Interspeech 2018
2-6 September 2018, Hyderabad, India

Effectiveness of Voice Quality Features in Detecting Depression

Amber Afshan[1], *Jinxi Guo*[1], *Soo Jin Park*[1], *Vijay Ravi*[1], *Jonathan Flint*[2], *Abeer Alwan*[1]

[1]Dept. of Electrical Engineering, University of California Los Angeles, USA
[2]Dept of Psychiatry and Biobehavioral Sciences, UCLA School of Medicine, Los Angeles, USA

amberafshan@g.ucla.edu, lennyguo@g.ucla.edu, sj.park@ucla.edu, vijaysumaravi@g.ucla.edu, jflint@mednet.ucla.edu, alwan@ee.ucla.edu

Abstract

Automatic assessment of depression from speech signals is affected by variabilities in acoustic content and speakers. In this study, we focused on addressing these variabilities. We used a database comprised of recordings of interviews from a large number of female speakers: 735 individuals suffering from depressive (dysthymia and major depression) and anxiety disorders (generalized anxiety disorder, panic disorder with or without agoraphobia) and 953 healthy individuals. Leveraging this unique and extensive database, we built an i-vector framework. In order to capture various aspects of speech signals, we used voice quality features in addition to conventional cepstral features. The features (F0, F1, F2, F3, H1-H2, H2-H4, H4-H2k, A1, A2, A3, and CPP) were inspired by a psychoacoustic model of voice quality [1]. An i-vector-based system using Mel Frequency Cepstral Coefficients (MFCCs) and another using voice quality features was developed. Voice quality features performed as well as MFCCs. A score-level fusion was then used to combine these two systems, resulting in a 6% relative improvement in accuracy in comparison with the i-vector system based on MFCCs alone. The system was robust even when the duration of the utterances was shortened to 10 seconds.

Index Terms: depression detection, computational paralinguistics, voice quality features, i-vectors

1. Introduction

The deployment of automatic assessment systems would transform the ability to diagnose, treat and prevent major depressive disorders (MDD). MDD affects almost one in five women and one in twelve men in their lifetime [2] and was recently recognized as the world's leading cause of disability [3]. Yet current pharmacological [4] and psychological therapies [5] provide limited efficacy, and only about half of those suffering from MDD are identified and offered treatment [6]. An obstacle preventing effective use of existing therapies, and impeding the discovery of better ones, is the difficulty of diagnosing MDD. Diagnosis is still made on the basis of a clinical interview and mental status examination, a method with relatively low reliability; screening instruments are hampered by poor specificity and sensitivity and no reliable biomarkers exist. Further complicating the problem, MDD remains a syndromal diagnosis, leaving open the possibility that it consists of a number of different conditions, each with different etiologic pathways and treatment responses; indeed, there is mounting evidence that MDD is not monolithic [7]. Early intervention before the onset of severe symptoms can alleviate MDD's worst consequences including suicide.

One possible source of information for improving diagnosis, and recognizing subtypes, is the characterization of MDD from a person's speech. Changes in the way people talk reflect alterations in mood, but attempts to use this information have not so far been clinically useful. Depression can be characterized by prosodic abnormalities and/or articulatory and phonetic errors [8]. There are links between depression and alterations in the dynamics of vocal tract resonances or formants [9, 10] and there have been studies using prosodic [11], voice quality [12], and spectral features [13, 14].

In recent years, speech technology has been used to perform automatic identification of depression. Some studies investigated using single phoneme or word-level utterances for recognizing depression [15, 16]. A few studies investigate non-speech patterns, diadochokinesis patterns or nonsense words [11] and other studies have used either read speech [17, 18] or spontaneous speech [17, 16, 15].

Our work focuses on building algorithms to enable reliable automated detection of MDD from speech signals, with a special focus on voice quality features.

1.1. Related Work on Voice Quality and i-vectors

There have been several studies showing that voice quality contains information about the mental state of a person [19, 8]. In depression detection, commonly used voice quality measures include jitter, shimmer, the small cycle-to-cycle variations in glottal pulse amplitude in voiced regions, harmonic-to-noise ratio, and the ratio of harmonics to inharmonic components [19]. These features are related to vocal fold vibration, which is influenced by vocal fold tension and subglottal pressure.

