VARIABILITY COMPENSATION IN SMALL DATA: OVERSAMPLED EXTRACTION OF I-VECTORS FOR THE CLASSIFICATION OF DEPRESSED SPEECH

Nicholas Cummins[1,2], Julien Epps[1,2], Vidhyasaharan Sethu[1] , Jarek Krajewski[3]

[1]School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia
[3] Experimental Industrial Psychology, University of Wuppertal, Wuppertal, Germany
n.p.cummins@unsw.edu.au, j.epps@unsw.edu.au

ABSTRACT

Variations in the acoustic space due to changes in speaker mental state are potentially overshadowed by variability due to speaker identity and phonetic content. Using the Audio/Visual Emotion Challenge and Workshop 2013 Depression Dataset we explore the suitability of i-vectors for reducing these latter sources of variability for distinguishing between low or high levels of speaker depression. In addition we investigate whether supervised variability compensation methods such as Linear Discriminant Analysis (LDA), and Within Class Covariance Normalisation (WCCN), applied in the i-vector domain, could be used to compensate for speaker and phonetic variability. Classification results show that i-vectors formed using an over-sampling methodology outperform a baseline set by KL-means supervectors. However the effect of these two compensation methods does not appear to improve system accuracy. Visualisations afforded by the t-Distributed Stochastic Neighbour Embedding (t-SNE) technique suggest that despite the application of these techniques, speaker variability is still a strong confounding effect.

Index Terms— Depression, Acoustic Variability, I-vectors, Linear Discriminant Analysis, Within Class Covariance Normalisation, t-Distributed Stochastic Neighbour Embedding

1. INTRODUCTION

A wide range of acoustic information is modulated onto speech signals; this potentially places an upper-bound on the accuracy of a speech based depression classification system. Acoustic variability that arises due to speaker characteristics, channel effects and phonetic content has been shown to have detrimental effects on the accuracy of recognition of a range of paralinguistic information such as long-term speaker traits including age and gender [1], temporary speaker traits such as intoxication [2] and sleepiness [3], as well as transient speaker states such as emotion [4]. In emotion recognition, in particular, it has been shown that speaker variability affects the feature space distribution of emotional data [4]. Both automatic emotion and depression recognition systems share common traits; a continuous negative affect is a key symptom of depression [5]. However depression is more steady-state compared with the transient nature of emotions, with individuals inflicted for weeks or months rather than seconds or minutes [6].

In speaker recognition, i-vectors, together with a range of complementary transforms designed to further reduce errors arising from intersession variability, have become a pseudo standard due to their ability to compress both speaker and channel variability into a low-dimensional feature space [7], [8]. However little work has been done exploring the suitability of this paradigm for modelling paralinguistic tasks which often have substantially (both in terms of number of speakers and duration) smaller amounts of training data when compared with those used in speaker recognition.

Motivated by results showing that both speaker variability and phonetic variability have negative effects on depression classification [9], [10], we investigate the suitability of i-vectors for modelling depressed speech as well as the ability of the paradigm to reduce the effects of variability not related to depression.

2. RELATION TO PRIOR WORK

Whilst a range of prosodic [11], [12], voice quality [13], spectral features [9], [14] and Gaussian Mixture Model (GMM) based supervectors [10] have been established for use in an automatic speech based depression classifier, there are only a small number of papers which have investigated the effects of unwanted acoustic variability on depressed speech classification.

Results presented in [9] show that per-speaker normalisation offers no improvement for a depression classifier indicating that, as in emotion recognition, speaker variability has stronger effects than variability due to depression. Work in [15] shows that depression classification is susceptible to both speaker and channel effects.

Recent results, found using the Audio/Visual Emotion Challenge (AVEC) and Workshop 2013 Depression Dataset, show that Nuisance Attribute Projection (NAP) applied to Kullback-Leibler (KL-means) supervectors may be able to help reduce effects due to phonetic variability in a depression regression system [10]. Whilst this paper focuses on i-vectors for depressed speech classification, we also apply our final i-vector system configuration to the depression scale prediction challenge (Section 5.4) to allow comparison with results presented in [10].

