INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition

Bogdan Vlasenko[1,2], *Hesam Sagha*[1], *Nicholas Cummins*[1], *Björn Schuller*[1,3]

[1]Chair of Complex & Intelligent Systems, Universität Passau, Germany
[2]Idiap Research Institute, Martigny, Switzerland
[3]Machine Learning Group, Imperial College London, U.K.
bogdan.vlasenko@uni-passau.de

Abstract

Whilst studies on emotion recognition show that gender-dependent analysis can improve emotion classification performance, the potential differences in the manifestation of depression between male and female speech have yet to be fully explored. This paper presents a qualitative analysis of phonetically aligned acoustic features to highlight differences in the manifestation of depression. Gender-dependent analysis with phonetically aligned gender-dependent features are used for speech-based depression recognition. The presented experimental study reveals gender differences in the effect of depression on vowel-level features. Considering the experimental study, we also show that a small set of knowledge-driven gender-dependent vowel-level features can outperform state-of-the-art turn-level acoustic features when performing a binary depressed speech recognition task. A combination of these preselected gender-dependent vowel-level features with turn-level standardised openSMILE features results in additional improvement for depression recognition.

Index Terms: Depression, Gender, Vowel-Level Formants, Speech Motor Control, Classification

1. Introduction

It has been predicted that, within the next 15 years, unipolar depression along with heart disease, will become one of the leading causes of disabilities worldwide [1]. Despite this increasing prevalence, diagnostic tools remain rooted, almost exclusively, in patient-based questionnaires. Such tools are open to a range of subjective biases including the skill and experience of a clinician and the reliability of a patient's own insights on their current mental state [2]. With the aim of enhancing current diagnostic techniques, investigations into new approaches for objectively detecting and monitoring depression based on measurable biological, physiological or behavioural signals are a highly active and growing area of research [3–6].

Very recent research suggest that depression impacts speech motor control [7, 8]. Depression, similar to many speech motor control disorders [9], can be characterised by prosodic abnormalities, articulatory and phonetic errors [3]. Formants, the dominant components of the speech spectrum, are considered a major marker of speech motor control disorders [10]. Unsurprisingly, there are strong links between alterations in formants' dynamics and depression [8, 11, 12]. Results presented in [11] indicate that speech affected by depression has significantly reduced second formant locations; in particular the diphthong /ɑɪ/. More recent results reveal significant reductions in the *Vowel Space Area* (VSA) measured using the first and the second formant coordinates of the /i/, /ɑ/, and /u/ of speech affected depression [8]. Formant dynamics are used to form the *Vocal Tract Correlation* (VTC) features which have been used to accurately predict an individual's level of clinical depression [13].

Interestingly, preliminary investigations into the similarity and differences between depressed and sleepy speech suggest that the effects of depression on formant features may differ between the genders [12]. This result is supported, in part, by studies which show the usefulness of performing gender dependent classification when using formant and spectral features [14, 15]. Such results are not unexpected; formant distributions are expected to differ between genders due to physiological differences and variations in emotionality [16, 17]. Whilst there are some evidences for differentiation in depression symptoms between men and women (e. g., appetite and weight [18]), possible potential acoustic differences have received very little attention.

Gender-specific emotion recognisers have been shown to perform better than those with mixed gender emotional models [19]. This is also demonstrated by [20] who shows that the combined performance of a gender-dependent recogniser is better than that of a gender-independent. In [17], the authors described a gender-dependent analysis of vowel-level formants for straight-forward classification of emotional arousal.

Herein, we investigate the effects of depression on formant dynamics analysed on a per gender basis. Performing vowel-level formant analysis, we extract a set of gender dependent formant features, and then test their suitability for detecting depression state. This work builds on previous results offered in [21], which demonstrate the potential of using phoneme based features for depression detection. Further, by analysis of the formant features, we also aim to capture effects relating to depression-induced changes in speech motor control [7, 8], to aid the classification of depression affected speech. Earlier phoneme-level analysis provided an outstanding classification performance for cross-corpus emotion recognition [22, 23].

2. Depression Corpus

The corpus used to generate all the experimental results presented in this paper is the *Distress Analysis Interview Corpus - Wizard of Oz* (DAIC-WOZ) database. This corpus has recently been made publicly available as part of the 2016 *Audio-Visual Emotion Challenge and Workshop* (AVEC 2016) [24]. The dataset contains audio-visual recordings of interviews of 189 participants in English with an animated virtual interviewer operated via a Wizard-of-Oz paradigm [25]. Moreover, each participant was assigned a single depression value using the PHQ-8 self-assessed depression questionnaire [26].

