INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An Investigation of Emotional Speech in Depression Classification

Brian Stasak[1,2], *Julien Epps*[1,2], *Nicholas Cummins*[1] *and Roland Goecke*[3]

[1] School of Elec. Eng. & Telecomm., University of New South Wales, Sydney, Australia
[2] National Information Communications Technology (NICTA)
[3] Human-Centred Technology, University of Canberra, Canberra, Australia

brian.stasak@student.unsw.edu.au, j.epps@unsw.edu.au, n.p.cummins@unsw.edu.au, roland.goecke@ieee.org

Abstract

Assessing depression via speech characteristics is a growing area of interest in quantitative mental health research with a view to a clinical mental health assessment tool. As a mood disorder, depression induces changes in response to emotional stimuli, which motivates this investigation into the relationship between emotion and depression affected speech. This paper investigates how emotional information expressed in speech (i.e. arousal, valence, dominance) contributes to the classification of minimally depressed and moderately-severely depressed individuals. Experiments based on a subset of the AVEC 2014 database show that manual emotion ratings alone are discriminative of depression and combining rating-based emotion features with acoustic features improves classification between mild and severe depression. Emotion-based data selection is also shown to provide improvements in depression classification and a range of threshold methods are explored. Finally, the experiments presented demonstrate that automatically predicted emotion ratings can be incorporated into a fully automatic depression classification to produce a 5% accuracy improvement over an acoustic-only baseline system.

Index Terms: Depression Classification, Data Selection; Automatic Emotion Ratings.

1. Introduction

In recent years, the problem of automatically detecting and monitoring depression using behavioral signals, through paralinguistic speech cues, has gained considerable popularity [1]. Typical speech based depression detection systems in the literature focus on eliciting a paralinguistic marker directly from a speech signal [2-5]. Surprisingly, despite strong links between changes in continuous affective measures and depression [6-8], there is little research exploring the benefits features derived from affective scores could introduce to a speech based depression classification system.

Depression has been consistently linked with changes in speech motor control [1]. Therefore, it is unsurprising that both spectral and formant based features are shown to be useful when classifying either the presence or severity of depression. Cummins et al. [2], using modulation spectrum based features, achieved a 2-class classification accuracy of 67%. Similar results were recorded by pairing Mel Frequency Cepstral Coefficients (MFCC) with a Gaussian Mixture Model (GMM) based classifier [3]. Combining formant-based features with either a GMM or Support Vector Machine (SVM) classifier, Helfer et al. [4] reported 2-class accuracies between 70-86%. Similar results were demonstrated using TEO-based features with a GMM classifier [5].

Speech depression recognition using continuous affect scores has potential as a future clinical noninvasive mental health assessment tool, assisting in discerning between healthy individuals from others suffering from clinical depression [1]. Depression prediction results presented by Perez et al. [9] suggest that a small number of dimensional affect scores, provided by human-listeners can produce comparable performance with large brute-forced acoustic feature spaces when predicting a speaker's level of depression. However, this research neither provided individual details on ratings-based feature performances, nor explored using these ratings to investigate emotional information for the improvement of depression classification systems.

Arousal is described as a conscious affective experience based on a varied degree of subjective mental activation or interest. Increased arousal is represented by stronger acoustic formant intensity and fundamental frequency vocal tension, resulting in a perceivable rise in pitch. Individuals with depression often produce low levels of arousal which result in monotonous vocalizations that are characterized by a decrease in spectral and formant dynamics [1]; indicating speech affected by depression is potentially associated with low arousal levels [10].

Dominance is an individual's perceived assertiveness, authority, and/or aggressive vocal characteristics. A high degree of dominance is useful in parenting, emergency, or threatening situations. Individuals with depression use more perceptually submissive speech, and therefore often exhibit less dominance than healthy individuals [11].

Valence entails an individual's interpretation of pleasantness or unpleasantness. Contrary to arousal, valence has been shown to have a stronger correlation with the semantic context of what is spoken than how it is prosodically spoken [12, 13]. Depression suppresses inhibition of responses to negative-valence stimuli. Therefore, individuals with depression are perceived by others to have lower valence scores than in healthy populations [14].

