NOISE AND REVERBERATION EFFECTS ON DEPRESSION DETECTION FROM SPEECH

Vikramjit Mitra, Andreas Tsiartas, Elizabeth Shriberg

SRI International, Menlo Park, CA, USA.

{vikramjit.mitra, andreas.tsiartas, elizabeth.shriberg}@sri.com

ABSTRACT

Speech-based depression detection has gained importance in recent years, but most research has used relatively quiet conditions or examined a single corpus per study. Little is thus known about the robustness of speech cues in the wild. This study compares the effect of noise and reverberation on depression prediction using 1) standard mel-frequency cepstral coefficients (MFCCs), and 2) features designed for noise robustness, damped oscillator cepstral coefficients (DOCCs). Data come from the 2014 Audio-Visual Emotion Recognition Challenge (AVEC). Results using additive noise and reverberation reveal a consistent pattern of findings for multiple evaluation metrics under both matched and mismatched conditions. First and most notably: standard MFCC features suffer dramatically under test/train mismatch for both noise and reverberation; DOCC features are far more robust. Second, including higher-order cepstral coefficients is generally beneficial. Third, artificial neural networks tend to outperform support vector regression. Fourth, spontaneous speech appears to offer better robustness than read speech. Finally, a cross-corpus (and cross-language) experiment reveals better noise and reverberation robustness for DOCCs than for MFCCs. Implications and future directions for real-world robust depression detection are discussed.

Index Terms—*depression detection, noise robustness, reverberation robustness, cross-corpus, mental health, AVEC Challenge, spontaneous speech, read speech, MFCC, DOCC*

1. INTRODUCTION

Speech has become an important source for health and mental health assessment and monitoring, as it can be obtained and analyzed in a noninvasive, natural, and inexpensive manner. Speech has been shown to reflect a range of speaker state information via features at various stages of the production process [1-14]. Current clinical analysis of depression is predominantly interview based, where such assessments provide an objective score for each patient [1], based on which further diagnosis and treatment are carried out. Such subjective clinical assessments are both labor- and time-intensive. Automatic detection of depression can help medical practitioners monitor changes in depression status, and prioritize follow-up with clinicians.

Recent studies [15, 16] have claimed that the speech of subjects suffering from Major Depressive Disorder (MDD) shifts compared to non-MDD subjects, indicating that speech can be a useful source for extracting bio-signatures of MDD. Several researchers have explored detecting MDD bio-signatures from speech. A wide array of features has been explored in the literature, in particular standard mel-cepstral features (MFCCs) [17, 18]; prosodic features (such as pitch, energy, and speaking rate, etc.) [19, 20, 21] and traditional speech-property-based features, such as formants, formant bandwidths, spectral energies, spectral tilt, etc. [18, 19, 20, 21, 22]. Correlation-structure features have been proposed by [23], and demonstrated impressive MDD detection accuracy on speech. Other studies [24] have used the popular MFCC features along with their velocity and acceleration coefficients.

Studies have also used both audio and video modalities [24, 25] for MDD detection. Work in [25] demonstrated that using both audio and video modalities improves MDD detection accuracy compared to using each modality alone. Yet speech is often easier to record and archive compared to video, and is also expected to be more invariant. It is also typically understood to be more private. Hence, speech-based MDD detection strategies can be expected to be cheaper in cost and relatively easier to prototype. Research in [24] demonstrated that audio data gave slightly better results than video data; however, other studies [26, 27] have shown the reverse.

Though recent results indicate the feasibility of automatic prediction of depression from speech, how such automated methods would perform under unseen acoustic background conditions is not yet fully explored. Most studies use data collected in laboratory setups with little or no background noise or reverberation. The effect of noise on emotion prediction has been explored in [41, 42, 43], but little work has been done to study the effect of noise and/or reverberation on depression data.

Cross-corpus experiments [40, 32] can serve as an evaluation of how different techniques generalize across corpora and such experiments reveal how systems behave under mismatched training-testing conditions. Environmental degradations in cross-corpus evaluations can potentially increase the difficulty of the prediction task. However, for the medical domain, and depression data in particular, only limited evaluation of the robustness of different algorithms across corpora has occurred, mainly due to medical data's limited availability and regulatory restrictions.