Voice quality features have not been effectively applied to automatic detection of depression. One of the main reasons being the definition of the term 'voice quality'. Voice quality has been represented using impressionistic labels, such as tense, harsh, and breathy which have different interpretations based on the researcher. Moreover, it is difficult to robustly extract voice quality features from the speech signal. One technique involves inverse-filtering to identify voice source characteristics by removing the effects of the vocal tract transfer function [20]. Avoiding the difficult inverse-filtering approach, other techniques have been proposed to estimate voice source characteristics. Jitter and shimmer are two of the most popular features in this direction [8] but it is unclear how they relate to the perception of voice quality.

Speaker and phonetic variability are shown to degrade the performance of depression detection systems [13, 21]. There have been studies using Gaussian Mixture Model (GMM) based supervectors, and Nuisance Attribute Projection (NAP) for Kullback-Leibler (KL-means) supervectors to reduce effects due to phonetic variability [13]. Recently, the total variability framework was introduced as an effective approach to capture the important variabilities in a low dimensional space [22]. Using this framework, the i-vector approach was developed [23] which has become the state-of-the-art system for speaker veri-

10.21437/Interspeech.2018-1399

fication. But little effort has been made in using the i-vectors for a depression detection task. The main reason is the lack of large databases to learn the Universal Background Model (UBM) and the total variability matrix for i-vector extraction. Some studies have worked around this problem and applied i-vectors for this task. Oversampling of the data is done in [24] to extract i-vectors for depression detection. Paula et al. in [25] perform experiments using MFCCs, Shifted-Delta-Cepstra (SDC), Rasta Perceptual Linear Prediction Coefficients (PLP), and spectral and prosodic features as input to an i-vector system with feature concatenation for estimating depression level. Further, experiments in a speaker independent setup are shown in [26]. A multimodal setup using video features as well with MFCC based i-vectors is described in [27]. These studies have shown the potential of i-vectors in improving the depression detection performance.

The rest of the paper is organized as follows. Sections 2 and 3 describe the database and acoustic features used in this paper, followed by Section 4 which describes the system used for depression identification. Section 5 presents the results and a discussion, while Section 6 concludes the paper.

2. Database

The depression database used in this study was developed as a part of the Experimental Research on Genetic Epidemiology (CONVERGE) study [28]. The CONVERGE study was designed for a genome-wide association of major depression disorders and thus focused on a few cases with increased genetic risk for MDD. In order to obtain a more genetically homogeneous sample only women were recruited to the study (the genetic correlation between males and females for depression is approximately 0.6)[29]. Each subject was interviewed by a trained interviewer assisted by a computerized assessment. The diagnoses of depressive (dysthymia and MDD) and anxiety disorders (generalized anxiety disorder - GAD, panic disorder with or without agoraphobia) were made with the Composite International Diagnostic Interview (Chinese version) [30], which classifies diagnoses according to the Diagnostic and Statistical Manual of Mental Disorders fourth edition (DSM-IV) criteria.

The database includes recordings of the interviews from 735 individuals classified as suffering from MDD and 953 healthy individuals. The database is in Mandarin. All the audio recordings were collected with a sampling rate of 16kHz. There are a total of 52 hours and 28 minutes of data. A large degree of phonetic and content variability characterize this database.

3. Acoustic Features

3.1. ComParE 2016 Acoustic Feature Set

The ComParE 2016 feature set has been used in paralinguistics analysis [31] and in previous depression research [27, 32]. This set consists of F0, energy, spectral, cepstral coefficients (MFCCs) and voicing related frame-level features which are referred to as low-level descriptors (LLDs). They also include zero crossing rate, jitter, shimmer, harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness. In total, this feature set contains 6373 static features resulting from the computation of various functionals over low-level descriptor contours. These functionals are statistical, polynomial regression coefficients and transformations on the low-level descriptors. We used the TUMs open-source openSMILE system to extract the ComParE16 features[33] .

3.2. Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficient (MFCCs) were extracted with a window size of 25 ms, a window shift of 10 ms, a pre-emphasis filter with coefficient 0.97, and a sinusoidal lifter with coefficient 22. A filter bank with 23 filters was used and 13 coefficients were extracted. Utterances were downsampled to 8 kHz before feature extraction. We also used the first and second derivatives of MFCCs.

