision and recall are reported.

	Accuracy	F_1	Precision	Recall
Acoustic	51.28	0.51	0.52	0.53
Visual	64.10	0.60	0.60	0.59
Fusion	**89.74**	**0.88**	**0.93**	**0.86**

4.1 Unimodal Evaluation

First, we evaluate the unimodal audiovisual features statistically. As seen in Table 2, we find significant differences for almost all features below the level of $p < 0.05$ for both acoustic and visual features. The acoustic features reveal more tense voice qualities for subjects with depression and less voice quality variations over the interviews. Further, the visual features reveal reduced movement variations for the head and reduced emotionality and increased neutrality respectively. Additionally, some features correlate significantly with the depression severity scale as measured by PHQ-9, with a negative trend, i.e. Pearson's $\rho <$ -0.32.

Based on these findings, we conducted a basic polynomial support vector machine classification experiment with a leave-one-speaker-out testing approach. The results are reported in Table 3. While the acoustic modality only reaches an average accuracy of 51.28% and a mean F_1 of 0.51, the visual features reach 64.10% accuracy with a mean F_1 of 0.60. The multimodal results are discussed below in Section 4.2.

4.2 Multimodal Evaluation

Similarly to the unimodal evaluation, we conducted a basic polynomial support vector machine classification experiment with a leave-one-speaker-out validation approach utilizing both modalities in an early fusion experiment. The feature vectors containing information of both the acoustic and visual modalities are concatenated and used for training of the classifier. The results are reported in Table 3. The early fusion experiment clearly outperforms the single modalities with a mean accuracy of 89.74% and a mean F_1 of 0.88. Only false reject errors were observed, i.e. four depressed subjects were classified as not-depressed.

4.3 Joint Multimodal Factor Analysis

Additionally to the early fusion classification, we conduct a standard factor analysis to assess if there exist joint multimodal factors within the extracted audiovisual features that explain a good amount of the variance within the data. The factor analysis revealed two common factors among the data. The values for each factor are computed as linear combinations of the observed unimodal features and the corresponding factor loadings above the threshold $\theta = 0.3$. The two extracted factors reveal strong significant effects between the groups of depressed and non-depressed subjects with Hedges' g $\in$ -1.04, -0.80, which is a fairly significant effect ($p \leq 0.02$). Additionally, the first factor shows a strong negative correlation with the depression severity as measured by PHQ-9 with Pearson's $\rho =$ -0.39. These effects are visualized in Figure 1.

The loadings for each of the two factors are reported in

Table 4: Statistics on the two identified factors found in the multimodal data. Both show statistically significant measures discerning participants with moderate to severe depression (N = 14) and participants without (N = 25). The mean and standard deviation (in parentheses) values as well as the corresponding p-values derived from independent t-tests and Hedges' g value. Additionally, we provide correlation effects of the descriptors with the assessed depression severity as measured by PHQ-9 and significance estimates of observed linear correlation.

	Depression	No-Depression	p-value	Hedges' g	Pearson's ρ	Corr. p-value
Factor 1	-1.812 (3.167)	1.015 (2.352)	0.003	-1.039	-0.390	0.014
Factor 2	-0.955 (1.489)	0.535 (1.987)	0.020	-0.798	-0.308	0.056

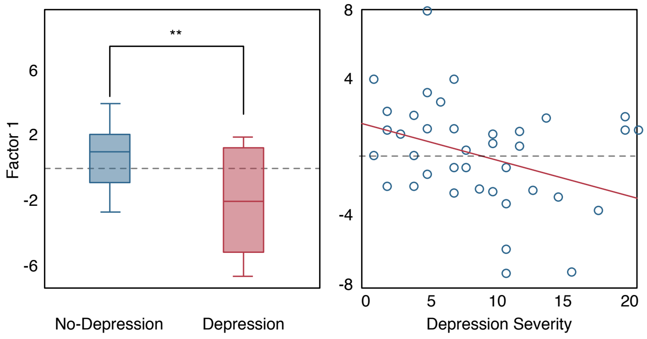

Figure 1: (a) Boxplot showing the distributions of the first joint multimodal factor for subjects with and without depression as measured with PHQ-9. (b) Scatterplot showing the correlations between depression severity and the first joint multimodal factor. It is seen that a correlation is observed with ρ = 0.39. The regression line fit to the data is shown in red.