The effects of noise and/or reverberation on the performance of many speech-based applications (such as automatic speech recognition, speaker recognition, language identification, etc.) have been explored, but these effects are largely unexplored for speech-based automated analysis of mental health. Automated methods are expected to be exposed to varying environmental conditions, hence assessing how such methods would behave under varying background conditions is imperative.

In this work, we assess the robustness of an automated depression-prediction system by exploring robust acoustic features known to perform well in speech recognition tasks under noisy [28] and reverberant [29, 30] background conditions. Through performance analysis, we demonstrate that a combination of a robust acoustic feature and a suitable machine-learning algorithm achieves robust performance. Further, we show that under noisy and reverberated conditions, performance from spontaneous speech is almost always tends to be better than that of read speech, which is in line with some [31] but not all prior observation. Further, we explore how such systems perform for a totally different corpus, recorded in a different language, using an entirely different scoring scheme.

2. DATA

The dataset used in this work is an audio-visual depression corpus [27] distributed with the AVEC-2014 baseline system. It contains 300 videos of subjects (one subject per recording). This dataset includes 84 subjects, with some recorded more than once. The duration of each recording ranges from 20 to 50 min with an average duration of 25 min. The total duration of all clips is 240 hrs. The recordings took place in a number of quiet settings; however, they contain some ambient noise, reverberation and distortions introduced by the environment. The recordings consist of two sub-tasks: Northwind (read speech) and Freeform (spontaneous speech), which were supplied as 300 (2x150) audio-video files.

The dataset was split into three non-overlapping partitions of training, development, and test sets with 50 Northwind-Freeform pairs in each set, for a total of 300 task recordings. The training, development, and test sets had similar distributions in terms of age, gender, and depression levels. There was no session overlap between partitions; however, we did notice some speaker overlap across the partitions. The depression scores for the training and development set were distributed to the challenge participants by the organizers. The test set scores were not provided; please note that in this work, we present only results on the development set. The depression scores provided with the AVEC-2014 dataset consist of an individual's self-reported depression levels specified according to the Beck depression rating scale [39].

For cross-corpus analysis, we used data collected at Vanderbilt University (VU), 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. The details about the dataset can be obtained from [34]. Note that the clinicians used the HAM-D depression score to rate each subject in this dataset. The HAM-D depression score is the most widely used clinician-administered depression-assessment scale and was designed for use after an unstructured clinical interview.

We artificially corrupted the audio data with noise and reverberation. Fourteen different types of noise were used, including factory, babble, traffic, highway traffic, mixed crowd, city traffic, etc. The AVEC-2014 dataset was corrupted at three signal-to-noise ratios (SNRs): 20 dB, 10 dB, and 5 dB, whereas the VU-PTU dataset was corrupted at 15 dB and 5 dB.

Reverberation was added by using the setup distributed through the 2014 REVERB challenge [33]. For reverberation, twelve different room conditions were used, where the room types were small, medium, and large, each having two different room-impulse responses (RIRs) with two different microphone positions. Note that a distant microphone setup [33] was used in this study to make the problem of reverberation more adverse. In addition to reverberation, around 40dB of ambient room noise was added to simulate more realistic recording conditions. Note that the ambient noise SNR was uniformly between 20dB to 60dB.

3. FEATURES

As a baseline, we used mel-frequency cepstral coefficients (MFCCs), which tend to be used in most work on depression as well as emotion detection. MFCCs were computed using 24 mel-filterbanks and N-cepstral (N was varied from 13 to 24) coefficients that were concatenated with the energy coefficient.

As a comparison feature set we used Damped Oscillator Cepstral Coefficients (DOCC) [34]. This is a robust acoustic feature set that has demonstrated robustness in speech recognition under both noisy [35] and reverberated [36] conditions. DOCCs aim to model the dynamics of the hair cells within the human ear and have a longer-term memory than the MFCCs. 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. Inherently, the DOCCs perform a long-term filtering of bandlimited time domain signals and can filter our narrow-band noise and late-reverberation effects. Our studies on automatic speech recognition tasks have shown that DOCCs convincingly performs better than MFCC for both noisy and reverberated conditions, and this is the reason for using this feature in our experiments reported in this paper.