3.3. Voice Quality Features (VQual)

Based on extensive studies on the patterns of variability across speakers in source spectral shapes and glottal pulse shapes [34], a spectral model to represent the voice source contribution to perceived voice quality has been developed [35]. The model parameters include the fundamental frequency (F0), harmonic-to-noise ratios, and difference in harmonic amplitudes H1-H2, H2-H4, H4-H2k where the amplitudes of first, second, and fourth harmonics, and the harmonic nearest to 2 kHz as H1, H2, H4, and H2k. This model is perceptually valid in that listeners are sensitive to the parameters of the model. This set of parameters account for perceived voice quality (e.g., [36, 37, 38, 39]).

Inspired by this model, a feature set was developed for automatic speaker verification applications. The feature set, denoted as VQual, comprised of F0, first three formants (F1, F2, F3), H1-H2, H2-H4, H4-H2k, formant amplitudes A1, A2, A3 and cepstral peak prominence (CPP, [40]). The formants F1, F2, and F3 were added to capture the variation in vowel quality which differs substantially across (and occasionally within) speakers. CPP, a measure of signal periodicity, replaced the harmonic-to-noise ratios. This set of features was effectively applied to automatic speaker verification [41, 1]. The features were extracted every 10ms using VoiceSauce software [42]. We also added the features' first and second derivatives. Even though the feature set was originally developed to capture speaker identity, we expect this feature set to provide valuable information for automatic depression detection as well.

4. System Description

4.1. Gaussian Mixture Models for Classification

To model frame-level features we used Gaussian Mixture Models (GMMs). We trained GMMs for both the depressed and non-depressed cases i. e., the Expectation Maximization algorithm is used to cluster the data. After obtaining the GMMs, we used a maximum likelihood estimator to obtain the similarity of each test utterance to either class.

4.2. Total Variability Modeling

In the total variability space, the Universal Background Gaussian Mixture Model (UBM) which represents the feature distribution of the acoustic space, is adapted to a set of given speech frames based on the eigenvoice adaptation technique [43] in order to estimate utterance-dependent GMM parameters. The eigenvoice adaptation technique operates on the assumption that all the pertinent variability is captured by a low rank rectangular matrix T named the total variability matrix. The i-vector extraction can be represented as follows:

$$M = m + Tv \tag{1}$$

where m is the mean super-vector of the UBM. M is the mean centered super-vector of the speech utterance derived using the

0^{th} and 1^{st} order Baum-Welch statistics. v is the i-vector the representation of a speech utterance.

In this work, we consider binary classification of classes: depressed or non-depressed. We followed the approach described in [23] to extract the i-vectors considering these two classes.

4.3. Logistic Regression

Using i-vectors, we performed classification with logistic regression [44]. We learn the regression coefficients from training data by maximizing the log likelihood. We then applied the logistic regression algorithm to estimate the probability that a given utterance belongs to a particular class.

4.4. Fusion of Scores

Since MFCCs and Voice Quality features carry complementary information we built separate i-vector classification systems using those features. We then used a score-level (log probability) fusion approach to combine the results to test for further improvements. Here, we linearly combined the scores using the following equation:

$$s = \alpha s_v + (1 - \alpha) s_m \tag{2}$$

where s_m and s_v correspond to the logistic regression scores using MFCCs and VQual respectively, α ranging from 0 to 1 is the coefficient. The scores were scaled to have zero-mean and unit-variance prior to fusion.

5. Experiments and Results

5.1. The Experimental Setup

The data were split into train and test sets by randomly assigning 70% of the speakers to the train set and 30% to the test set. After MFCCs and VQual feature extractions, as described in Section 3, a UBM of 256 mixtures was trained for each feature set. Followed by total variability matrix calculation, and used it to extract i-vectors of dimension 600. Since we had an adequate amount of data available, we trained a UBM using the training data alone without any data augmentation [24]. The i-vectors were then classified using a logistic regression model trained using the i-vectors of the training data. We then linearly added the scores of MFCCs and VQual feature classifiers to obtain the score-level fusion results.

For the baseline systems, we trained 256 mixture GMMs. Thus, we maintained uniformity between the i-vector setup and the baseline. We also evaluated the performance of the ComParE 2016 setup on the CONVERGE data.

