EFFECTS OF FEATURE TYPE, LEARNING ALGORITHM AND SPEAKING STYLE FOR DEPRESSION DETECTION FROM SPEECH

Vikramjit Mitra, Elizabeth Shriberg

SRI International, Menlo Park, CA, USA.
{vikramjit.mitra, elizabeth.shriberg}@sri.com

ABSTRACT

Computational methods for speech-based detection of depression are still relatively new, and have focused on either a standard set of features or on specific additional approaches. We systematically study the effects of feature type, machine learning approach, and speaking style (read versus spontaneous) on depression prediction in the AVEC-2014 evaluation corpus, using features related to speech production, perception, acoustic phonetics, and prosody. Using a multilayer ANN we find that one feature type, MMEDuSA [2], results in a 25% relative error reduction over the AVEC-2014 baseline system [1] for both mean absolute error (MAE) and root mean squared error (RMSE). Other individual feature types perform comparably to the baseline, but have much lower dimensionality and simpler to interpret. Further improvements were achieved from fusing diverse features and systems. Finally, results suggest that the relative contribution of different feature types depends on whether the speech is spontaneous or read. Overall, spontaneous speech led to lower error rates than read speech, an important consideration for the collection of future clinical data.

Index Terms— Depression detection, robust signal analysis, acoustic features, articulatory features, prosody, neural networks, clinical data

1. INTRODUCTION

Depression affects a significant portion of the human population and can often be life-threatening [3, 4]. Early detection, evaluation and treatment [5, 6] can help to reduce depressive symptoms and improve the quality of human life. Currently depressive symptoms are mainly assessed through subjective evaluations that require expert clinician intervention. Such clinical assessments provide an objective score to each patient [7], upon which further diagnosis and treatment are based. Such subjective clinical assessments are both labor- and time-intensive. Automatic detection of depression can help medical practitioners monitor for changes in depression status, and then prioritize follow-up with clinicians.

In [8, 9] researchers have analyzed acoustic "bio-signatures" of major depressive disorder (MDD) before and after treatment with antidepressant medication; other studies [10, 11] have noted that the speech of subjects suffering from MDD shifts compared to non-MDD subjects. Detection of bio-signatures of MDD from speech has been explored by several researchers in the past. A wide array of features has been explored in the literature, in particular standard mel-cepstral features (MFCCs) [12, 13], prosodic features (such as pitch, energy, and speaking rate, etc.) [14, 15, 16], and traditional speech property based features such as formants, formant bandwidths, spectral energies, spectral tilt, etc. [13, 14, 15, 16, 17]. Correlation structure features have been proposed recently [18], and demonstrated impressive MDD detection accuracy. Other studies [19] have used traditional MFCC features along with their velocity and acceleration coefficients. Studies have also used both audio and video modalities [19, 20] for MDD detection and demonstrated impressive accuracies. It was demonstrated in [20] that use of both audio and video modalities improves the accuracy of a MDD detection compared to using each modality alone. Speech is often easier to record and archive compared to video and is also expected to be more invariant. Hence speech-based MDD detection strategies can be expected to be cheaper in cost and relatively easier to prototype. In [19] the authors demonstrated that the audio data can give slightly better results than video; however, other studies [1, 21] have shown the reverse.

In this work we conduct a systematic study that varies feature type, machine leaning approach, and the speaking task (read versus spontaneous). Our goal is to *better understand how these factors interact to influence automatic detection performance*. We present an audio-based neural network (NN) model for depression score prediction, where the scores are based on an individual's self-reported depression levels specified according to the Beck depression rating scale [22]. We analyze a wide array of speech-based features and evaluate their performance on the 2014 Audio-Visual Emotion recognition Challenge (AVEC) [1].

2. DATA

The AVEC-2014 dataset is an audio-visual depression corpus [1] that contains 300 videos of subjects (one subject per recording) recorded by a webcam and a microphone. This dataset includes 84 subjects, with some subjects recorded more than once: 18 subjects appear in three recordings, 31 in two, and the remaining 34 in only one recording. The duration of each recording ranged 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 recordings took place in a number of quiet settings; however, we observed some ambient noise, reverberation and distortions introduced by the background into the audio recordings.

The recordings consisted of speech that were spoken out loud while solving a task: counting numbers from one to ten; reading out loud; singing; telling a story from the subject's own past; and telling an imagined story. The recordings in the AVEC-2014 challenge subset consist of only two tasks: Northwind (Read speech) and Freeform (spontaneous speech), which were supplied as separate recordings, resulting in a total of 300 (2x150) videos. The set of source videos is largely the same as that used for the AVEC-2013 challenge; however, five pairs of previously unseen

recordings were used by the organizers to replace a small number of videos used in the 2013 challenge.