As part of the AVEC challenge, the corpus is divided into training, development, and test partitions [24]. All the participants have been assigned into one of two classes *depressed*

http://dx.doi.org/10.21437/Interspeech.2017-887

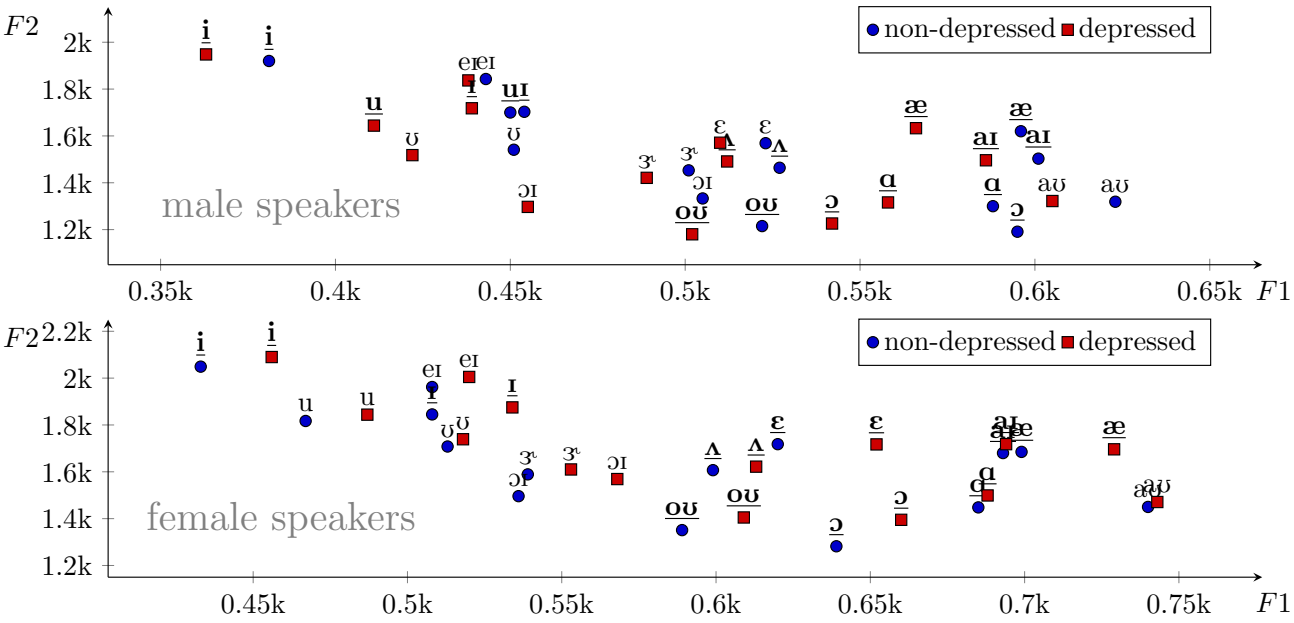

Figure 1: *Positions of average $F1$ and $F2$ values of English vowels in $F1/F2$ [Hz/Hz] space in the training and development partitions of the DAIC-WOZ depression corpus. Abbreviations: $F1$ - first formant, $F2$ - second formant. The formants values for indicative vowels selected by our analysis are underlined.*

Table 1: *Distribution, in terms of gender and depression status, of the participants in the DAIC-WOZ database for training, development and test sets. The total number of participant turns in a particular division is given in parenthesis. Also given is the total length (hh:mm) of each partition*

Gender	Class	Train	Dev.	Test	hh:mm
Male	N-Dep.	55(9018)	12(2085)	18(3333)	10:37
	Dep.	8(1129)	4 (709)	5 (644)	1:44
Female	N-Dep.	32(4786)	16(3029)	20(3978)	9:44
	Dep.	13(1899)	3 (821)	4 (818)	2:33
Total		107	35	47	24:38

(Dep.) and *non-depressed* (N-Dep.) based on the PHQ-8 scores; the average score of each class is 2.75 and 15.9, respectively. The total division of the participants in terms of depression class and gender is given in Table 1.