Due to the aforesaid links between affective states and depression, emotion features derived from continuous affect rating measures are suggested in this paper to investigate depressed speech classification. This is achieved by selecting rating and speech segments specific to particular affect regions. Lastly, a fully automatic depression classification system based on insights from this investigation is considered.

http://dx.doi.org/10.21437/Interspeech.2016-867

2. Experiment Data

A subset of the *Audio-Visual Emotion Challenge* (AVEC) 2014 corpus was used for this research due to its continuous affect ratings, depression scores, feature sets, and use in the aforementioned speech depression classification studies [2-4, 9, 15]. AVEC 2014 includes studio-quality speech audio recorded from 84 male and female speakers. Each speaker undertook a *Beck's Depression Inventory-II* (BDI-II) questionnaire [16]. Supplied with the AVEC 2014 data were 2068 standard baseline acoustic features, including 78 *Low Level Descriptors* (LLD) and many functionals per speaker file. The affect ratings were gathered from 4-5 listeners individually and they continuously rated every such file for arousal, valence, and dominance.

All manually rated values were scaled between -1 and 1 on a per-rater basis. Individual ratings were compiled using a weighted average to produce continuous (30 per second) arousal, valence, and dominance gold-standard ratings per file. Only the 'freeform' speech was used in these experiments because previous AVEC 2014 and other speech depression classification research indicated better performance on natural speech [9, 15, 17-19]. For full details of the AVEC 2014 corpus the reader is referred to [15].

3. Experimental System Configurations

3.1 Depression Classification Using Speech/Ratings

In the following experiments, depression classification was favored over prediction as per [9, 15] because clearly defined classes potentially reveal more about the effects of depression on speech and continuous manual rating trends. Moreover, the ordinality of BDI scores complicates accurate prediction (i.e. a score of 10 is not necessarily twice as depressed as a score of 5). In Section 4, interaction between emotion and depression is investigated using combinations (concatenations) of these feature sets:

- Baseline features (2068 functionals total) – *Baseline Functionals* (BF) [15]
- Low Level Descriptors features (78 total) [15]
- Manually generated features (arousal, valence, and dominance gold-standard continuous ratings, aggregated on a per-file basis using means, medians, and standard deviations; 9 total) – *Manual Ratings* (MR)

Depression classification was performed by MATLAB 2016b software *Support Vector Machine* (SVM). SVM approaches have been shown to give good speech depression classification performance [1, 20]. All experiments utilized an SVM linear kernel with default hyper-parameter settings. Similarly to Cummins et al. [21], two-class labels were determined by assigning BDI 0-9 as 'none to low depression' and BDI 19-65 as 'moderate to high depression'. There were roughly 40 speakers per class and 100 training/development files[1]; file length varied from approximately 6 seconds to 4 minutes. Five-fold cross validation, performed using this 2-class division, was used in all classification tests, with an average accuracy reported.

The scope of the proposed depression classification experiments is shown in Figure 1. All experiments evaluated baseline, LLD, and manual ratings-based individual features along with various combinations. In addition, the effects of data selection were examined specifically for ratings-based and LLD features.

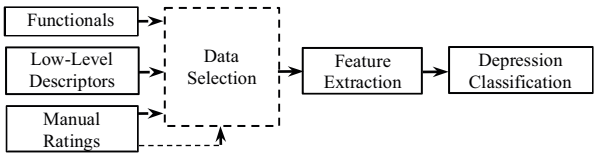

Figure 1: *System configuration for experiments applying arousal/valence/dominance thresholds as a means of data selection for depression classification. Dashed lines indicate components that may or may not be used in given experiments.*

3.2 Emotion-Based Data Selection

Based on the discussion in Section 1, it is reasonable to ask whether depression discrimination is improved by selecting segments of speech residing exclusively in particular emotional regions. Convenient means of selecting these emotional subsets is to threshold features based on the rated arousal, valence, or dominance per rating time increment. In order to select various mild and severe emotional subsets, four alternate threshold types were applied (Figure 2):

a) Upper Bounds: lower emotion values retained.
b) Lower Bounds: higher emotion values retained.
c) Extroverted Bounds: mild (center) emotion values retained.
d) Introverted Bounds: severe (fringe) emotion values retained.