We resampled all data to 16 kHz. The challenge data was split into three 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. 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, hence in this work we present only the results from the development set. The performance metrics used for the AVEC-2014 challenge were mean absolute error (MAE) and root mean squared error (RMSE).

3. FEATURES

We explored a diverse set of features that operate at multiple time scales: some are frame-level low-level descriptors, while others are speech segment-level global descriptors. The features also represent a broad range of complexities with respect to processing prior to final feature dimensionality. We restricted ourselves to automatically extractable features that do not rely on words for two reasons: privacy and practicality. Word features also require speech recognition, which may or may not be available at high enough performance levels for a particular individual or context. We explored three broad classes of features: (1) ***spectral features (SFs)***, (2) ***articulatory and phonetic features (APFs)*** and (3) ***prosodic features (PFs)***.

Spectral features (SFs) include Gammatone Cepstral Coefficients (GCC). These features use a bank of gammatone filters to analyze speech; the resulting bandlimited signal within an analysis window of 26 ms is used to produce the gammatone spectra. A Discrete Cosine Transform (DCT) of the gammatone spectra is used to generate the GCC feature vector.

Damped Oscillator Cepstral Coefficients (DOCC) [23] use a biologically plausible model of auditory hair cells that tries to capture the perceptually relevant information from audio. In DOCC processing the incoming speech signal is analyzed by a bank of gammatone filters (in this work, we used a bank of 40 gammatone filters equally spaced on the equivalent rectangular bandwidth (ERB) scale), which splits the signal into band-limited subband signals. These subband signals are used as the forcing functions to an array of forced damped oscillators whose response is used as the acoustic feature.

Normalized Modulation Cepstral Coefficients (NMCC) [24] are perceptually-motivated noise-robust acoustic features that track the amplitude modulation (AM) of subband speech signals. The AM trajectories of the subband speech signals are used to generate a modulation spectrum, whose cepstral information is used as the NMCC feature set.

Modulation of Medium Duration Speech Amplitudes (MMeDuSA) [2, 25] tracks the subband AM signals of speech, and uses a medium duration analysis window (~52 ms). It also captures the overall summary modulation, which helps in tracking speech activity and detecting vowel prominence/stress.

Articulatory and Phonetic features ***(APFs)*** include that capture manner and place of articulation as well as information pertaining to phonetic correlates of the speech acoustics. Articulatory Features (AFs) [26, 27] track the configuration of the speech production system, i.e., the vocal tract, dynamically in time. As depression affects a speaker's production system AFs can potentially capture relevant signatures of depression from speech. In this work, we used a deep neural network (DNN) [26] to estimate the AFs in the form of vocal tract constriction variables (aka TVs) from speech. The AFs are an 8 dimensional feature and capture (1) Lip Aperture; (2) Lip Protrusion; (3) Tongue Tip Constriction Degree; (4) Tongue Tip Constriction Location; (5) Tongue Body Constriction Degree; (6) Tongue Body Constriction Location; (7) Velic Opening; and (8) Glottal Opening.

Acoustic Phonetic (AP) features [28] represent acoustic phonetic information and are analyzed using a 10 ms window with 5 ms frame rate. For the experiments reported in this paper 13 APs were used that represent information such as reflection coefficients, Hilbert envelope, periodic energy, aperiodic energy [29], nasal energy [30], etc. These provide information regarding voice quality, energy contour, etc., which can potentially act as biomarkers of depression in speech.

Prosodic features (PFs) capture information relevant to the prosodic structure of speech and hence lack the fine-grained information of the spectral features. Pitch tracks (KaldiF0) [31] were obtained using the Kaldi pitch recognition toolkit [32] where the output contains a 2-D information of pitch tracks and a normalized cross-correlation function that gives indication of voicing. Depression usually results in speech with lower pitch dynamics.

Energy contour features (encon) [33] aim to capture rhythmicity as well as overall speaking rate (without relying on phone recognition) by looking at the periodicity of energy peaks. The motivation behind using encon for the proposed task in this paper is that depressed speech may have a slower overall rate and be more temporally monotonous. This feature models the contour of 10 ms windows of the first two coefficients (c_0 and c_1) from an MFCC front end; each cepstral stream is mean-normalized over the utterance, making it robust to absolute level differences over both entire sessions and within-session segments. DCT is then taken over a 200 ms sliding window with a 100 ms shift. Vector components comprise the first five and two bases from the DCT over each window of c_0 and c_1, respectively.