The acoustic features were mean- and variance-normalized on a per-subject basis. We create a fixed length representation in the form of i-vectors (similar to our work in [37, 31]) for each conversation channel. The i-vector subspace had 30 dimensions; final i-vectors were length normalized before being used by the classifiers.

4. DEPRESSION-SCORE PREDICTION MODEL

For our initial experiments, we used support vector regression (SVR) [38] for predicting depression scores (in the Beck depression rating scale) from speech. The SVR training was performed by using the scikit-learn Python package; where the SVR had a polynomial kernel of order 20. Our initial exploration with different SVR kernels [37] revealed that the polynomial kernel was the optimal kernel for the given task, and hence we used it as the default kernel for all reported experiments.

In addition to SVRs, we trained separate artificial neural networks (ANNs) for each feature type and training condition. The nets were trained using back-propagation with a scaled conjugate gradient algorithm, where the inputs were the 30D i-vectors, and the targets were the Beck depression rating scores. Note that the ANNs had linear activation for the input and output layers, with tan-sigmoid activation between the hidden layers. The performance of the ANNs was evaluated with Pearson's product moment correlation (PPMC) coefficient, mean absolute error (MAE), and root mean squared error (RMSE), as these were the performance metrics used in AVEC-2014 [27].

For the cross-corpus analysis, we trained ANN models by using the AVEC-2014 training data, and then employed the trained models to predict the depression scores for the VU-PTU dataset. We report the PPMC between the model-predicted depression scores and the HAM-D depression scores of the VU-PTU dataset.

5. RESULTS AND DISCUSSION

We examined effects of the conditions in Table 1 for both MFCCs and DOCCs, in each case finding an optimal feature size. Results using the SVR system are provided in Table 2. As shown, MFCCs with 17 cepstral features and DOCCs with 20 cepstral features gave better performance amongst the different cepstral dimensions explored in this work. Table 2 also shows that trend is somewhat noisy, which may have been due to the limited data size of the AVEC-2014 corpus. Note that the higher cepstral features capture source information (i.e., information relevant to the speaker's

speech production system), whereas the lower cepstral features typically capture linguistic information. Interestingly, the Table 2 results may indicate that higher cepstral coefficients help to obtain better depression-prediction performance as depression impacts the speaker's production mechanism. Based on the Table 2 observations, we use 17 cepstral features for MFCCs (MFCC17) and 20 cepstral features for DOCCs (DOCC20) in all following experiments.

Table 1. Train/test conditions. X = 20, 10, 5.

Condition	Training	Testing
AVEC	original	original
AVEC-mismatch-[X]dB	original	noisy@[X]dB
AVEC-mismatch-reverb	original	reverberated
AVEC-match-[X]dB	noisy@[X]dB	noisy@[X]dB
AVEC-match-reverb	reverberated	reverberated

Table 2. Depression-prediction performance for AVEC development data using MFCC and DOCC with different cepstral dimensions using SVR model.

Feature Name	#Cepstra	MAE	RMSE	r_{PPMC}
MFCC	13	8.65	10.97	0.44
MFCC	**17**	**8.48**	**10.17**	**0.54**
MFCC	20	8.83	10.82	0.41
MFCC	24	9.05	10.94	0.42
DOCC	13	7.74	9.31	0.64
DOCC	17	7.87	9.85	0.58
DOCC	**20**	**7.75**	**9.27**	**0.67**
DOCC	24	7.70	9.53	0.63

Next, we performed mismatched and matched train-test evaluations on the AVEC data using MFCC17 and DOCC20. Figure 1 shows the line-plot of the PPMC obtained from matched and mismatched training-testing conditions at different SNR levels using the SVR model.