Table 5: Factor loadings for each feature. Loadings are not listed if absolute value is below θ = 0.3.

	Factor 1	Factor 2
NAQ		0.8768
QOQ		0.9924
OQ$_{NN}$		
std NAQ		0.3436
std QOQ	0.9651	
HeadRotation	0.9491	
HeadMovement		
EmotionNeutral	0.8092	
EmotionVariability	0.5908	

Table 5. It is revealed that the first factor indeed is a multimodal factor that receives contributions of both the audio and visual modalities. Loadings are not listed if the absolute value is below θ. Two features, were not selected for any of the two factors.

5. DISCUSSION

In the following we discuss the findings in more detail. Based on the unimodal analysis we could identify several acoustic and visual indicators of depression. In particular, we observe increased tenseness in the voice of depressed subjects as measured with the standard voice quality measures NAQ, QOQ, and OQ$_{NN}$. All three measures show the same tendency of more tense voices for depressed subjects, as reported in Table 2. All observed p-values are below a threshold of 0.05 and Hedges' g, a reliable measure of effect size, ranges $\in$ [-1.07, -0.81], which corresponds to a moderate to strong effect. Additionally, these voice quality measures show significant negative correlations with depression severity with Pearson's $\rho <$ -0.30.

Within the speaker-independent classification experiments, the acoustic features only score below chance, which is disappointing. However, we believe that the chosen parameters for the support vector classification did yield odd support vectors, as with a simple linear discriminant analysis the acoustic features yield an accuracy of 76.92% with a mean F_1 measure of 0.75.

Psychomotor retardation and reduced emotional expressivity are two common concepts found in related work on depression assessment that are indicative of psychological disorders [24, 6]. The visual features that we extracted were chosen based on these two concepts, as introduced in Section 3.2. Our automatically extracted behavior descriptors reveal reduced head movements and emotional expressivity as measured with our multimodal sensing framework for depressed subjects (cf. Table 2). Not all visual features show linear correlation effects with depression severity (i.e. reduced emotional expressivity measures).

For the visual features the linear discriminant analysis did not yield acceptable results as the linear correlation analysis shows that there is a higher dimensional relationship between the observed behavior descriptors and the depression severity.

Both individual modalities, are outperformed by the simple early feature-level fusion classification results that yield almost 90% accuracy in the speaker-independent classification experiment. The support vector machine only confuses four depressed subjects as non-depressed subjects in this experiment.

In order to investigate, potential multimodal behavior indicators we computed the main factors within the at feature-level combined observations. We found two main factors to be present in the data. One of the two factors indeed shows strong multimodal loadings and combines both measures of the acoustic and visual modalities (cf. Table 5). Further, this factor shows strong statistical significance in the t-test and the strongest observed Pearson ρ = -0.390 (cf. Table 4). Also, the factor combines a measure of voice quality variation (i.e. standard deviation of NAQ) with the visual factors representing reduced emotional expressivity to form an audiovisual indicator for reduced multimodal expressivity. This is a motivating result that we seek to investigate further with additional modalities and observations in the future. We hope to find meaningful combination of features within these kinds of multimodal joint factors from varying modalities that correspond to concepts observed by psychologists in the past, such as psychomotor retardation, or positive affective attenuation and negative affective potentiation [6].

6. CONCLUSIONS

Based on our research goals we can identify the following main contribution of this work: **1** We could find unimodal nonverbal indicators of depression using acoustic and visual measures. Additionally, some of them show strong linear correlations with depression severity. **2** Further, multimodal early feature-level fusion of acoustic and visual measures in a subject-independent study revealed considerably better classification results close to 90% in accuracy than unimodal classification performance. **3** A basic factor analysis reveals two underlying factors within the combined feature vectors. The first joint factor is indeed a multimodal factor that shows the strongest statistical significance within this work and statistical significant differences for the participant groups depression/no-depression. These results are encouraging and we would like to investigate this type of multimodal factor analysis in the future and incorporate indicators of different modalities, such as body movement or verbal content of the speech. Additionally, we seek to investigate more sophisticated multimodal analysis paradigms and spatio-temporal algorithms and indicators.