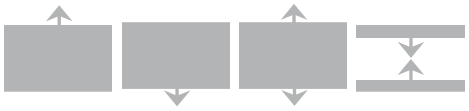

Figure 2: *From left to right, examples of (a) upper, (b) lower, (c) extroverted, and (d) introverted threshold types for emotion-based data selection. Gray indicates the retained ratings region.*

4. Results and Discussion

4.1 Effect of Manual Emotion Rating Information

As seen in Figure 3 (black dotted line 'All'), for standalone ratings-based features, arousal demonstrated better depression classification results than valence and dominance. Similarly, prior studies indicated a strong relationship between speech arousal and BDI prediction [15, 21]. Notably, three manual arousal ratings-based features achieved 70% performance accuracy alone, while 2068 acoustic baseline features attained 77.5% accuracy. This indicates that ratings-based features can compactly summarize some key depression information.

Manual ratings-based features concatenated with acoustic baseline features (Figure 3, black dashed line) consistently produced equal or better depression classification results over baseline functionals (arousal + valence + dominance same as standalone arousal result). Dominance ratings-based features performed the worst for depression classification. This was expected based on previous literature showing difficulty procuring above chance emotion classification results using 'freeform' speech dominance information [15].

[1] Class labels available per request

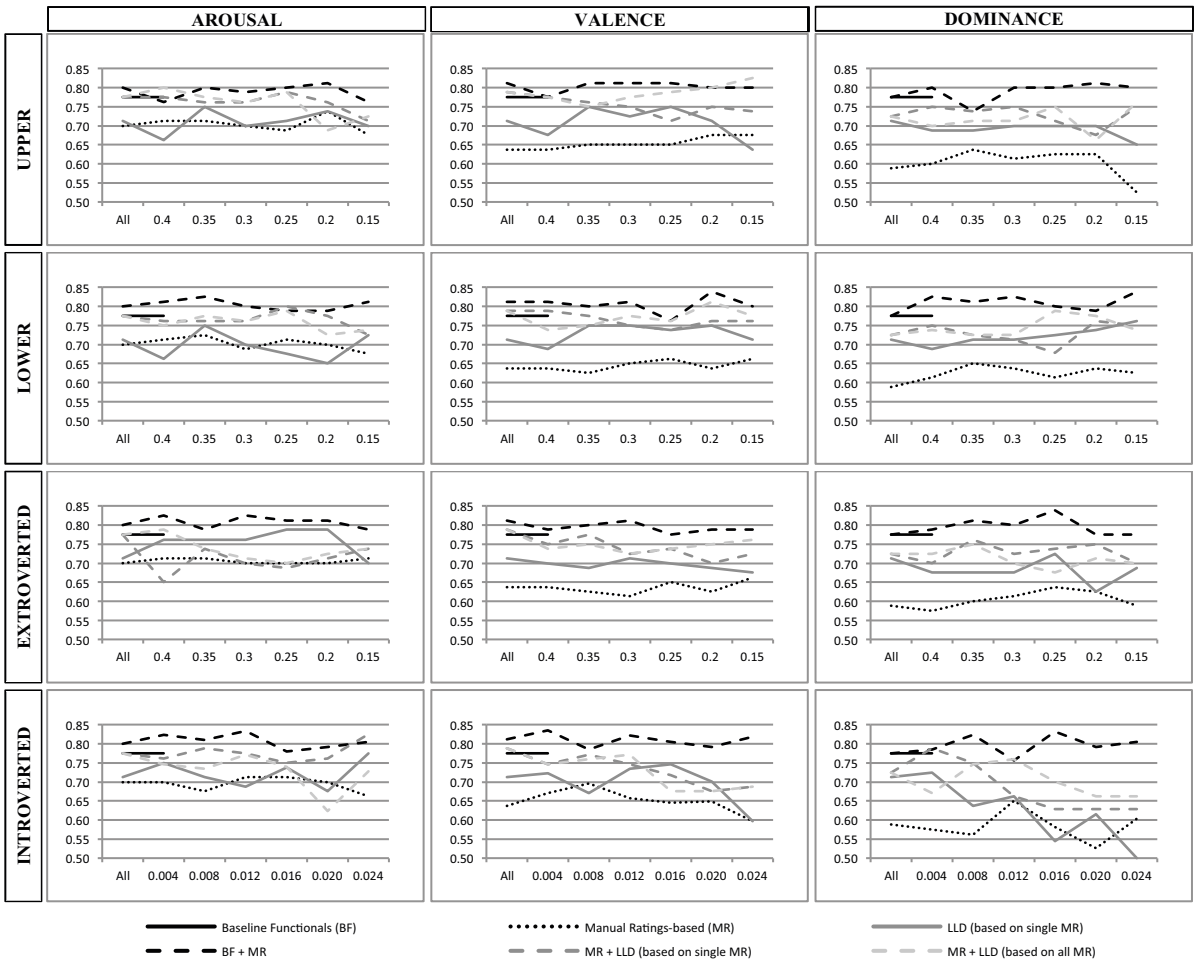

Figure 3: *Depression classification accuracies versus threshold value (four threshold types) for arousal, valence, and dominance ratings-based features, speech features, and a combination. The LLD based on single manual rating (MR) is the selection of any LLD feature frame whose corresponding manual rating was within the given data selection threshold. The LLD based on all MR selects only those LLD frames for which all three manual ratings were within the threshold.*

4.2 Effect of Emotional Thresholds

When compared with using *all* rating features, improvements in classification performance were recorded by applying *data selection* to standalone arousal, valence, and dominance ratings-based features (Figure 3, black dotted lines). Using manual ratings-based and baseline features, the upper, lower, and extroverted data selection with any threshold between 0.40 to 0.20 (retaining roughly 95% to 75% of the ratings) produced equal or better performance than baseline results. The introverted bounds when paired with baseline features did not exhibit wide threshold parameter ranges that consistently resulted in better classification performance. This suggests that discarding features (both rating-based *and* acoustic) corresponding to severe (fringe) emotions was helpful in general.