Spectral tilt (tilt) features capture vocal effort in a manner somewhat robust to extrinsic session variability, using methods developed in [33]. These features were extracted for voiced frames. Voicing was determined using a logistic regression classifier using number of zero crossings, log energy, number of peaks in the autocorrelation of the window signal, and standard deviation of the inter-peak distance, where the voicing threshold was set to 0.5. The five component spectral tilt features include H2-H1, F1-H1, and F2-H1 (where H1, H2 are the lower-order harmonics and F1, F2 are the first two formants), 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 of the difference between the maximum of the log power spectrum and the maximum in the 2 kHz to 3 kHz range.

F0 peaks (f0peaks) is an intonation-related feature that uses the pitch track to obtain pitch peak distributions, and various statistics on the location of pitch peaks relative to each other and to segment boundaries. The motivation behind this feature is that if depressed speakers sound less animated, this should result in fewer peaks spaced more widely apart (a measure of speaking rate) and the peaks may be less extreme than for non-depressed speakers. More details are in [34]

Note that encon, tilt and f0peaks are highly sparse features that capture very specific information at certain temporal locations of speech. Apart from the features detailed above, we also explored standard MFCC features. Table 1 provides a summary of the different features. The acoustic features (DOCC, MMeDuSA, GCC, NMCC and MFCC), AFs, APs and KaldiF0 were mean- and variance-normalized on a per-subject basis. In our prior experiment [34] we found i-vectors [36] to be an effective representation of the acoustic features. To compensate for the limited amount of data available in the AVEC-2014 dataset we trained a small Universal Background Model (UBM), which had 16 Gaussian components and the i-vector subspace had only 30 dimensions. The i-vectors were length normalized before being fed to the ANNs. For the other features (tilt, encon and f0peaks) we obtained a fixed-length representation using statistics over the feature distributions (mean, variance, min, max) as well as statistics on distances between feature extraction regions (to capture durational characteristics).

Table 1. Summary of all the features explored in this study

Name	Type	Dimension
DOCC	acoustic	13
NMCC	acoustic	13
MMeDuSA	acoustic	16
MFCC	acoustic	13
GCC	acoustic	13
AF	articulatory	8
AP	Acoustic phonetic	12
tilt	vocal effort (sparse prosodic)	5
encon	rhythmicity (sparse prosodic)	7
KaldiF0	pitch (prosodic)	2
f0peaks	rhythmicity, rate, pitch (sparse prosodic)	9

4. DEPRESSION SCORE PREDICTION

The features detailed in section 3 were transformed to a fixed length representation before being input to the ANN based depression score model. The i-vector representations had 30 dimensions, whereas the others had the following: tilt = 24 dims; f0peaks = 27 dims and encon = 21 dims.

We trained a separate ANN for each of the feature types and explored optimizing the number of neurons in each layer by using a leave-one-out strategy. The nets were trained with greedy layer-wise learning, using back propagation with scaled conjugate gradient algorithm, where the inputs were the features 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, MAE and RMSE. All of the i-vector systems demonstrated best performance when 2 hidden layers were used, while the rest demonstrated better performance with only one hidden layer.

After all the individual feature-based ANNs were trained we performed m-way depression score fusion (where fusion was performed by simple (a) averaging the scores, or (b) taking the median of the scores) amongst all the subsystems.

5. RESULTS AND DISCUSSION

As mentioned earlier, the AVEC-2014 data contained two partitions: (1) read speech (identified as Northwind) and (2) spontaneous speech (identified as Freeform). For each of the ANNs, we obtained the performance metric over the whole development set, as well as over the read and spontaneous partitions individually as well. Table 2 presents the optimal ANN configurations for each of the features explored. In tables 3 a, b and c we present the obtained performance measures (r_{PPMC}, MAE and RMSE) from each of the systems for the whole, spontaneous and read part of the development data. Tables 3 a, b and c show that there is no single feature that performs the best across conditions. While DOCC overall performs better than other features for read speech, the same is true for MMeDuSA for spontaneous speech. Some features (e.g., encon, AF, MMeDuSA) performed better for spontaneous speech, whereas others such as the APs and f0peaks performed better for read speech. For read speech the acoustic phonetic information (such as formants, periodic/aperiodic energy etc.) may be easier to capture because of the well-behaved nature and more careful articulation of speech, which may have contributed to the better performance of the APs. Overall, results indicate that it is easier to capture depression from spontaneous speech rather than from read speech. This suggests that speakers suppress their state somewhat when reading, because of the irrelevant nature of the content or the attention to reading, or both.