Acknowledgements

This work is supported by DARPA contract W911NF-04-D-0005 and U.S. Army Research, Development, and Engineering Command and. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

7. REFERENCES

[1] P. Alku, T. Bäckström, and E. Vilkman. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, 1992.

[2] P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[3] T. Baltrusaitis, P. Robinson, and L. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2610–2617, 2012.

[4] L. Batrinca, G. Stratou, L.-P. Morency, and S. Scherer. Cicero - towards a multimodal virtual audience platform for public speaking training. In *Proceedings of Intelligent Virtual Agents (IVA) 2013*, pages 116–128. Springer, 2013.

[5] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin. Psychomotor retardation in depression: biological underpinnings, measurement, and treatment. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(2):395–409, 2011.

[6] L. M. Bylsam, B. H. Morris, and J. Rottenberg. A meta-analysis of emotional reactivity in major depressive disorder. *Clinical Psychology Review*, 28:676–691, 2008.

[7] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Ying, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, 2009.

[8] N. Cummins, J. R. Epps, M. J. Breakspear, and R. Goecke. An investigation of depressed speech detection: Features and normalization. In *Proceedings of Interspeech 2011*. ISCA, 2011.

[9] C. d'Alessandro and N. Sturmel. Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude. *Sadhana*, 36(5):601–622, 2011.

[10] J. K. Darby, N. Simmons, and P. A. Berger. Speech and voice parameters of depression: a pilot study. *Journal of Communication Disorders*, 17(2):75–85, 1984.

[11] H. Ellgring. *Nonverbal communication in depression.* Cambridge University Press, Cambridge, 1989.

[12] M. Elliott, M. A. Clements, J. W. Peifer, and L. Weisser. 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.

[13] L. A. Fairbanks, M. T. McGuire, and C. J. Harris. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of Abnormal Psychology*, 91(2):109–119, 1982.

[14] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. G. Gailey, and C. Levinton. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of Psychiatric Research*, 27(3):309–319, 1993.

[15] T. Hacki. Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. *Folia Phoniatrica*, pages 43–48, 1989.

[16] J. A. Hall, J. A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinician-patient interaction. *Applied and Preventive Psychology*, 4(1):21–37, 1995.

[17] L. V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981.

[18] J. Kane, S. Scherer, M. Aylett, L.-P. Morency, and C. Gobl. Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7982–7986. IEEE, 2013.

[19] J. Kane, S. Scherer, L.-P. Morency, and C. Gobl. A comparative study of glottal open quotient estimation techniques. In *to appear in Proceedings of Interspeech 2013*. ISCA, 2013.

[20] K. Kroenke, R. L. Spitzer, and J. B. W. Williams. The phq-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[21] C. Kublbeck and A. Ernst. Face detection and tracking in video sequences using the modifiedcensus transformation. *Image and Vision Computing*, 24(6):564 – 572, 2006.

[22] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305, 2011.

[23] L.-P. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *8th IEEE International Conference on Automatic Face Gesture Recognition (FG08)*, pages 1–8, sept. 2008.

[24] J. E. Perez and R. E. Riggio. *Nonverbal social skills and psychopathology*, pages 17–44. Nonverbal behavior in clinical settings. Oxford University Press, 2003.

[25] J. T. M. Schelde. Major depression: Behavioral markers of depression and recovery. *The Journal of Nervous and Mental Disease*, 186(3):133–140, 1998.

[26] S. Scherer, J. Kane, C. Gobl, and F. Schwenker. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech and Language*, 27(1):263–287, 2013.

[27] S. Scherer, J. P. Pestian, and L.-P. Morency. Investigating the speech characteristics of suicidal adolescents. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 709–713. IEEE, 2013.

[28] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Proceedings of Interspeech 2013*. ISCA, 2013.

[29] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*. IEEE, 2013.

[30] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann. Smartbody: behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 1*, AAMAS '08, pages 151–158. International Foundation for Autonomous Agents and Multiagent Systems, 2008.