The application of i-vectors to paralinguistic speech classification problems may be complicated by more than just the lack of previous investigation on comparatively small databases. Speaker traits like depression often only have examples of one class (i.e. low or high depression but not both) from a single speaker among the training/development data [14]. Compared with emotion or speaker recognition, in which training databases exist with examples of many emotions or channels per speaker [4], a different approach will be required.

Whilst i-vectors, and the related techniques of Joint Factor Analysis and Latent Factor Analysis, have been used in other paralinguistic classification tasks such as age and gender analysis [1], [16] and emotion classification [4], [17], to the best of the authors' knowledge this paper is the first paper to explore the suitability of i-vectors for the classification of speech affected by depression. Further, depression data, including AVEC, often pose the additional challenge that speech utterances from only a single level of depression per speaker are available.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

3. I-VECTOR PARADIGM

Given a Universal Background Gaussian mixture Model (UBM) to represent the feature distribution of the acoustic space, individual speech utterances can be used to adapt this UBM and the resulting GMM represented as supervectors. This allows for the application of a number of linear vector space operations but is held back by the inherent high dimensionality of supervectors. The i-vector space is a low dimensional subspace onto which supervectors are mapped via a linear transformation while retaining most of the variability (useful information) present in the supervector space and has been used extensively in speaker recognition [7], [8]. In the context of depression classification, it is expected that the UBM approximately models the phonetic structure of the acoustic space and hence the supervectors and consequently the i-vectors capture variations in this structure due to other factors including level of depression, speaker identity, channel effects, etc. The i-vector model is given by:

$$\Phi_x = \Phi + \mathbf{T}\boldsymbol{\psi} \quad (1)$$

where Φ is the supervector corresponding to the UBM, Φ_x is the supervector corresponding to a particular utterance, $\boldsymbol{\psi}$ is the i-vectors, and $\mathbf{T}$ is the projection matrix [7], [8].

3.1. Estimation of T Matrix

Mathematically, the i-vector model is a factor analysis method and the $\mathbf{T}$ matrix is estimated from training data. In general, results from speaker recognition show that the more training data used in the training of the $\mathbf{T}$ space, the more accurate the final system [18]. Typically of paralinguistic speech systems, as the database we are using has a limited training set (50 files, Section 4.1), to provide more supervector instances for the estimation of $\mathbf{T}$ we used an oversampling technique wherein multiple overlapping segments of speech (subfiles) were extracted from each file (Figure 1) [10].

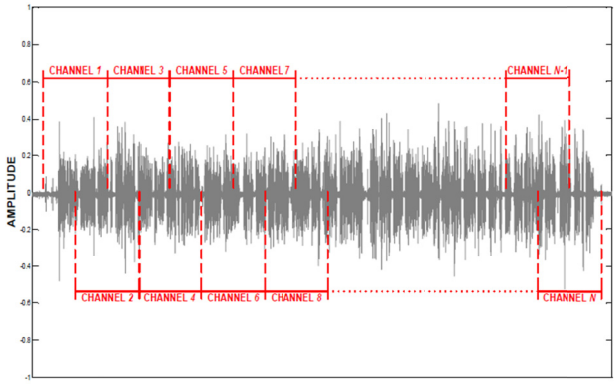


Figure 1: Example of creating different speech segments (subfiles) or 'channels' from a single sufficiently long audio file for extracting supervectors used to estimate T. Figure reproduced from [10].

3.2 Variability Compensation

To attempt to mitigate the effects of speaker and channel variability while retaining variability due to depression in the i-vector space, two forms of compensation were applied to the oversampled extracted i-vectors. The first, Linear Discriminant Analysis (LDA), attempts to find the linear projection from the i-vector space that maximises separation between classes (levels of depression in this case), while the second, Within Class Covariance Normalisation (WCCN), attempts to minimise variations within a class [7], [8].

Both LDA and WCCN were applied on a per class basis; using subfiles labelled with a class (low or high levels of depression) basis. WCCN was also trialled on a per-file basis; the transform matrix was trained using subfiles labelled on a per-file basis.

4. EXPERIMENTAL SET-UP

4.1. Corpora

All experiments presented used a subset of the Audio/Visual Emotion Challenge and Workshop (AVEC) 2013 dataset [19]. The full data set comprises 150 recordings, divided into training, development and testing partitions each of 50 files (recordings). Each recording has an associated Beck Depression Inventory (BDI) score, a self-reported measure of depression, with clinical validity, which rates the severity of cognitive, affective and somatic symptoms, to give a patient a score which reflects their level of depression [20]. For an in-depth description of the corpus, the reader is referred to [10], [19].