During our analysis, we split the individual DAIC-WOZ recordings into individual participant turns based on the provided transcriptions. The break-down of the number of turns per gender, per partition is also provided in Table 1. As one could see from the Table 1 dataset contains 189 dialogs with therapist, and 32 249 turns extracted from aligned textual transcriptions provided with the dataset.

A range of state-of-the-art audio classification approaches have been tested on this data as part of AVEC challenge. These include: VTC features [27]; the i-vector paradigm [28]; and a deep neural network which combined both convolutional and Long Short Term Memory (LSTM) layers [29].

3. Vowel-Level Formant Analysis

In the first stage of our evaluation, we automatically estimated the phoneme boundaries using *forced alignment* provided by HTK [30]. Mono-phone *Hidden Markov Models* (HMMs) were trained on acoustic material presented in the DAIC-WOZ corpus. To execute a vowel level analysis, a phoneme level transcription is needed; which requires a corresponding lexicon containing phonetic transcription of words presented in the corpus. As the DAIC-WOZ corpus does not provide such a lexicon, phonetic transcriptions were taken from the CMU Pronouncing Dictionary. Transcriptions for missing words were generated with grapheme to phoneme system (G2P). For estimating phoneme alignment we used 2.5 ms analysis window.

Upon automatic extraction of phoneme borders, we estimate contours for the *first formant* ($F1$) and the *second formant* ($F2$) values. Formant contours were extracted via the Burg algorithm using PRAAT speech analysis software [31]. The following setup for the algorithm was used: the maximum number of formants tracked = 5, the maximum frequency of the highest formant = 6000 Hz, the time step between two consecutive analysis frames = 1 ms, the effective duration of the analysis window = 25 ms, and the amount of pre-emphasis = 50 Hz.

Afterwards, we estimated and took average $F1$ and $F2$ values for each vowel segment presented in the training, development and test samples of the DIAC database. To characterise the changes of the vowels' quality under the influence of the speaker's depressive state, we estimated the mean of the first and the second formants for each vowel (15 vowels in the ARPA-bet non-stressed phoneme set) individually. This resulted in 2x15 = 30 pairs of mean and standard deviations for the average $F1$ and $F2$ values extracted. The random variables which represent the average $F1$ and $F2$ features are approximately normally distributed. Finally, two sets (one per gender) of 10 gender-dependent vowel-level formant features, which are highly

Table 2: *The 10 most indicative gender-dependent formant features extracted on vowel segments from the training and development partitions of the DAIC-WOZ depression corpus. Abbreviations: $F1$ - first formant, $F2$ - second formant. Note, all z-test scores correspond with a significant value of $p < 0.001$*

Gender	ARPA(IPA)[Z-score]
Male	$F1$: AE({æ})[7.99], AO({O})[7.44], IY({i})[6.86], UW({u})[6.41], OW({oU})[5.71], AA({A})[5.70] AY({aI})[5.52], AH({2})[5.51], IH({I})[5.41] $F2$: AH({2})[5.60]
Female	$F1$: IH({I})[11.38], EH({E})[9.57], IY({i})[8.36], AE({æ})[7.53], AH({2})[6.27], OW({oU})[6.02] $F2$: AO({O})[8.80], AY({aI})[7.38], OW({oU})[6.94], AA({A})[6.49]

indicative of the effects of depression in speech, were selected using the z-test (Table 2).

As one can see from Figure 1, the vowel-level mean values for the first and the second formants are different for depressed and non-depressed speech. As expected, the results differ for each gender; for male speakers we see displacement of mean values to the left (i. e., lower $F1$) for depressed speech, as in the case with low-arousal emotional speech described in [17]. In a case of low-arousal detection based on vowel-level formant features, female and males have common tendency: average values for $F1$ are shifted left for indicative vowels. For female speakers, on the other hand, we see an opposite tendency; displacement to the right side (i. e., higher $F1$). This observation forms the basis for our decision to perform gender-dependent analysis for more reliable depression detection analysis.

Considering high level of Z-test scores with significance values $p < 0.001$, we decided to use only utterances which contain indicative vowels. As the result, during the second stage we used just 24185 utterances with indicative vowels out of 32249. Each of 24185 utterances has a different length and number of indicative vowels in transcriptions. For generating fix length vowel-level features, we estimated gender dependent mean values, later called template values, for each indicative vowel for the training, development and test data. The template values were estimated on the speech material of the whole dialogue with therapist. These mean values will be used as template features for each selected utterance. If an utterance contains an indicative vowel, then instead of the template value, we use weighted average of both template (weight equal to 10) and an estimated turn level mean average (weight corresponds to number of indicative vowels in the turn).

