CROSS-CORPUS DEPRESSION PREDICTION FROM SPEECH

Vikramjit Mitra[1], Elizabeth Shriberg[1], Dimitra Vergyri[1], Bruce Knoth[1], Ronald M. Salomon[2]

[1]SRI International, Menlo Park, CA, USA.
[2]Northwestern University, Feinberg School of Medicine, Chicago, IL, USA
[1]{vikramjit.mitra,elizabeth.shriberg,dimitra.vergyri,bruce.knoth}@sri.com, [2]rsalomon@nmff.org

ABSTRACT

Research on detecting depression from speech has advanced in recent years, but most work has focused on the analysis of one corpus at a time. Given that clinical corpora are typically small, it is important to explore approaches that generalize across corpora and that could ultimately be adapted to new data. We study a new corpus of patient-clinician interactions recorded when patients are admitted to a hospital for suicide risk and again when they are released. To train prediction models, we use the 2014 AVEC challenge German speech dataset, which differs from our data in many factors (including language, context, speakers, and recording conditions). Results reveal that some of the AVEC-trained models predict scores for the clinical data that correlate with both HAM-D depression scores and with the pre-/post-admission ordering. A KL-divergence analysis within the clinical data confirms that the same feature set captures changes correlated with the HAM-D scores. Finally, read versus spontaneous speech samples in both corpora behave differently with respect to the best features and modeling approaches. Implications for the cross-corpus prediction of depression are discussed.

Index Terms—*depression detection, cross-corpus modeling, mental health, acoustic features, prosodic features, articulatory features, phonetic features, AVEC Challenge.*

1. INTRODUCTION

Speech offers important benefits for mental health monitoring, because it can be obtained and analyzed in a noninvasive, natural, and inexpensive manner, and can be used in telemedicine applications for remote assessment. The speech signal carries important information that may assist psychiatrists with clinical assessment. Central controls of laryngeal, pharyngeal, and nasal structures generate objectively measureable expressions of emotional stress [1], [2], [3], [4]. Mental health problems including schizophrenia [5], [6], [7], depression [8], [9], [10], and psychopathy [11] can affect speech prosody.

Prior studies have attempted to identify speech characteristics that can be used to detect different psychological conditions. For depression, measures of loudness, word production rate, and pause duration, which are highly vulnerable to non-specific effects of motor retardation, were used in [12]. Silverman [13] proposed to use vocal parameters of speech to detect suicide risk. Follow-up studies used speech spectral measures to successfully differentiate between the speech of near-term suicidal patients and depressed controls. France et al. [14] used long-term averages of extracted formant information. Yingthawornsuk et al. [15] used the percentages of the total power, the highest peak value, and its frequency location at which the percentages of the total power were found. Ozdas et al. [16] used lower-order mel-cepstral coefficients in Gaussian mixture models and unimodal Gaussian models. Work by Keskinpala et al. [17], [18] investigated energy in frequency-band features.

In this study we analyze a new corpus of real patient-clinician interactions related to suicide risk. Suicide is a serious public health problem that can have lasting harmful effects on individuals, families, and communities [19]. When a psychiatrist sees a patient, suicide risk is evaluated as part of a clinical interview. Researchers and psychiatrists assess mood by different techniques, including the Hamilton depression rating scale (HAM-D) [20] and the Beck depression inventory scale (BDI) [21]. Our study uses pairs of interview recordings (at admission and at release) per patient. In both cases, a clinician calculated the HAM-D depression score. The interactions per patient involve the same clinician and contain both interview and read speech portions, allowing speaking style comparisons for each speaker. We compare admission and release interviews for each speaker, and find consistent feature differences across patients. For privacy reasons, we use only non-lexical features. We also use BDI score predictor models trained with the 2014 Audio-Visual Emotion Recognition Challenge (AVEC-2014) dataset [22] to predict the HAM-D depression scores for the new patient-clinician interaction dataset. It should be noted that though BDI and HAM-D are both instruments used by clinicians to assess depression, the literature shows that these should be viewed as complementary instruments, since the first is patient-rated and the second clinician administered and they address somewhat different characteristics of personality. Studies report correlations of 0.4 to 0.7 between the two scores [23]. Moreover the 2014 AVEC data differs from the patient-clinician data in language, as AVEC-2014 is entirely in German and the patient-clinician data is in English.