The make-up of this corpus provides some unique challenges. One is large variety in terms of both file length (ranging from 5 to 27 minutes in length) and speech tasks (vocal exercises, free and read speech tasks, noting that not all files include all tasks) contained within each file. This introduces a large degree of phonetic variability within each file [10]. Specifically induced changes in speaker affect are also present throughout each file [19]; this could potentially introduce another source of unwanted variation into our system.

To form a suitable two-class classification data set for preliminary experiments (i.e. one in which differences in acoustic variability between each class are most clearly due to the effects of depression), herein referred to as the development classification (DVC) partition, all files (twenty files from twenty distinct speakers) from the development set, with BDI scores between 0-9 (indicating mild to non-existent levels of depression), were selected to form the 'low' class. A further twenty files from twenty other speakers (out of 21 such files available in the development set - the longest one was left out to balance the two classes as best as possible), with BDI greater than 19 (severe depression), were selected from the development partition to form the high class.

4.2. Experimental Settings

The experimental settings (unless otherwise stated) of the classification system were as follows: the frame level features were MFCCs extracted as per [10]. A baseline classification accuracy was found using KL-means supervectors; the reader is referred to [21] for the extraction methodology. All UBMs were trained with 10 iterations of the EM algorithm from the entire training partition. The entire AVEC training set was also used to train the $\mathbf{T}$ matrix; i-vectors were then extracted for the two-class DVC partition described in Section 4.1. The number of UBM mixtures, MAP-adaptation iterations, $\mathbf{T}$ matrix dimensionality and LDA dimensionality were set empirically using the training partition and cross-fold validation on DVC partition.

LIBSVM with a linear kernel [22], and default user settings were used for all Support Vector Machine (SVM) testing and training. All training vectors were normalized to a range of [0, 1], and test vectors were normalized by the same factors as the training vectors. The choice of MFCCs as features was made for two reasons; firstly MFCCs, in combination with GMMs, have proved successful for low/high levels of depression [14], [15]. Secondly the use of MFCCs ensures that the feature space dimensionality is low enough to allow UBMs to be trained such that these UBMs are representative of the acoustic/phonetic structure of speech.

4.3. Evaluation Method

A similar cross fold validation scheme to that presented in [10] was used to generate the classification accuracy results. In this method 100 trials of 5-fold cross-validation are employed, where in each trial the DVC partition was randomly split into the 5 folds, which were then used as the SVM training and test permutations. Results are reported in terms of average accuracy across all trials (where each trial is the average of the 5 folds of cross-validation), minimum and maximum accuracy from each set of trials as well as the standard deviation in each trial set. Classification systems are referred to either as 'standard', where a single supervector / i-vector is extracted per file or 'oversampled'. When using the oversampling technique multiple scores were generated per file (one per each supervector/i-vector) and the median score was used to generate one prediction per file.

5. RESULTS

5.1 Baseline System Accuracy

As in [10], an initial series of comprehensive tests were run on the DVC partition using an uncompensated, KL-means supervector system. This system was chosen as our baseline as it was the most consistent performing system in [10]. Results confirmed that KL-means extracted from a 128 mixture GMM using 5 iterations of MAP also gave the best standard system (one supervector per file) result (70.30%) (Table 1). An oversampled uncompensated KL-means system (multiple supervectors per file) gave an accuracy of 67.85%, with the subfiles extracted from 60 second segments with a 10 second overlap (Table 1).

5.2 i-vector Classification

To determine the suitability of the i-vector paradigm for modelling depression, an exhaustive series of tests was run on a standard i-vector system (only *one* i-vector per file) to determine the ideal dimensionality of **T**. None of these classification accuracies were able to better either KL-means baselines (standard or oversampled) and the best accuracy (58.88%) was found for a **T** matrix dimensionality of 100 (Table 1).