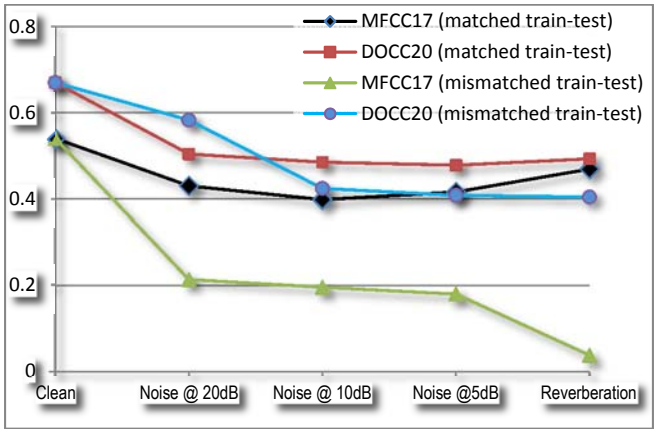

Figure 1. PPMC (correlation) for different training-testing conditions with SVR models using MFCC17 and DOCC20 feature. The vertical axis represents PPMC; the horizontal axis represents different environmental conditions.

In addition to SVRs, we also explored ANN models. ANNs can model nonlinearity quite well and can perform complex functional mappings with a high degree of accuracy. Tables 3 and 4 show the results obtained from the mismatched and matched evaluations using the SVR and ANN models.

Tables 3 and 4 share several pieces of interesting information. First, the matched train-test conditions always give better performance over mismatched train-test conditions. For mismatched train-test conditions, the PPMC for MFCC goes down significantly compared to that in Table 2. Table 3 shows the impact of unseen environmental conditions, indicating how susceptible automated speech applications are to noise and reverberation. The mismatched reverberation condition seemed to be most detrimental for MFCCs, where the PPMC suffered an almost a 93% decrease, and the RMSE increased by 30% compared to the clean baseline in Table 2. Importantly, DOCCs did not demonstrate such catastrophic degradation in performance, despite witnessing a fall in PPMC and a rise in the MAE and RMSE scores. For both MFCC and DOCCs, the mismatched reverberation condition seemed to be the most challenging one. Table 4, on the contrary, to the mismatched results in Table 3, shows that both DOCCs and MFCCs perform reasonably well under matched training-testing conditions. MFCCs in the matched condition performed much better, demonstrating ~ 20% reduction in PPMC and ~ 11% increase in RMSE. Even in the matched condition, DOCCs performed better than the MFCCs, showing both higher PPMC and lower RMSE and MAE than the latter.

Table 3. Depression-prediction performance for AVEC mismatched train-test conditions using MFCC and DOCC features with the SVR and ANN models.

	Condition	MAE		RMSE		r_{PPMC}	
		SVR	ANN	SVR	ANN	SVR	ANN
MFCC17	Clean	8.48	**8.43**	10.47	**10.14**	0.54	**0.56**
	AVEC-mismatch-20dB	**9.60**	11.33	**12.36**	14.02	0.21	**0.23**
	AVEC-mismatch-10dB	**10.26**	10.92	**12.41**	12.95	**0.20**	0.19
	AVEC-mismatch-5dB	**10.44**	11.05	**12.53**	13.54	**0.18**	0.07
	AVEC-mismatch-reverb	**11.60**	12.12	**13.27**	14.34	0.04	**0.08**
DOCC20	Clean	7.75	**7.11**	9.27	**8.67**	0.67	**0.69**
	AVEC-mismatch-20dB	8.28	**7.76**	10.23	**9.50**	0.58	**0.62**
	AVEC-mismatch-10dB	**8.66**	**8.66**	**11.26**	11.46	0.42	**0.47**
	AVEC-mismatch-5dB	9.83	**9.48**	**11.85**	11.89	0.41	**0.43**
	AVEC-mismatch-reverb	9.30	**9.00**	11.47	**11.19**	0.40	**0.45**

Table 4. Depression-prediction performance for AVEC matched train-test conditions using MFCC and DOCC features with the SVR and ANN models.