2. DATA

We used data collected at Vanderbilt University, at the Emergency Room and Psychiatric Treatment Unit (PTU) offices. The patients were interviewed for 15 to 30 minutes about their feelings and life events. Then they were asked to read aloud a half page of text called "The Rainbow Passage." This short reading took 1–3 minutes. If a patient was admitted for therapy at the hospital, one or two follow-up sessions were also recorded; the last was the release interview from the facility.

We had at least two recorded interviews, corresponding to admittance to and release from the psychiatric facility, for 17 patients, 11 female and 6 male. The admittance interviews lasted on average 35 minutes, for a cumulative total of approximately 10 hours for all patients. The release interviews lasted an average of 21.5 minutes, totaling approximately 6 hours for all patients. After each interview, the clinician calculated the HAM-D depression score for the patient. The HAM-D is the most widely used clinician-administered depression assessment scale. It was designed for use after an unstructured clinical interview.

The AVEC-2014 challenge dataset is a subset of the AVEC-2013 audio-visual depression corpus [22], which contains 150 videos of subjects performing human-computer interaction tasks. The total number of subjects in the entire dataset is 84. Some subjects were recorded more than once: 18 subjects appear in three recordings; 31 appear in two; and the remaining 34 appear in only one recording. The duration of each recording ranges from 20 minutes to 50 minutes, with an average duration of 25 minutes. The total duration of all clips is 240 hours. The average age of the subjects was 31.5 years, with a standard deviation of 12.3 years and a range of 18 to 63 years. The data contained each participating individual's self-reported depression score according to the Beck depression rating scale [21]

The audio data was collected using a headset microphone connected to a computer and sampled at various sampling rates. The challenge data was split into three partitions of training, dev, and test sets, with 50 read-spontaneous pairs in each set, for a total of 300 task recordings. Partitions had similar distributions in terms of age, gender, and depression levels and no session overlap. Because AVEC organizers distributed depression labels for only the train and dev partitions, we used these partitions for this study.

3. FEATURES

We restricted ourselves to automatically extractable features that do not rely on words for reasons of privacy and practicality. For some patients, specific word usage is sensitive. Word features also require speech recognition, which may or may not be available at high enough performance levels for a particular individual or context. For each segment in the patient channel, we computed a range of features types intended to capture potential speech changes between their admission and release interviews. An important potential confound in this study is extrinsic (i.e., not due to the speaker) variability from first to last recording session. Although the same interviewer was used across sessions, it is possible that microphone gain, distances between speakers, or other aspects of the set up could change, and the differences between the interviews could partially reflect these changes. In real applications, such extrinsic issues will always be a factor, so we looked at a broad range of features, and we indicate for each feature our informal guess at robustness to extrinsic variation. A summary of features is given in Table 1.

Damped Oscillator Cepstral Coefficients (DOCC) [24], aim to model the dynamics of the hair cells within the human ear. In DOCC processing, speech is analyzed by a gammatone filter bank (GFB) that splits the signal into subbands. These subbands are used as the forcing functions to an array of damped oscillators whose response is used as the acoustic feature.

Normalized Modulation Cepstral Coefficient (NMCC) [24] is a perceptually motivated feature that tracks the amplitude modulation (AM) of subband speech signals. In NMCC processing, the speech signal is analyzed by using a time-domain GFB, and the AM trajectories of the resulting subband signals are used to produce a modulation spectrum, whose cepstral representation is used as the NMCC feature set.

Modulation of Medium Duration Speech Amplitudes (MMeDuSA) [[26], [27]] estimates the AM signals from bandlimited speech using a medium duration analysis window. Along with modulation-spectrum-based traditional cepstral features, it generates summary modulation information that plays an important role in tracking speech activity as well as in locating events such as vowel prominence/stress, etc.

Gammatone Cepstral Coefficients (GCCs) use GFBs to analyze the speech signal. The power of the resulting subband signals over an analysis window of 25 ms is computed, and their cepstral representation is used as the GCC feature.

Table 1. Feature details. The last column indicates expected robustness to non-speaker related variation between the two sessions.