5.2. Results

Results obtained for different classifier setups are summarized in Table 1. We perform the classification by using the feature sets individually, followed by using i-vectors for the each of the feature sets. ComParE16 feature set performed better than MFCCs and VQual for classification. It can be seen that i-vectors improved the accuracies by 26.86%, 29.66% and 7.54% for MFCCs, VQual and ComParE16 features respectively. Thus, proving that i-vectors are able to successfully decrease the impact of speaker and phonetic variability in speech.

Also note that VQual i-vectors provided results comparable to MFCC i-vectors, and they improved the performance by 6% (relative) when fused with scores from the MFCC i-vectors system. Thus, proving our hypothesis that VQual features provide information complementary to that provided by MFCCs. In contrast, score fusion did not improve the results in the case of using features only but rather it worsened it. Note that we also fused ComParE16 i-vector system with the fused MFCC and VQual i-vector system. The results remained almost the same as the fused MFCC and VQual i-vector system. Additionally, we also concatenated MFCC and VQual features and used them in the i-vector framework. The accuracy from this system was not on par with the score level fusion system.

It is not always feasible to obtain lengthy speech recordings from subjects. Hence, we investigated the robustness of the system as the length of the input speech is decreased. To do this, we split the test utterances into smaller segments of 10s, 20s, 30s and 40s and used these segments to perform classification. Figure 1 depicts the changes in accuracy with test-utterance duration. For these experiments we trained on full segments (1.8 min) and used three different i-vector setups MFCCs, VQual and score level fusion of MFCCs and VQual. As expected the overall performance of the classifiers decreased with the decrease in the test-utterances duration. But interestingly, we can see that fusion results continued to outperform the individual MFCC i-vectors and VQual i-vector systems.

Further, we observe that VQual features when fused with MFCCs provide a significant improvement. As the duration of the test utterance decreases this relative improvement increases from 8% when the complete utterance is provided to 15% when the test duration is 10s

5.3. Discussion

It is difficult to detect depression using frame-level features. Usage of utterance-level information improved the results. This can be seen in our experiments using i-vectors. Moreover, this improvement is also due to the normalization in the total variability space. Thus, decreasing the impact of speaker and phonetic variability on system performance. Note that even though, the ComParE16 features include across utterance statistics, the performance improvement is not as much as using i-vector framework.

Combining the MFCC and VQual features into a single model by concatenation did not perform as well as the score level fusion approach. So, we used score-level fusion for our further experiments.

As expected, both MFCCs and VQual features performed worse as test utterances became shorter. But, the VQual features were able to improve the performance of the system through score fusion by providing complementary information to MFCCs. We can continue to detect depression with an accuracy of 77% even when the test utterances were 10 seconds long and the accuracy is as high as 95% when the test utterances were 1.8 minutes long.

6. Conclusion

This study proposed the use of voice quality features (VQual), which account for perceived voice quality, for depression detection. We used the VQual feature set in combination with MFCCs at the score-level to obtain improvement over each system. We showed improved performance when i-vectors are used for the depression detection, and discussed the robustness of our setup as the duration of the test utterance decreased.

Future work will include using auto encoders [45] to learn the most effective features for detecting depression. Addition-

Table 1: *Results on depression detection using different feature sets with and without the i-vector framework. Boldface indicate the numbers which are the best among all the experiments*

Features	Precision	Recall	F1-score	Accuracy
MFCCs	0.4070	0.9206	0.5645	0.6272
VQual	0.3614	0.8692	0.5105	0.5792
ComParE16	0.8016	0.8311	0.8161	0.8064
Score Fusion (MFCCs & VQual)	0.3930	0.9346	0.5533	0.6252
MFCC i-vectors	0.8281	0.9860	0.9002	0.8958
VQual i-vectors	0.8807	0.8692	0.8749	0.8758
ComParE16 i-vectors	0.9018	0.8551	0.8778	0.8818
MFCC & VQual i-vectors	0.90175	0.91589	0.90877	0.90782
Score Fusion (MFCC & VQual) i-vectors	**0.9263**	0.9766	**0.9508**	**0.9479**
Score Fusion (MFCC, VQual & ComParE16) i-vectors	0.9193	**0.98131**	0.94929	0.94589