When identifying which acoustic frames to keep or drop, manual ratings were successfully used as a data selection criterion for acoustic features. Performance gains over LLD (all data retained) are noticeably observed for thresholds applied to ratings-based features and LLD based on all MR features (Figure 3, light gray dashed line). In many instances, these performed better than or equal to the baseline results. Thus, LLD frames for which all three manual ratings were within the threshold is more advantageous for performance gains than a single rating (Figure 3, light gray dashed line).

There was a sizeable performance impact when rating-based features were combined with a small set of acoustic LLD features (Figure 3, solid gray line versus light gray dashed line). Without data selection, arousal or valence ratings-based and LLD based on all manual ratings together demonstrated 5-7% performance gains. However, dominance ratings-based features only had a boost in performance after thresholds were applied to dominance ratings-based features and LLD features. This finding infers dominance requires data selection for a gain in depression classification performance.

4.3 Arousal, Valence, and Dominance Insights

A heightened level of depression usually results in reduced vocal dynamic energy [1], and in turn generally lower perceived speech arousal ratings. As thresholds rose and more lower rating scores were removed from LLD and/or arousal ratings-based features, a falling trend in arousal LLD performance was noted. Individuals with depression generally speak with perceptibly less dominance [11]. Therefore, it was speculated that as additional lower threshold dominance

ratings were removed, depression classification performance would decrease. The standalone dominance ratings-based features and LLD features generally had a steep drop off in classification performance when less than 80% of the ratings were retained. Speakers with depression typically have lower perceived valence ratings than healthy populations [14]. Consequently, introverted thresholds on standalone valence ratings-based features demonstrated up to 5% improvements with speech depression performance by retaining the outermost ratings, which included the severe (fringe).

5. Depression Classification Based on Automatic Emotion Prediction

The insights in Section 4 point to the advantage of using both ratings-based and acoustic features; this more evident in the case with consolidated small speech feature sets (i.e. LLD). A key question is how the insights from Sections 4.1 to 4.3 can be applied to an automatic system design. Overall, individual arousal manual ratings-based features performed best when compared with valence and dominance consolidating previous literature indicating that speech arousal is a strong indicator for depression [1]. Of course, manual ratings are not feasible for an automatic system, however automatic prediction of arousal and valence ratings has seen considerable research activity in the past few years [22-24]. Automatic prediction of arousal ratings in particular is supported by recent system proposals, including some based on very low-dimensional acoustic features which successfully emulated human-listening rating standards [25, 26].