Name	Type	Extraction region	Feature dim.	Est. robustness to extrinsic variation
tilt	vocal effort	voiced frames in segment	5	medium
dle onset and offset	vocal effort	voiceless->voiced transitions in segment	6	high
encon	rhythmicity	200 ms window in segment	7	high
f0	pitch	frame	1	high
f0pk	pitch at peaks	frames at peaks in segment	1	high
f0pk-stats	rhythmicity, rate, pitch	peak locations in segment	9 (stats)	high
DOCC	acoustic	26 ms window at 10 ms frame rate	13	high
NMCC	acoustic	26 ms window at 10 ms frame rate	13	high
MMeDuSA	acoustic	52 ms window at 10 ms frame rate	16	high
GCC	acoustic	26 ms window at 10 ms frame rate	13	medium

The encon feature [28] captures rhythmicity by looking at the periodicity of energy peaks within each segment. This feature models the contour of 10 ms c0 and c1 output from a mel-frequency cepstral coefficient (MFCC) frontend; each cepstral stream is mean-normalized over the utterance, making it robust to absolute level differences over sessions and within-session segments. A discrete cosine transform (DCT) is then taken over a 200-ms sliding window with a 100-ms shift. Vector components comprise the first 5 and 2 bases from the DCT over each window of c0 and c1, respectively.

Tilt features aim to capture vocal effort in a manner somewhat robust to extrinsic session variability by using methods developed in [28]. Features are extracted for voice frames. Voicing is determined by using a logistic regression classifier trained with the number of zero crossings, log energy, number of peaks in the autocorrelation of the window signal, and standard deviation of the inter-peak distance; the voicing threshold is set to 0.5. The five component tilt features include H2-H1, F1-H1, F2-H1, which reflect lower-order harmonics and formants given the microphone and room conditions. The last two features are measures of the spectral slope per frame and the difference between the maximum of the log-power spectrum and the maximum in the 2 kHz–3 kHz range.

The delta log energy (dle) [28] features target session-normalized vocal-effort detection using a sparse feature (output only once per voiced-voiceless transition). The feature is the difference in log energy at the transition, with an updated implementation from [28]. Dle-onset features are triggered at each boundary from voiceless to voiced speech; dle-offset features occur at each boundary from voiced to voiceless speech.

Pitch-related features include f0, f0pk, and f0pk-stats features. F0 is computed using default parameter settings for the snack PRAAT-style pitch tracker [29] and is used only for voiced regions according to the snack output. We expect pitch features to be robust to extrinsic variability. The f0-peak features record only the subset of pitch values found by an automatic peak-picking

algorithm [30] run within each segment. Statistics computed for the f0pk-stats features include both pitch-level and pitch-peak distribution information. Pitch level includes the mean, max, and standard deviation of the peak pitches in the segment. Pitch peak distributions are intended to capture not pitch but rather the temporal distribution of pitch-accented syllables in the segment. These features include: peak count; peak rate (count divided by segment duration); mean and maximum interpeak distances; and the location of the maximum peak in the segment (e.g., early versus late), both as a percentage of the distance into the segment and as raw distance into the segment.

4. FEATURE MODELING AND ANALYSIS

The Vanderbilt University (VU) PTU data, though controlled in some aspects, is not large enough to train background models. To better understand how the features mentioned in the last section correlate with that of a subject's psychological state before and after treatments, we first focused on analyzing the individual features. We performed per-speaker analysis of each acoustic feature from the waveforms recorded before and after therapy. As mentioned earlier, the VU-PTU dataset included 17 speakers for whom paired data existed from before and after therapy. The data contained HAM-D depression scores ranging from 3 to 35, and based on clinician suggestion those scores can be quantized into four broad classes: (1) ND: no depression; (2) MD: mild depression; (3) D: moderate depression; and (4) SD: severe depression. For four speakers, the broad depression classes were the same before and after therapy; for four speakers, it changed from SD to ND; for one speaker, it changed from MD to ND; for four speakers, it changed from SD to D; for three speakers, it changed from SD to MD; and for the remaining speaker, it changed from D to ND. Hence, the amount of improvement in HAM-D depression levels varied across subjects in the dataset, which renders the database quite sparse in terms of observed depression levels per speaker and their variance before and after therapy.