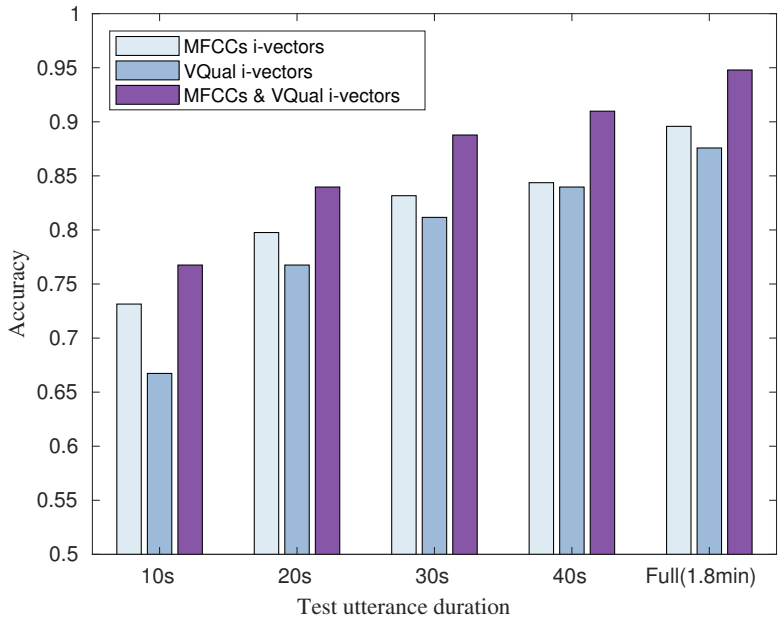

Figure 1: *Effect of test utterance duration on the performance of the system*

ally, as the CONVERGE data is large enough one more interesting analysis would be using deep neural networks to perform the detection.

7. References

[1] S. J. Park, G. Yeung, J. Kreiman, P. A. Keating, and A. Alwan, "Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems," *Interspeech*, pp. 1522–1526, 2017.

[2] American Medical Association, "for Treatment of Mental Disorders in the World Health Organization," *Jama*, vol. 291, no. 21, pp. 2581–2590, 2004.

[3] World Health Organization, "Depression and other common mental disorders: global health estimates," *World Health Organization*, pp. 1–24, 2017.

[4] P. A. Vöhringer and S. N. Ghaemi, "Solving the Antidepressant Efficacy Question: Effect Sizes in Major Depressive Disorder," *Clinical Therapeutics*, vol. 33, no. 12, pp. B49–B61, 2011.

[5] P. Cuijpers, A. Van Straten, E. Bohlmeijer, S. D. Hollon, and G. Andersson, "The effects of psychotherapy for adult depression are overestimated: A meta-analysis of study quality and effect size," *Psychological Medicine*, vol. 40, no. 2, pp. 211–223, 2010.

[6] K. B. Wells, R. D. Hays, M. A. Burnham, W. H. Rogers, S. Greenfield, and J. E. Ware, "Detection of depressive disorder for patients receiving prepaid or fee-for-service care. Results from the Medical Outcomes Study." *Jama*, vol. 262, pp. 3298–3302, 1989.

[7] K. S. Kendler, S. H. Aggen, and M. C. Neale, "Evidence for multiple genetic factors underlying DSM-IV criteria for major depression," *JAMA Psychiatry*, vol. 70, no. 6, pp. 599–607, 2013.

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

[9] B. Stasak, J. Epps, and R. Goecke, "Elicitation design for acoustic depression classification: An investigation of articulation effort, linguistic complexity, and word affect," *INTERSPEECH*, vol. 2017-Augus, pp. 834–838, 2017.

[10] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Automatic modelling of depressed speech: Relevant features and relevance of gender," *INTERSPEECH*, no. 444, pp. 1248–1252, 2014.

[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," *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[12] S. Scherer, G. Stratou, J. Gratch, and L. P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD," *INTERSPEECH*, pp. 847–851, 2013.

[13] N. Cummins, J. Joshi, and R. Goecke, "Diagnosis of Depression by Behavioural Signals : A Multimodal Approach Categories and Subject Descriptors," *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 11–20, 2013.

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

[15] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting Depression using Vocal, Facial and Semantic Communication Cues," *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pp. 11–18, 2016.

[16] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological Psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[17] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," *ICASSP*, no. May, pp. 7547–7551, 2013.

[18] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 3, pp. 17–20, 2004.

[19] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," *13th Annual Conference of the International Speech Communication Association*, pp. 1058–1061, 2012.