Table 2. ANN configurations for each feature-based system.

Feature Name	ANN input dim.	# of Neurons	
		Layer-1	Layer-2
DOCC	30	700	300
NMCC	30	700	300
MMeDuSA	30	500	400
MFCC	30	900	200
GCC	30	700	300
AF	30	1000	500
AP	30	600	300
tilt	24	50	-
encon	21	300	-
KaldiF0	30	700	200
f0peaks	27	26	-

Table 3.a Depression prediction performance for the full development set.

Class	Feature Name	MAE	RMSE	r_{PPMC}
SF	DOCC	7.871	**9.433**	**0.628**
	NMCC	8.053	9.926	0.583
	MMeDuSA	**7.673**	9.656	0.600
	MFCC	8.120	10.104	0.567
	GCC	7.686	9.651	0.616
APF	AF	7.912	10.284	0.540
	AP	8.882	11.245	0.437
PF	KaldiF0	8.606	11.018	0.442
	tilt	8.938	10.829	0.424
	encon	10.251	12.581	0.330
	f0peaks	10.098	12.331	0.341

The finding that spontaneous speech is better than read speech for depression detection has important potential impact for clinical data collection. Clinical studies that include speech often use only read speech passages in their protocol – presumably for convenience, privacy, and cross-speaker control. What we find

suggests that it would be useful to also use spontaneous speech, modulo privacy issues. Note that because our features do not use any word information, once features are extracted one could preserve privacy by simply discarding the original signal.

Table 3.b Performance for the spontaneous portion of the development set.

Class	Feature Name	MAE	RMSE	r_{PPMC}
SF	DOCC	7.835	9.576	0.614
	NMCC	7.694	9.495	0.625
	MMeDuSA	**6.778**	**8.674**	**0.695**
	MFCC	7.989	10.000	0.582
	GCC	7.527	9.452	0.647
APF	AF	7.107	9.328	0.629
	AP	9.177	11.616	0.399
PF	KaldiF0	8.129	10.528	0.493
	tilt	8.152	9.806	0.571
	encon	9.136	11.030	0.482
	f0peaks	9.937	12.696	0.308

Table 3.c Performance for the read portion of the development set.

Class	Feature Name	MAE	RMSE	r_{PPMC}
SF	DOCC	7.906	**9.288**	**0.642**
	NMCC	8.412	10.339	0.542
	MMeDuSA	8.569	10.546	0.508
	MFCC	8.251	10.208	0.557
	GCC	**7.845**	9.847	0.587
APF	AF	8.716	11.159	0.460
	AP	8.586	10.861	0.478
PF	KaldiF0	9.083	11.487	0.397
	tilt	9.725	11.763	0.265
	encon	11.365	13.961	0.201
	f0peaks	10.260	11.954	0.376

In addition to the studies presented above, we performed m-way score-level fusion of multiple systems; results are shown in Table 4 where we can see that best m-way fusion from the ANN systems gave us 32% and 35% relative reduction in MAE and RMSE compared to the AVEC audio-only baseline, and 21% and 19% relative reduction in MAE and RMSE compared to the AVEC video-only baseline. We also observed some improvement in performance compared to our AVEC-2014 submission [34], where a relative reduction of 1.6% and 2.5% in MAE and RMSE and a relative increase of 1.7% in PPMC score were obtained. The systems that were selected for the best m-way fusion were NMCC, MMeDuSA, encon, DOCC, AF, KaldiF0 and AP. It was also observed that fusing the scores using the median rather than the average was more useful.

Table 4. Depression prediction performance from different systems

System	Mode	MAE	RMSE	r_{PPMC}
AVEC-2014 baseline	audio	8.93	11.52	-
AVEC-2014 baseline	video	7.58	9.31	-
SRI's submission to AVEC-2014	audio	6.10	7.71	0.778
m-way fusion of ANN	audio	6.00	7.52	0.791
Fusion of ANN + SRI's submitted system to AVEC-2014	audio	**5.87**	**7.37**	**0.800**

Finally we combined our AVEC-2014 submission with the systems developed in this work and observed further performance improvement, where 3.8% and 4.4% relative reduction in MAE and RMSE and a 2.8% improvement in PPMC was observed compared to our AVEC-2014 submission system. The systems that got selected for this combination were MMeDuSA and AF.