In order to gain some understanding about the data distribution and the relevance of the individual features to depression levels, we investigated the Kullback-Leibler divergence (KLD) between a given feature from before and after therapy that demonstrated a change in depression levels. For single-order GMM $\lambda(\mu,\Sigma,w)$, with a full covariance matrix Σ, the KLD has an analytical solution [31]:

$$D_{KL}(P||Q) = tr\left(\Sigma_P\Sigma_Q^{-1}\right) + tr\left(\Sigma_Q\Sigma_P^{-1}\right) - 2S + tr[\left(\Sigma_P^{-1} + \Sigma_Q^{-1}\right)\left(\mu_P - \mu_Q\right)\left(\mu_P - \mu_Q\right)^T] \quad (1)$$

where $P(i)$ and $Q(i)$ are the probability distributions of two discrete random variables, and $\lambda_P(\mu_P,\Sigma_P,w_P)$ and $\lambda_Q(\mu_Q,\Sigma_Q,w_Q)$ are their single-order GMM models. Single-order GMM models that were adapted from their single-order UBM models (given a feature set) were trained. To normalize the per-session variability, we created N (where $N > 3$) utterance-based audio segments for each speaker i and session s (here, session means recordings before or after therapy; each speaker had two sessions s_1 and s_2). All the features were computed for each of those segments, and we computed $D_{KL,i,s}$ where i is the speaker label and s is the session. Given this, the intra-speaker KL-divergence was computed as the geometric mean of all the KL-divergences between the segments of a given speaker and given session, as shown in (2):

$$D_{KL,i,s_{k|k=1,2}}(S_k) = \sqrt[2(N-1)]{\prod_{j=1,l=1,l\neq m}^{N,N} D_{KL,i,s_k}(S_j||S_l)} \quad (2)$$

Given the intra-speaker divergences, we computed the normalized divergence for each speaker as:

$$D_{KL,i_{NORM}}(S_1||S_2) = \frac{D_{KL,i}(S_1||S_2)}{\sqrt{D_{KL,i,s1}D_{KL,i,s2}}} \quad (3)$$

For depression-score prediction experiments, all acoustic features (DOCC, MMeDuSA, GCC, and NMCC) were mean- and variance-normalized for each speaker. In our prior experiment [32], we found i-vectors [33] to be an effective representation of the acoustic features. To compensate for the limited amount of data available from both AVEC-2014 dataset and the VU-PTU data, we constrained the Universal Background Model (UBM) to have 16 Gaussian components and the i-vector subspace to have only 30 dimensions. The i-vectors were length normalized in our experiments. Note that in [32] we demonstrated that MFCC failed to perform as well as the acoustic features used in this work, and hence they were not used in the experiments presented here.

For the remaining features, we obtained a fixed-length representation by computing statistics over feature distributions, including mean and variance as well as distances between measurement locations that capture speaking rate information (e.g., distance between pitch peaks; durations of voiced regions).

5. RESULTS AND DISCUSSION

We used formula (3) to compute the normalized divergence for each feature, by speaker. Table 2 shows the normalized KLD, where the values were averaged (using the geometric mean and arithmetic mean) across speakers, for conditions in which a speaker transitioned from SD to D, and MD to ND. The last two columns of Table 2 also show the same for conditions in which the speakers did not show any change in broad depression levels between pre- and post-therapy. As observed from Table 2, most of the features did reflect increased KLD values for subjects who showed a change in depression levels. The VU-PTU dataset also was split for each speaker-session condition into read speech (reading) and spontaneous speech (interview). Tables 3 and 4 show the normalized KLD measures for conditions with and without change in depression levels at those two splits.

Tables 3, 4, and 5 reveal that most features gave higher KLD values for interview sessions than for reading sessions. This may suggest that spontaneous speech is better than read speech for analysis of depression, but further work is warranted. Also note the differences in KLD values for the different features, which may be due to the dynamic range difference of the features. Some of the features (such as tilt, encon, dle_onset, and dle_offset) showed higher KLD values for cases where broad depression levels differed, which is an expected and encouraging observation. Though most of the cepstral features demonstrated confusing observations, GCC and MMeDuSA showed lower KLD values for cases where depression levels stayed the same compared to where depression levels differed in Tables 2 and 3 (i.e., full data and interview), but showed exactly the reverse trend in Table 4, (i.e., reading). One hypothesis is that, in read speech, the biomarkers for depression may not be distinct in cepstral features; hence, the pattern is different.