For the oversampled system, we ran an exhaustive series of tests to determine the optimal parameters in terms of window size, overlap and **T** matrix dimensionality. Results indicated that the best setting was a 60 second window with a 10 second overlap. A sample of these results is shown in Figure 2. The best accuracy (74.85%) was achieved with a **T** dimensionality of 200; this is a relative improvement of 6.5% and 10.3% over the single and oversampled KL-means baselines respectively (Table 1). Perhaps more importantly, oversampling seems to allow relatively competitive i-vector systems to be developed where a 'standard' i-vector implementation provides close to chance-level accuracy.

Table 1: Standard and oversampled classification accuracies generated using uncompensated KL-means and i-vectors found using the DVC partition of the AVEC 2013 corpus.

Vector	System	mean	min	max	st.dev.
KL-means	Standard	70.30	57.50	80.00	4.01
	Oversampled	67.85	55.00	77.50	4.61
i-vectors	Standard	58.45	45.00	72.50	6.06
	Oversampled	74.85	60.00	85.00	5.41

We speculate that the improvement when using the oversampled i-vectors compared with the standard system is due to the increase in training samples (despite some redundancy) used to estimate the **T** matrix. 50 files were used to train the standard system whilst 1505 subfiles were used for the **T** matrix in the oversampled system. Further, the superior performance of the i-vector system when compared with the supervector system suggests that variations in the i-vector space may be more robust to phonetic content and speaker characteristics compared with those in the supervector space. Also, given that this system makes no attempt at speaker normalisation and given that i-vectors feature heavily in state-of-the-art speaker recognition systems, it is reasonable to infer from Table 1 that the use of the oversampled i-vector paradigm makes the depression classification system more robust to phonetic variability.

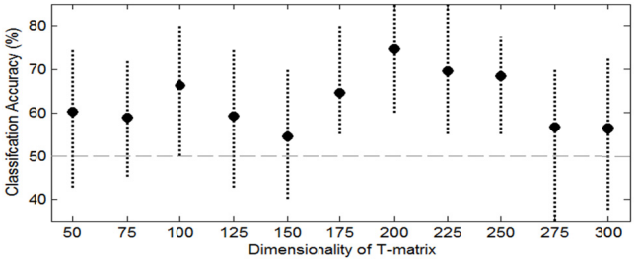


Figure 2: Classification accuracy for different dimensions of T, estimated using the oversampled technique, subfiles extracted from 60 second segments with a 10 second overlap, on the DVC partition of the AVEC 2013 corpus.

5.3 Effect of Variability Compensation

While the i-vector representation offers some inherent robustness to undesired variability, further variability compensation may be carried out in the i-vector space, and the application of two methods to the best performing system from section 5.2 (200 dimensional oversampled i-vector system with an estimated classification accuracy of 74.85%) was comprehensively evaluated. Specifically, the use of LDA and WCCN individually and LDA in combination with WCCN in the i-vector space were explored. Further, WCCN was applied on a per-depression class basis as well as a per-speaker basis (i.e., within speaker covariance normalisation). All results from these analyses (Table 2), however, indicate that these methods have a negative impact on system accuracy. Moreover, their performance was also lower than that of the KL-means baselines from Section 5.1. Section 5.4 attempts to shed some light on the reasons behind this drop in performance.

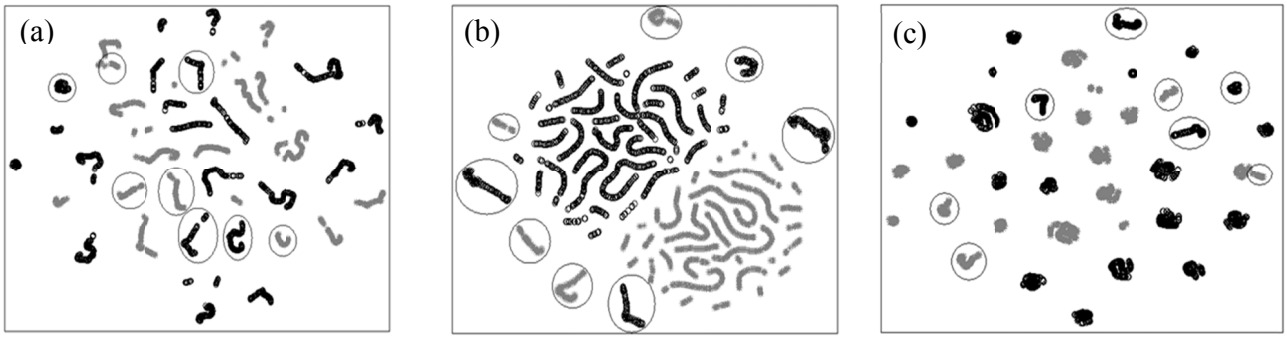
Table 2: Classification accuracy for a range of different variability compensation techniques, using 200-dimension i-vectors extracted via oversampling, found on the DVC partition of the AVEC 2013 corpus.