	Condition	MAE		RMSE		r_{PPMC}	
		SVR	ANN	SVR	ANN	SVR	ANN
MFCC17	Clean	8.48	**8.43**	10.47	**10.14**	0.54	**0.56**
	AVEC-mismatch-20dB	8.90	**8.76**	11.29	**10.80**	0.43	**0.50**
	AVEC-mismatch-10dB	8.62	**8.46**	11.30	**11.04**	0.40	**0.45**
	AVEC-mismatch-5dB	9.46	9.10	11.52	11.39	**0.42**	0.40
	AVEC-mismatch-reverb	**8.56**	8.90	**10.86**	11.17	**0.47**	**0.47**
DOCC20	Clean	7.75	**7.11**	9.27	**8.67**	0.67	**0.69**
	AVEC-mismatch-20dB	8.41	**8.27**	10.50	**10.09**	0.50	**0.56**
	AVEC-mismatch-10dB	**8.13**	8.29	10.57	**10.20**	0.49	**0.54**
	AVEC-mismatch-5dB	9.13	**8.29**	11.07	**10.26**	0.48	**0.53**
	AVEC-mismatch-reverb	8.63	**8.57**	10.68	10.47	0.49	**0.52**

The role of reverberation is noteworthy: in the mismatched condition, reverberation was detrimental for MFCCs; but for the

matched condition, its impact was smaller, and less severe than for noise. This finding suggests that reverberation may be effectively countered by properly training the models with reverberated data. Unlike noise, reverberation effects can be easily added to training data, as their characteristics are far less diverse than those of noise.

Tables 3 and 4 also show that DOCC20 performed reasonably well under all conditions, with its worst performance on the AVEC-mismatch-reverb and AVEC-mismatch-5dB conditions, where PPMC was reduced by ~ 40% and 39%, respectively, compared to baseline, i.e., the clean condition results. Finally, the performances of all the systems were found to degrade with higher levels of noise (i.e., with lower SNR values), which is typically expected.

For our experiments using ANN models have used single hidden layer neural nets with 700 neurons in the hidden layer. Typically, adding an extra layer or two results in improving the performance [31] with a chance of data over-fitting given the limited size of the training corpus. The ANN results are also provided in tables 3 and 4, which show their performance under mismatched and matched train-test conditions.

Comparing the results in tables 3 and 4 we can see that for most of the conditions, ANNs resulted in higher correlation score compared to its SVR counterpart, which is in line with our earlier observations in [31]. For DOCC features the ANNs in general gave better performance than SVRs, even for noisy and reverberated conditions and for MFCCs the trend was opposite. This may indicate that different features may behave differently with different modeling strategies and each such combination may capture complementary information that can potentially benefit late fusion of systems. In general, matched condition results are far better than the mismatched ones, indicating that the magnitude of the mismatch between training-testing conditions effect the performance of the models where the performance degradation is somewhat proportional to the degree of mismatch. We also observed that the DOCCs performed better than MFCCs.

We also analyzed performance of the matched and mismatched train-test datasets using ANN models on the spontaneous- and read-speech parts of the AVEC dev data, with the results given in Table 5. Table 5 indicates that even under noisy and reverberated conditions, spontaneous speech gave better performance compared to read speech. Note that the models were trained with both spontaneous and read speech, as we observed performance degradation when the models were trained with either of them separately. That finding may be due to having less data for model training when selecting one part over the other.

Next, we explore how the systems behave in an entirely mismatched scenario, where the recording conditions, language and content are different between the training and testing data.

Table 5. Performance differences between read and spontaneous speech for DOCC20 features using ANN models.

Condition	MAE		RMSE		r_{PPMC}	
	Read	Spont.	Read	Spont.	Read	Spont.
AVEC-mismatch-20dB	8.03	**7.50**	9.82	**9.17**	0.61	**0.65**
AVEC-mismatch-10dB	8.71	**8.61**	12.04	**10.84**	0.46	**0.50**
AVEC-mismatch-5dB	10.82	**8.15**	13.16	**10.46**	0.33	**0.55**
AVEC-mismatch-reverb	9.19	**8.82**	11.76	**10.60**	0.42	**0.50**
AVEC-match-20dB	8.59	**7.96**	10.72	**9.41**	0.47	**0.64**
AVEC-match-10dB	8.49	**8.07**	10.60	**9.77**	0.49	**0.59**
AVEC-match-5dB	8.50	**8.15**	10.40	**9.98**	0.47	**0.52**
AVEC-match-reverb	9.01	**8.13**	10.60	**9.77**	0.49	**0.59**

For that purpose we set up a cross-corpus analysis, in which we trained ANN models using the AVEC training data (without noise or reverberation), and then used the noisy and reverberated VU-PTU data for evaluation. Note that the AVEC data contains speech spoken by German speakers in a residential setup, whereas the VU-PTU data contains speech by English speakers in a clinical facility. Table 6 shows the results from the cross-corpus analysis.