For depression-score estimation, we trained feature-specific, single-layer artificial neural nets (ANNs) using the AVEC-2014 training set. The number of neurons was optimized using the AVEC-2014 dev set. Table 5 shows the Pearson's product moment correlation (PPMC) coefficient between the VU-PTU HAM-D scores and the ANN output BDI score. Also note that, in this experiment, the dle onset and offset features were combined to train and test a single ANN. Further, we performed m-way score fusion (fusion performed by simple averaging of the scores) among the eight systems shown in Table 5, and the best fusion came from five systems (DOCC, dle, encon, f0_peaks, and NMCC); and its

PPMC score is shown in the last row of Table 5. Figure 1 shows the scatter plot of the estimated Beck depression score and the target HAM-D scores from the best system fusion, which shows that the estimated Beck depression scores are correlated with the target HAM-D scores. For the results shown in table 5 and the scatter plot shown in Figure 1, we have used both the interview (spontaneous) and reading parts of the VU-PTU dataset.

Table 2. Averaged normalized KLD values for subjects who showed a change in depression level and for those who did not exhibit a change in depression level pre- and post-therapy.

Features	Normalized KLD for sessions in which depression levels changed		Normalized KLD for sessions in which depression levels did not change	
	Geometric mean	Arithmetic mean	Geometric mean	Arithmetic mean
tilt	56.85	89.24	11.16	33.69
f0_peaks	6.05	12.99	15.12	16.92
encon	15.00	38.60	7.17	14.60
dle_onset	21.76	91.13	6.00	42.20
dle_offset	41.20	139.82	3.45	18.57
DOCC	12.43	22.21	11.01	25.61
GCC	7.24	8.98	6.51	8.12
MMeDuSA	8.16	12.07	6.16	7.06
NMCC	6.78	8.58	7.40	8.77

Table 3. Averaged normalized KLD values for interviews for subjects who showed a change in depression level and for those who did not pre- and post-therapy.

Features	Normalized KLD for sessions in which depression levels changed		Normalized KLD for sessions in which depression levels did not change	
	Geometric mean	Arithmetic mean	Geometric mean	Arithmetic mean
tilt	69.34	102.81	12.54	47.72
f0_peaks	8.19	15.60	13.67	14.33
encon	15.85	37.68	8.86	18.69
dle_onset	42.40	89.81	4.71	55.16
dle_offset	30.28	213.26	5.56	29.39
DOCC	17.16	26.52	11.77	26.21
GCC	7.27	8.73	6.48	8.27
MMeDuSA	8.26	12.46	6.93	8.07
NMCC	7.62	8.47	7.76	9.66

Table 4. Averaged normalized KLD values for reading for subjects who showed a change in depression level and for those who did not pre- and post-therapy.

Features	Normalized KLD for sessions in which depression levels changed		Normalized KLD for sessions in which depression levels did not change	
	Geometric mean	Arithmetic mean	Geometric mean	Arithmetic mean
tilt	1.22	1.96	0.26	0.32
f0_peaks	13.17	19.80	16.06	36.73
encon	16.63	48.55	9.55	13.00
dle_onset	74.70	94.94	4.26	56.34
dle_offset	17.60	198.40	2.40	6.59
DOCC	15.57	21.37	22.18	43.75
GCC	10.17	11.13	12.92	15.99
MMeDuSA	12.78	13.53	13.73	19.36
NMCC	11.21	12.54	13.63	15.92

Finally, we also computed the PPMC between the pre-/post-admission differential HAM-D scores and the estimated BDI scores across all speakers using the best fusion outputs, the correlation was 0.62, which is in line with [23]. Note that according to [23] HAM-D and actual BDI scores exhibit a correlation of 0.4-0.7, while HAM-D and the predicted BDI scores (in this study) exhibit a correlation of 0.62. Also it needs to be emphasized that the training and the testing data are not merely different in the sense of acoustic conditions but comes from an entirely different language and culture, specifically.

Table 5. Pearson's product moment correlation (PPMC) coefficient between the VT-PTU HAM-D scores and the ANN output BDI score from different systems and system-fusion.