LDA dim	LDA only	LDA+ WCCN(class)	LDA+ WCCN(speaker)
No LDA		64.53	66.13
190	62.83	**66.18**	64.05
120	63.43	65.45	64.13
40	63.20	64.05	63.18

5.4 Visualisation of the i-vector space

In addition to classification experiments, the t-Distributed Stochastic Neighbour Embedding (t-SNE) technique [23] was employed to visualise the i-vector space before and after the application of LDA and WCCN. The t-SNE technique allows visualization of high dimensional data in 2 or 3 dimensions by attempting to preserve local structure of the high dimensional space in the low dimensional space [23].

The t-SNE plots shown in Figure 3, help us gain an understanding of why LDA and WCCN have such a negative impact on i-vector system accuracy. Figure 3a shows an uncompensated i-vector space taken from one fold of a cross fold



Figures 3: (a) t-SNE plot of uncompensated i-vector system, (b) LDA compensated i-vector system, and (c) WCCN (per speaker) compensated i-vector system. Black points represent low class, grey high class. Test samples are circled.

validation test (test partitions for all folds were chosen randomly and of equal size). We speculate that as the number of distinct clusters approximately equals the number of speakers in the DVC partition, that each individual cluster in the visualisation represents the oversampled i-vectors from a single speaker.

LDA has the effect of separating the training samples into two distinct class groups very well (Figure 3b). However this separation does not appear to generalise to unseen speakers (clusters corresponding to test samples depicted in the plots within circles) and consequently classification accuracy on these samples suffers. Further, the test samples still appear to be organised in speaker clusters and are well separated from the two training classes. We speculate this could be due to effects of the speaker identity that are not successfully mitigated by the transform. Similar results were seen for WCCN applied on a per-class basis (not shown here).

The effects of WCCN applied on a per-speaker basis (Figure 3c) are a reduction in the 'length' of each speaker cluster when compared with Figure 3a but do not alter the global structure of the i-vector space and consequently do not achieve any 'useful normalisation'. We speculate that WCCN on a per-speaker basis is reducing phonetic variability within each file. However, as the i-vector representation itself appears to do this to a large extent and as within-speaker covariance normalisation does not actually improve results we speculate that speaker identify is the major confounding factor.

5.5 Depression Scale Prediction using i-vectors

Finally, for transparency, we compare the proposed systems with those presented in [10] and baseline corpus results [19], we tested both an oversampled uncompensated i-vector system as well as an oversampled i-vector WCCN (per speaker) system for their ability to estimate depression scale (i.e. regression as opposed to 2-class classification) on two sets, the AVEC development and test partitions. All system accuracies are reported in terms of Root Mean Square Error (RMSE).The reader is referred to [10] for the setup of these tests.

Table 3: RMSE's generated using the oversampling method compared with system accuracies from [10].

System	RMSE	
	Devel.	Test
Baseline [19]	10.75	14.12
KL-means (standard) [10]	9.00	10.17
KL-means (oversampled+ NAP) [10]	8.94	13.34
i-vectors (oversampled)	10.34	11.37
i-vectors (oversampled *WCCN speaker*)	10.13	11.58

On the development set, both i-vector systems beat the challenge baselines, but were unable to beat the KL-means supervector system. On the test set, the proposed i-vector systems outperformed the challenge baseline and the oversampled KL means, NAP compensated, supervector system but not the standard KL-means supervector system. Interestingly, WCCN applied on a per-speaker basis lowers the RMSE as evaluated on the development set but not on the test set. It should be noted that the i-vector systems were set up and optimised for a classification task rather than a regression task.