[20] M. Fröhlich, D. Michaelis, and H. W. Strube, "SIMsimultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 479–488, 2001.

[21] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," *INTERSPEECH*, pp. 2997–3000, 2011.

[22] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," *INTERSPEECH*, pp. 1559–1562, 2009.

[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[24] 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," *ICASSP*, pp. 970–974, 2014.

[25] P. Lopez-Otero, L. Dacia-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for depression detection," *2nd International Workshop on Biometrics and Forensics, IWBF 2014*, 2014.

[26] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "Assessing speaker independence on a speech-based depression level estimation system," *Pattern Recognition Letters*, vol. 68, pp. 343–350, 2015.

[27] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features," *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pp. 43–50, 2016.

[28] Y. Li, S. Shi, F. Yang, J. Gao, Y. Li, M. Tao, G. Wang, K. Zhang, C. Gao, L. Liu, K. Li, K. Li, Y. Liu, X. Wang, J. Zhang, L. Lv, X. Wang, Q. Chen, J. Hu, L. Sun, J. Shi, Y. Chen, D. Xie, J. Flint, K. S. Kendler, and Z. Zhang, "Patterns of co-morbidity with anxiety disorders in Chinese women with recurrent major depression," *Psychological Medicine*, vol. 42, no. 6, pp. 1239–1248, 2012.

[29] K. S. Kendler, C. O. Gardner, M. Gatz, and N. L. Pedersen, "The sources of co-morbidity between major depression and generalized anxiety disorder in a Swedish national twin sample," *Psychological Medicine*, vol. 37, no. 3, pp. 453–462, 2007.

[30] W. Ter Smitten, MH and Smeets, RMW and Van den Brink, "Composite international diagnostic interview (CIDI), version 2.1," *Amsterdam: World Health Organization*, no. 343, pp. 343–345, 1998.

[31] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," *INTERSPEECH*, vol. 08-12-Sept, pp. 2001–2005, 2016.

[32] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[33] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," *Proceedings of ACM Multimedia*, pp. 1459–1462, 2010.

[34] J. Kreiman, B. R. Gerratt, and N. Antoñanzas-Barroso, "Measures of the Glottal Source Spectrum," *Journal of Speech Language and Hearing Research*, vol. 50, no. 3, p. 595, 2007.

[35] M. Garellek, R. Samlan, B. R. Gerratt, and J. Kreiman, "Modeling the voice source in terms of spectral slopes," *The Journal of the Acoustical Society of America*, vol. 139, no. 3, pp. 1404–1410, 2016.

[36] J. Kreiman and B. R. Gerratt, "Perceptual sensitivity to first harmonic amplitude in the voice sourcea)," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2085–2089, 2010.

[37] M. Garellek, R. A. Samlan, J. Kreiman, and B. R. Gerratt, "Perceptual sensitivity to a model of the source spectrum," *Proceedings of Meetings on Acoustics*, vol. 19, no. May, pp. 1–5, 2013.

[38] J. Kreiman and B. R. Gerratt, "Perceptual interaction of the harmonic source and noise in voice," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 492–500, 2012.

[39] M. Garellek, P. Keating, C. M. Esposito, and J. Kreiman, "Voice quality and tone identification in White Hmong," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 1078–1089, 2013.

[40] J. Hillenbrand and R. A. Houde, "Acoustic Correlates of Breathy Vocal QualityDysphonic Voices and Continuous Speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[41] S. J. Park, C. Sigouin, J. Kreiman, P. Keating, J. Guo, G. Yeung, F.-y. Kuo, and A. Alwan, "Speaker Identity and Voice Quality : Modeling Human Responses and Automatic Speaker Recognition," *Interspeech*, pp. 1044–1048, 2016.

[42] Y.-L. Shue, P. Keating, and C. Vicenik, "VOICESAUCE: A program for voice analysis." *The Journal of the Acoustical Society of America*, vol. 126, no. August, p. 2221, 2009.

[43] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE transactions on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005.

[44] P. McCullagh, "Generalized linear models," *European Journal of Operational Research*, vol. 16, no. 3, pp. 285–292, 1984.

[45] A. Pal and S. Baskar, "Speech emotion recognition using Deep Dropout Autoencoders," *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, no. March, pp. 1–6, 2015.