Features	#Neurons in ANN	r_{PPMC}
tilt	50	-0.1644
f0_peaks	25	0.2714
encon	300	0.3461
dle (onset + offset)	50	0.4125
DOCC	700	0.4502
GCC	700	0.2250
MMeDuSA	500	0.1350
NMCC	700	0.1142
Best fusion	-	0.6252

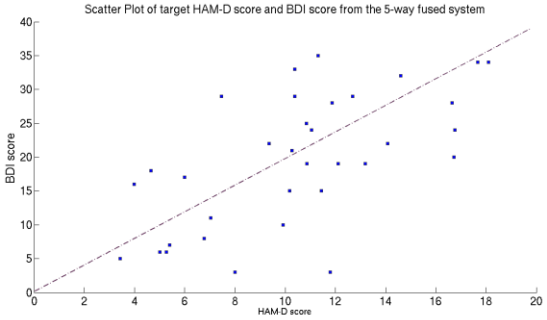

Figure 1. Correlation between the obtained BDI score and the target HAM-D scores from the 5-way fused system for all speakers in the VU-PTU dataset.

6. CONCLUSION

We ran a cross-corpus study of depression score prediction, using clinical annotations as a gold standard. A novel English corpus with depression scores served as test data and the 2014 AVEC challenge German speech dataset served as training data. Despite corpus differences in language, context, recording conditions, and depression scoring instrument, the AVEC-trained models predict scores for the clinical data that correlate both with HAM-D depression scores and with the pre-/post-admission ordering. A KL-divergence analysis within the clinical data confirmed that the same feature set captures changes correlated with the HAM-D scores. Finally, read versus spontaneous speech samples in both corpora behaved differently with respect to the best features and modeling approaches, suggesting that clinical depression collections should consider both types of data in their protocols. Overall we find promising results for cross-corpus prediction of depression when a range of speech features beyond standard MFCCs is employed. The correlation is in line with (or even better than) that expected from cross-corpus human annotations.

7. ACKNOWLEDGMENTS

We thank Michelle H. Sanchez, Martin Graciarena, Andreas Kathol, Colleen Richey, Mitchell McLaren and Kristin Precoda for help and suggestions, and Suman Ravuri for assistance with the dle implementation. This research was partially supported by NSF Grant #IIS-1162046.

8. REFERENCES

[1] F.C. Merewether, M. Alpert, "The components and neuroanatomic bases of prosody," J. of Comm. Disord., Vol. 23(4-5), pp. 325–336, 1990. Review.

[2] A.J. Friedhoff, M. Alpert, R. Kurtzberg, "An electro-acoustic analysis of the effects of stress on voice," J of Neuropsychiatr., Vol. 5, pp. 266–272, 1964.

[3] M. Alpert, R. Kurtzberg, A. Friedhoff, "Transient voice changes associated with emotional stimuli," Arch. Gen. Psychiatry, Vol. 8, pp. 362–365, 1963.

[4] C. Sobin, M. Alpert, "Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy," J. of Psycholinguist Res., Vol. 4, pp. 347–365, 1999.

[5] J.C. Borod, M. Alpert, A. Brozgold, C. Martin, J. Welkowitz, L. Diller, E. Peselow, B. Angrist, A. Lieberman, "A preliminary comparison of flat affect schizophrenics and brain-damaged patients on measures of affective processing," J. of Comm. Disord., Vol.2, pp. 93–104, 1989.

[6] R.J. Shaw, M. Dong, K.O. Lim, W.O. Faustman, E.R. Pouget, M. Alpert, "The relationship between affect expression and affect recognition in schizophrenia," Schizophr. Res., 37(3), pp. 245–250, 1999.

[7] F. Tolkmitt, H. Helfrich, R. Standke, K.R. Scherer, "Vocal indicators of psychiatric treatment effects in depressives and schizophrenics," J. Comm Disorders, Vol.15, pp. 209–222, 1982.

[8] J.K. Darby, N. Simmons, P.A. Berger, "Speech and voice parameters of depression: a pilot study," J of Commun. Disord.,17(2), pp. 75–85, 1984.

[9] M. Garcia-Toro, J.A. Talavera, J. Saiz-Ruiz, A. Gonzalez, "Prosody impairment in depression measured through acoustic analysis," J Nerv. Ment. Dis., 188(12), pp. 824–829, 2000.

[10] M. Alpert, E.R. Pouget, R.R. Silva, "Reflections of depression in acoustic measures of the patient's speech," J Affect Disord., 66, pp. 59–69, 2001.