6. CONCLUSIONS AND FUTURE WORK

In a systematic study varying feature type, machine learning, and speaking style in an AVEC data set, we demonstrated using multiple features that ANNs can be used to predict depression levels from speech. Overall we found that most feature types perform better using spontaneous rather than read speech. This finding suggests that where feasible, clinical data collections should include spontaneous speech (rather than only read speech) in their protocols.

We also demonstrated that further performance improvement can be achieved by selecting read versus spontaneous speech depending upon which feature is used, and that ANNs are a good learning approach. A direct comparison with Support Vector Regression (SVR) results revealed that the ANNs are comparable and in fact sometimes better than the SVR for some features (e.g., AFs, APs, MMeDuSA etc.). In future work we plan to explore training separate models for read speech and spontaneous speech and performing both within-split and across-split detection experiments to understand how the features and systems behave under matched and mismatched spontaneous-read speech train-test conditions. We also plan to investigate the differences in depression recognition performance between read and spontaneous speech using additional speech data sets, to assess generalizability.

7. ACKNOWLEDGEMENTS

We are grateful to Mitchell McLaren, Martin Graciarena, Andreas Kathol and Colleen Richey for their help and suggestions. This research was partially supported by NSF Grant # IIS-1162046.

8. REFERENCES

[1] 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

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

[3] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision, Washington, DC, American Psychiatric Association, 2000.

[4] M.M.Weissman, S. Wolk, R.B. Goldstein, D. Moreau, P. Adams, S. Greenwald, C.M. Klier, N.D. Ryan, R.E. Dahl, P. Wichramaratne, "Depressed adolescents grown up," Journal of the American Medical Association, 1999; 281(18):1701–1713.

[5] J. March, S. Silva, S. Petrycki, J. Curry, K. Wells, J. Fairbank, B. Burns, M. Domino, S. McNulty, B. Vitiello, J. Severe, "Treatment for Adolescents with Depression Study (TADS) team. Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents with Depression Study (TADS) randomized controlled trial," Journal of the American Medical Association, 2004; 292(7):807–820.

[6] J.A. Bridge, S. Iyengar, C.B. Salary, R.P. Barbe, B. Birmaher, H.A. Pincus, L. Ren, D.A. Brent, "Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment, a meta-analysis of randomized controlled trials," Journal of the American Medical Association, 2007; 297(15):1683–1696.

[7] 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.
[8] J. Darby and H. Hollien, “Vocal and speech patterns of depressive patients,” Folia phoniat, vol. 29, pp. 279–291, 1977.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] E. M. II, M. A. Clements, J. W. Peifer, and L. Weisser, “Criticalanalysis 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.
[23] V. Mitra, H. Franco, M. Graciarena, “Damped Oscillator Cepstral Coefficients for Robust Speech Recognition,” Proc. of Interspeech, pp. 886–890, 2013.
[24] 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.
[25] V. Mitra, M. McLaren, H. Franco, M. Graciarena, N. Scheffer, “Modulation Features for Noise Robust Speaker Identification,” Proc. of Interspeech, pp. 3703–3707, 2013.
[26] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, “Articulatory features from deep neural networks and their role in speech recognition,” Proc. of ICASSP, pp.3041-3045, Florence, 2014.
[27] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, “Articulatory Information for Noise Robust Speech Recognition,” IEEE Trans. on ASLP, Vol. 19, Iss. 7, pp. 1913–1924, 2010.
[28] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks", PhD thesis, University of Maryland College Park, December 2004.
[29] O. Deshmukh, J. Singh, C. Espy-Wilson, “A novel method for computation of periodicity, aperiodicity and pitch of speech signals,” Proceedings of the 34th International Conference on Acoustics, Speech and Signal Processing, 17–21 May, Montreal, Canada, pp. 117–120, 2004.
[30] T. Pruthi, C. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization," Proc. of Interspeech, pp. 1925–1928, 2007.
[31] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal and S. Khudanpu "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition," in Proc. of ICASSP, 2014.
[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The Kaldi speech recognition toolkit," in Proc. ASRU, 2011.
[33] E. Shriberg, A. Stolcke, S. Ravuri, “Addressee Detection for Dialog Systems Using Temporal and Spectral Dimensions of Speaking Style,” Proc. of Interspeech, 2013.
[34] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, M. Gracierana, “The SRI AVEC-2014 evaluation system,” in Proc. of the 4th International Audio/Visual Emotion Challenge and Workshop, ACM Multimedia, 2014.
[35] P. Boersma, D. Weenink, “Praat: doing phonetics by computer,” Version 5.1.05, url: http://www.praat.org/, 2009
[36] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. on Speech and Audio Processing, 2011, 19, 788-798.