To explore whether the efficacy of manual ratings and their derived features could be replicated using an automatic process, the following emotion prediction was proposed (Figure 4). To predict the arousal, valence, and dominance ratings, *Geneva Minimalistic Acoustic Parameter Set* (GeMAPS) features were extracted from train/development audio files for prediction of arousal (similarly to the AV+EC 2015 reference emotion prediction system) [27]. As in Huang et al. [24], these were passed to a *Relevance Vector Machine* (RVM) system trained iteratively on a leave-one-file-out basis to predict arousal ratings per file. Smoothing and delay compensation (6 seconds) were also applied. The automatic RVM arousal ratings achieved a correlation coefficient of 0.33 and 11.2 Root Mean Square Error (RMSE). This RMSE value was close to the original audio portion of the AVEC baseline (11.52 RMSE) [15] (note, valence and dominance had correlation coefficients of less than 0.10).

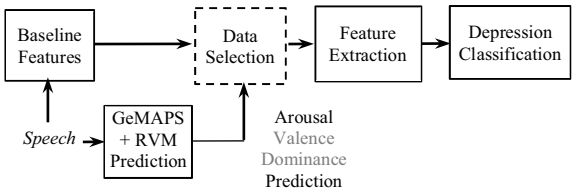

Figure 4: *Fully automatic depression classification using emotion prediction and emotion-based data selection.*

After the automatic arousal ratings were generated from GeMAPS features and an RVM system, data selection was applied using the upper, lower and extroverted bounds. Results showed that standalone automatically generated arousal rating-based features attained depression classification accuracy comparable with standalone manually generated arousal ratings-based features (Table 1).

Furthermore, results indicate that when combining automatically generated arousal and GeMAPS features, subsequent thresholding can produce fully automated systems whose performances are equal to that of manual ratings-based features. The automatically predicted arousal rater-based features improved with upper, lower, and extroverted data selection using narrower thresholds settings of 0.35 to 0.15 (retaining roughly 95% to 75% of the ratings).

Table 1. *Depression classification percentage accuracy based on arousal (A) from either manual ratings or emotion prediction and/or acoustic baseline features (BF).*

	Manual All	Auto All	Auto .35	Auto .30	Auto .25	Auto .20	Auto .15	Auto .10
UPPER								
A	70.0	67.5	71.2	71.2	71.2	75.0	71.2	67.5
A + BF	80.0	76.2	82.5	78.8	82.5	77.5	75.0	82.5
LOWER								
A	70.0	67.5	67.5	67.5	72.5	71.2	73.8	73.8
A + BF	80.0	76.2	76.2	76.2	82.5	78.8	81.2	77.5
EXTROVERTED								
A	70.0	67.5	71.2	70.0	71.2	70.0	70.0	68.8
A + BF	80.0	76.2	78.8	82.5	77.5	75.0	82.5	78.8

6. Conclusions

Given continuous affect ratings or scores, this research suggests that features derived from them carry complementary information to conventional acoustic features when classifying depression via speech. By applying different thresholds to manual or automatic ratings, ratings-based features can achieve performance gains. Experiments presented demonstrate higher performance than the baseline alone; irrespective of the data selection type or manual/automated approach, thresholds retaining roughly 75% to 95% over the rating values attained equal or better performance compared to without a threshold.

The application of automatic prediction of ratings-based arousal features was useful even though the correlation with the manual ratings was not high. The results show that for arousal, automated features derived from GeMAPS provide a boost in performance when emotion-based data selection is applied. More research on other databases is warranted. Furthermore, investigation into evaluating explicitly controlled speech segments (e.g. word/phrase data selection analysis) might reveal more insight in regards to informational regions of more value for depression classification.

7. Acknowledgements

This research was funded in part by the ARC Discovery Project grant DP130101094. NICTA is funded by the Australian Government as represented by the Department of Broadband Communication, and the Digital Economy, and the Australian Research Council through the ICT Centre of Excellence program. Thanks to fellow UNSW PhD colleagues Ting Dang and Zhaocheng Huang.

8. References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, 2015, pp. 10-49.

[2] N. Cummins, J. Epps, and E. Ambikairajah, "Spectro-temporal analysis of speech affected by depression and psychomotor retardation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver – Canada, 2013, pp. 7542-7546.