[11] S.M. Louth, S. Williamson, M. Alpert, E.R. Pouget, R.D. Hare "Acoustic distinctions in the speech of male psychopaths," J Psycholinguist Res., 27(3), pp. 375–384, 1998.

[12] A.C. Trevino, T.F. Quatieri, N. Malyska, "Phonologically-based biomarkers for major depressive disorder," EURASIP Journal on Advances in Signal Processing, pp. 2011-2042, 2011.

[13] S.E. Silverman, "Vocal parameters as predictors of near-term suicidal risk," U.S. Patent 5, 148, 483, Sept. 1992.

[14] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, D.M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE Transaction on Biomedical Engineering, Vol. 47(2), pp. 829–837, 2000.

[15] T. Yingthawornsuk, H. Kaymaz Keskinpala, D. France, D.M. Wilkes, R.G. Shiavi, R.M. Salomon, "Objective estimation of suicidal risk using vocal output characteristics," Proc. of Interspeech, pp. 649–652, 2006.

[16] A. Ozdas, R.G. Shiavi, D.M. Wilkes, M.K. Silverman, S.E. Silverman, "Analysis of vocal tract characteristics for near-term suicidal risk assessment," Methods of Information in Medicine, Vol. 43, pp. 36–38, 2004.

[17] H. Kaymaz Keskinpala, T. Yingthawornsuk, D.M. Wilkes, R.G. Shiavi, R. M. Salomon, "Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands," Fifteenth European Signal Processing Conference (EUSIPCO), pp. 2229–2233, September 2007

[18] H. Kaymaz Keskinpala, T. Yingthawornsuk, D.M. Wilkes, R.G. Shiavi, R.M. Salomon, "Distinguishing high risk suicidal subjects among depressed subjects using mel-frequency cepstrum coefficients and cross validation technique," 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2007), pp. 157–160, 2007.

[19] Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Web-based Injury Statistics Query and Reporting System (WISQARS) [online], 2010.
Available from: www.cdc.gov/injury/wisqars/index.html

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

[21] A. Beck, R. Steer, R. Ball, and W. Ranieri, "Comparison of Beck depression inventories -ia and -ii in psychiatric outpatients," Journal of Personality Assessment, 67(3), pp. 588-597, 1996.

[22] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, "AVEC 2014—3D dimensional affect and depression recognition challenge," Proc. of AVEC2014, 2014.

[23] P. Richter, J. Werner, A. Heerlein, A. Kraus, H. Sauer: "On the validity of the Beck Depression Inventory", Psychopathology 31, pp. 160–168, 1998.

[24] V. Mitra, H. Franco, M. Graciarena, "Damped oscillator cepstral coefficients for robust speech recognition," Proc. of Interspeech, pp. 886–890, 2013.

[25] V. Mitra, H. Franco, M. Graciarena, A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," Proc. of ICASSP, pp. 4117–4120, 2012.

[26] V. Mitra, M. McLaren, H. Franco, M. Graciarena, N. Scheffer, "Modulation features for noise robust speaker identification," Proc. of Interspeech, pp. 3703–3707, 2013.

[27] V. Mitra, H. Franco, M. Graciarena, D. Vergyri, "Medium duration modulation cepstral feature for robust speech recognition," Proc. of ICASSP, pp. 1768–1772, 2014.

[28] E. Shriberg, A. Stolcke, S. Ravuri, "Addressee detection for dialog systems using temporal and spectral dimensions of speaking style," Proc. of Interspeech, 2013.

[29] P. Boersma, D. Weenink, "Praat: Doing phonetics by computer," Version 5.1.05, url: http://www.praat.org/, 2009.

[30] H. Nam, L. Goldstein, E. Saltzman, D. Byrd, "TADA: An enhanced, portable task dynamics model in Matlab," J. of Acoust. Soc. Am., 115(5), pp. 2430, 2004.

[31] P.J. Moreno, P.P. Ho, and N. Vasconcelos "A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications," In Advances in Neural Information Processing Systems 16, Cambridge, MA, 2004. MIT Press.

[32] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, M. Gracierana, "The SRI AVEC-2014 evaluation system," Proc. of AVEC2014, 2014.

[33] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. on Speech and Audio Processing, 19, pp. 788–798, 2011.