Table 6. Cross-corpus depression-prediction performance for ANN models trained with AVEC data and evaluated on VU-PTU data using MFCC and DOCC features.

	Test Condition	MAE	RMSE	r_{PPMC}
MFCC17	VU-PTU	8.688	10.123	0.177
	VU-PTU+noise @15dB	10.665	12.577	0.125
	VU-PTU+noise @5dB	9.592	11.450	0.112
	VU-PTU+reverb	11.110	13.205	0.049
DOCC20	VU-PTU	7.265	8.789	0.449
	VU-PTU+noise @15dB	9.102	11.646	0.186
	VU-PTU+noise @5dB	8.187	10.630	0.185
	VU-PTU+reverb	9.041	11.017	0.365

Note that the MAE and RMSE do not provide meaningful quantitative analysis regarding performance here, as the predicted depression scored from the ANN outputs (in Beck's scale) and the target depression scores of the VU-PTU data (HAM-D ratings) are quite different. However, the MAE and RMSE hint at how closely the predicted and target depression scores are from the cross-corpus analysis. Table 8 shows that DOCCs performed much better overall than MFCCs. Interestingly for the cross-corpus analysis, MFCCs did not perform as poorly as they did on the in-corpus analysis of the 5 dB noisy condition. As shown, DOCCs performed quite well under reverberated conditions, indicating the resilience of this feature set under reverberant distortion.

6. CONCLUSION

We reported a series of experiments for predicting depression scores from speech, using the AVEC-2014 corpus under matched and mismatched conditions, and including controlled noise and reverberation corruption at different levels. Results showed clearly that selecting robust features (in this case, DOCCs over the standard MFCCs) adds resilience to system performance. ANNs were found to be more robust than SVRs. In an analysis of speaking style we found that spontaneous speech gave better performance than read speech. This finding was observed even for noisy and reverberated conditions but deserves further study given differences in data sizes. Finally, we found that noise usually impacts cross-corpus performance more adversely than does reverberation. We demonstrated that using suitable robust features and modeling strategies mitigates performance degradation from varying background conditions. In future work, we intend to explore noise effects in naturalistic data, feature optimization, feature fusion, and adaptation techniques to find ways to improve robustness of automatic depression-prediction models.

7. REFERENCES

[1] 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. Elsevier, 2012, pp. 227–237.