6. CONCLUSION

Speaker variability, phonetic content and intersession (recording setup) variability have all previously been shown to introduce unwanted variability, to both depression classification and prediction systems [9], [10], [15]. This paper presents an oversampled extraction technique for the i-vector systems in smaller datasets. Apart from permitting the development of useful i-vector systems in this context, this technique was found to be better suited to classifying high and low levels of depression than both an uncompensated KL-means supervector system and the standard i-vector (one i-vector per file) extraction techniques. We speculate that this is due to both the suitability of i-vectors for minimizing the effects of unwanted variability in a classification context given the possibility of estimating reasonable i-vector parameters via oversampling.

Further attempts to minimize unwanted variability through LDA and WCCN were unsuccessful. We speculate that as in [9] this is due to the speaker variability being stronger than the variability due to the effects of depression. Visualisations based on t-SNE support this speculation. In addition to classification systems, results from the regression analysis show that i-vector systems could outperform the challenge baseline and provide performance close to that of a competitive system operating on the entire AVEC data [10].

Future work includes testing the effects of a less naïve oversampling method. Given the result shown in the t-SNE plots in section 5.3, cluster-based classification techniques will also be trialled.

7. ACKNOWLEDGEMENTS

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. This work was partly funded by ARC Discovery Project DP130101094 (Epps) and the German Research Foundation (KR3698/4-1) (Krajewski).

8. REFERENCES

[1] M. Li, A. Metallinou, D. Bone, and S. Narayanan, "Speaker states recognition using latent factor analysis based Eigenchannel factor vector modeling," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 1937–1940.

[2] D. Bone, M. Li, M. P. Black, and S. S. Narayanan, "Intoxicated Speech Detection: A Fusion Framework with Speaker-Normalized Hierarchical Functionals and GMM Supervectors," *Comput. Speech Lang.*, p. NA, 2012.

[3] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing*, vol. 84, pp. 65–75, 2012.

[4] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker Variability In Speech Based Emotion Models - Analysis and Normalisation," in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7522–7526.

[5] R. J. Davidson, D. Pizzagalli, J. B. Nitschke, and K. Putnam, "Depression: Perspectives from Affective Neuroscience," *Annu. Rev. Psychol.*, vol. 53, no. 1, pp. 545–574, 2002.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Process. Mag. IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[7] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification.," in *10th Annual Conference of the International Speech Communication Association Interspeech2009*, 2009, vol. 9, pp. 1559–1562.

[8] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[9] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in *12th Annual Conference of the International Speech Communication Association Interspeech2011*, 2011, pp. 2997–3000.

[10] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of Depression by Behavioural Signals: A Multimodal Approach," in *The 21st ACM International Conference on Multimedia*, 2013, p. NA.

[11] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *J. Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[12] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biol. Psychiatry*, vol. 72, pp. 580 –587, 2012.

[13] S. Scherer, G. Stratou, J. Gratch, and L. Morency, "Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD," in *14th Annual Conference of the International Speech Communication Association Interspeech2013, Lyon, France, 25-29 Aug 2013*, 2013, pp. 847–851.

[14] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling Spectral Variability for the Classification of Depressed Speech," in *14th Annual Conference of the International Speech Communication Association Interspeech2013, Lyon, France, 25-29 Aug 2013*, 2013.

[15] D. Sturim, P. A. Torres-Carrasquillo, T. F. Quatieri, N. Malyska, and A. McCree, "Automatic Detection of Depression in Speech Using Gaussian Mixture Modeling with Factor Analysis," in *12th Annual Conference of the International Speech Communication Association Interspeech2011*, 2011, pp. 2983–2986.

[16] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 151–167, Jan. 2013.

[17] R. Xia, and Y. Liu, "Using i-Vector Space Model for Emotion Recognition," in *13th Annual Conference of the International Speech Communication Association Interspeech2012*, 2012, pp. 2230–33.

[18] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. IEEE Odyssey*, 2010.

[19] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "The Continuous Audio/Visual Emotion and Depression Recognition Challenge," in *The 21st ACM International Conference on Multimedia*, 2013, p. NA.

[20] 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.

[21] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *2006 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, (ICASSP' 06).*, 2006, vol. 1, pp. 97–100.

[22] C.-C. Chang, and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 1:27, 2011.

[23] L. J. P. Maaten, and G. E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. nov, pp. 2579–2605, 2008.