[3] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech," *INTERSPEECH 2013*, Lyon – France, 2013, pp. 857-861.

[4] B. Helfer, T. Quatieri, J. Williamson, D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," *INTERSPEECH 2013*, Lyon - France, 2013, pp. 2172-2176.

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

[6] A. Tellegen, "Structures of Mood and Personality and Their Relevance to Assessing Anxiety, with an Emphasis on Self-Report," in A.H. Tuma and J.D Maser (Eds.), *Anxiety and the Anxiety Disorders*, Hillsdale, NJ: Erlbaum, pp. 681-706, 1985.

[7] P. Lang, M. Greenwald, M. Bradley, and A. Hamm, "Looking at pictures: affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, 1993, pp. 261-273.

[8] M. Bradley and P. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, 1994, pp. 49-59.

[9] H. Perez, H. Escalante, L. Villasenor-Pineda, M. Montes-y-Gomez, D. Pinto-Avedano, and V. Reyes-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depressions recognition," *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, Orlando, FL – USA, 2014, pp. 49-55.

[10] J. Hall, J. Harrigan, and R. Rosenthal, "Nonverbal behaviour in clinician-patient interaction," *Applied and Preventive Psychology*, Vol. 4, Issue 1, Winter 1995, pp. 21-37.

[11] K. Osatuke, J. Mosher, J. Goldsmith, W. Stiles, D. Shapiro, G. Hardy, and M. Barkham, "Submissive voice dominate in depression: assimilation analysis of a helpful session," *Journal of Clinical Psychology: In Session*, vol. 63, no. 2, 2007, pp. 153-164.

[12] S. Karadogan and J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition of speech," *IEEE 3rd International Workshop on Cognitive Information Processing*, Parador de Baiona - Spain, 2012.

[13] L. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition and Emotion*, vol. 12, no. 4, 1998, pp. 579-599.

[14] J. Joorman and I. Gotlib, "Emotion Recognition in Depression: Relation to Cognitive Inhibition," *Cognition & Emotion*, vol. 24, no. 2, 2010, pp. 281-298.

[15] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, Orlando, FL – USA, 2014, pp. 3-10.

[16] A. Beck, R. Steer, R. Ball, and W. Ranieri, "Comparison of Beck depression inventories –ia and –ii in psychiatric outpatients," *Journal of Personality Assessment*, vol. 67, no. 3, 2004, pp. 588-597.

[17] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for the classification of depressed speech," *MIPRO 2014*, Opatija – Croatia, 2014, pp. 1331-1335.

[18] S. Scherer, L. Morency, J. Gratch, and J. Pestian, "Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis," *IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, Brisbane – Australia, 2015, pp. 4789-4793.

[19] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: a comparison between spontaneous and read speech," *ICASSP 2013*, Vancouver, B.C. – Canada, 2013, pp. 7547-7551.

[20] V. Sethu, J. Epps, and E. Ambikairajah, "Speech Based Emotion Recognition," in T. Ogunfunmi, R. Togneri, & M. Narasimhai (Eds.), *Speech and Audio Processing for Coding Enhancement and Recognition*, Springer, New York, 2014, pp. 197-228.

[21] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Variability compensation in small data: oversampled extraction of i-vectors for the classification of depressed speech," *IEEE International Conference on Acoustic*, *Speech*, *and Signal Processing (ICASSP)*, Florence – Italy, 2014, pp. 970-974.

[22] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015: The 5th international audio/visual emotion challenge and workshop," *MM '15 Proceedings of the 23rd ACM International Conference on Multimedia*, Brisbane – Australia, 2015, pp. 1335-1336.

[23] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," AVEC '15, Brisbane – Australia, Oct. 26th 2015.

[24] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction", AVEC '15, Brisbane – Australia, Oct. 26th 2015.

[25] D. Bone, C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: a rule-based framework with knowledge-inspired vocal features," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, 2014, pp. 201-213.

[26] C. Lee, D. Bone, and S. Narayanan, "An analysis of the relationship between signal-derived vocal arousal score and human emotion production and perception," *INTERSPEECH 2015,* Dresdan – Germany, 2015, pp. 1304-1308.

[27] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, K. Truong, 'The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing', *IEEE Trans. on Affective Computing*, 2015.