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

Review.
[3] 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.
[4] M. Alpert, R. Kurtzberg, A. Friedhoff, “Transient voice changes associated with emotional stimuli,” *Arch. Gen. Psychiatry*, Vol. 8, pp. 362–365, 1963.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] J. Darby and H. Hollien, “Vocal and speech patterns of depressive patients,” *Folia phoniat*, vol. 29, pp. 279–291, 1977.
[14] J. Darby, N. Simons, and P. Berger, “Speech and voice parameters of depression: A pilot study,” *J. Commun. Disorders,* vol. 17, pp. 75–85, 1984.
[15] A. Ozdas, R. G. Shiavi, D. M. Wilkes, M. K. Silverman, and 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.
[16] 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,” *IEEE Transactions on Biomedical Engineering,* vol. 51, no. 9, pp. 1530–1540, September 2004.
[17] L. 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 *IEEE Conference on Acoustics, Speech, and Signal Processing,* Dallas, TX, USA, 2010, pp. 5154–5157.
[18] H. K. Keskinpala, T. Yingtha wornsuk, D. M. Wilkes, R. G. Shiavi, and R. M. Salomon, “Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands,” in *European Signal Processing Conference*, Poznan, Poland, 2007, pp. 2229–2233.
[19] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, “Acoustical properties of speech as indicators of depression and suicidal risk,” *IEEE Transactions on Biomedical Engineering,* vol. 47, no. 7, pp. 829–837, July 2000.
[20] E. M. II, 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*, vol. 55, no. 1, pp. 96–107, January 2008.
[21] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre, “Detecting depression from facial actions and vocal prosody,” in *International Conference on Affective Computing and Intelligent Interaction*, 2009.
[22] M. H. Sanchez, D. Vergyri, L. Ferrer, C. Richey, P. Garcia, B. Knoth, W. Jarrold, "Using prosodic and spectral features in detecting depression in elderly males," *Proc. of Interspeech*, 2011.
[23] J. R. Williamson, R. Horwitz, T.F. Quatieri, B. Yu, B. S. Helfer, D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," *Proc. of AVEC 2013*.
[24] N. Cummins, V. Sethu, J. Joshi, R. Goecke, A. Dhall, J. Epps "Diagnosis of depression by behavioural signals: A multimodal approach," *Proc. of AVEC 2013*.
[25] H. Meng, H. Wang, H. Yang, M. Al-Shuraifi, Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," *Proc. of AVEC 2013*.
[26] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, "AVEC 2013 – The continuous audio/visual emotion and depression recognition challenge," *Proc. of AVEC 2013*.
[27] 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 AVEC*, 2014
[28] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Graciarena, “Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions,” in *Proc. of Interspeech*, 2014.
[29] V. Mitra, W. Wang, Y. Lei, A. Kathol, G. Sivaraman, and C. Espy-Wilson, “Robust features and system fusion for reverberation-robust speech recognition,” in *Proc. of REVERB Challenge*, 2014.
[30] V. Mitra, J. Van Hout, M. McLaren, W. Wang, M. Graciarena, D. Vergyri, and H. Franco, “Combating reverberation in large vocabulary continuous speech recognition,” *Proc. of Interspeech*, 2015.
[31] V. Mitra and E. Shriberg, “Effects of feature type, learning algorithm and speaking style for depression detection from speech,” *Proc. of ICASSP*, pp. 4774–4778, 2015.
[32] V. Mitra, E. Shriberg, D. Vergyri, B. Knoth and R.M. Salomon, “Cross-corpus depression prediction from speech,” in *Proc. of ICASSP*, pp. 4769-4773, 2015.
[33] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, “The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* 2013.
[34] V. Mitra, H. Franco, M. Graciarena, “Damped oscillator cepstral coefficients for robust speech recognition,” *Proc. of Interspeech,* pp. 886–890, 2013.
[35] V. Mitra and H. Franco, “Time-frequency convolution networks for robust speech recognition,” in *Proc. of ASRU 2015*.
[36] V. Mitra, J. Van Hout, W. Wang, M. Graciarena, M. McLaren, H. Franco and D. Vergyri, “Improving robustness against reverberation for automatic speech recognition,” in *Proc. of ASRU 2015*.
[37] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri and M. Gracierana, “The SRI AVEC-2014 evaluation system,” *4th International Audio/Visual Emotion Challenge and Workshop, ACM Multimedia*, 2014.
[38] H. Drucker, C.J. Burges, L. Kaufman, A. Smola, V. Vapnik, "Support vector regression machines. Advances in neural information processing systems," 9, pp. 155–161, 1997.
[39] 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):588{97, December 1996.
[40] Z. Zhang, F. Weninger, M. Wo ̈llmer, and B. Schuller, “Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Big Island, HY, USA, 2011, pp. 523–528
[41] Huang CW, Chen GM, Yu H, Bao YQ, Zhao L. “Speech Emotion Recognition under White Noise”. Archives of Acoustics 2013; 38 (4): 457-463.
[42] Schuller, Björn, Dejan Arsic, Frank Wallhoff, and Gerhard Rigoll. "Emotion recognition in the noise applying large acoustic feature sets." *Speech Prosody, Dresden* (2006): 276-289.
[43] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, “Emotion recognition from noisy speech,” in *Proc. ICASSP*, Toulouse, France, May 14–19 2006
[44] A. Tawari and M. Trivedi, “Speech emotion analysis in noisy real-world environment,” in *Proc. Int. Conf. Pat- tern Recognition*, Istanbul, Turkey, Aug. 23–26 2010.