4. General Acoustic Features

We compare our proposed features with a general (acoustic-pattern independent) acoustic feature set, *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [32], which has been developed by experts and widely used for paralinguistic tasks such as emotion recognition from speech [24, 33, 34]. The Low-Level Descriptor features and the Functionals (e. g., mean) are listed in the Table 3. Overall, this set provides 102 features.

5. Experimental Settings

The experimental settings (unless otherwise stated) for our classification experiments were as follows: we compare the efficacy of our extracted *vowel level formant features* (VL-Formants) for classifying speech affected by depression with the eGeMAPS audio feature set that been shown to be suitable for a range of paralinguistic classification tasks including depression recognition [33, 34]. We form *turn-level* representations of both feature sets (cf. Table 1). All fusion results reported are for feature fusion i. e., the concatenation of the both feature representations.

Table 3: *Set of Low-Level Descriptor features and the Functionals applied on them, used in the eGeMAPS feature set. (Coef. of Var. = Coefficient of variations)*

LLD	Functional
F_0 (Linear & semi-tone), Loudness	Mean, Coef. of Var., Percentile (20,50,80), Percentile Range (20–80), Mean/Std of Rising/Falling Slope
Spectral Flux, MFCC(1–4), Jitter, Shimmer, Harmonic to Noise Ratio, Harmonic differences, F_1, F_2, F_3 (bandwidth, amplitude, frequency), Voiced sounds: alphaRatio, Hamberg Index, Spectral Slope (0–500Hz, 500–1500Hz), MFCC(1–4)	Mean, Coef. of Var.
(Unvoiced) alphaRatio, Hamberg Index, Spectral Slope (0–500Hz, 500–1500Hz), MFCC(1–4)	Mean
Voiced segments, Loudness Peaks	Per second
(Un)Voiced Segment Length	Mean, std
Equivalent Sound Level	

All results are reported in terms of the AVEC-2016 development and test partitions [24]. For results reported as *development* the systems was trained with data from the training partition only. Whilst for the results reported as *test* the systems was trained with data from both the training partition and development partition. All classifications are performed using the *Liblinear* package [35], with the cost parameter, tuned separately for each experiment via a grid search. Feature standardisation, i. e., subtracting the mean and dividing by the standard deviation, is applied in an online manner. All classification results are reported in terms of F_1*-score*, the challenges official metric [24], for the two classes (*depressed* and *not-depressed*) calculated on a per turn level. Results for the eGeMAPs features are calculated and reported in both gender dependent and independent scenarios. Whilst the VL-Formants, as they are separate feature spaces for each gender (Section 3), are calculated and reported in a gender independent scenario.

Table 4: *Results for depression classification using either eGeMAPS, our gender dependent VL-Formants, and early fusion combination thereof. Performance is given in terms of F_1-score for* depressed *(*not-depressed*) classes. F_1-scores for* depressed*(*not-depressed*) speaker classification generated using the DAIC-WOZ Corpus according to AVEC 2016 development and test conditions. The results are based on gender dependent models. Gender independent testing is not performed on the VL-Fromants as extracted on a per gender basis*

F_1-Score	eGeMaps		VL-Formants		VL-Formants & eGeMaps	
	Gender Ind.	*Gender Dep.*	*Gender Dep.*		*Gender Dep.*	
DEVELOPMENT SET						
Male	0.14 (0.73)	0.48 (0.62)	**0.53**	(0.71)	0.52	(0.70)
Female	0.78 (0.86)	0.83 (0.90)	**1.00**	(1.00)	**1.00**	(1.00)
Overall	0.55 (0.79)	0.65 (0.77)	**0.75**	(0.87)	0.74	(0.86)
TEST SET						
Male	0.07 (0.79)	0.18 (0.64)	0.28	(0.51)	**0.21**	(0.60)
Female	0.65 (0.85)	0.80 (0.93)	**1.00**	(1.00)	**1.00**	(1.00)
Overall	0.44 (0.82)	0.46 (0.80)	0.54	(0.80)	**0.53**	(0.82)

6. Results

The advantages of performing gender dependant depression classification using the eGeMAPS feature sets can be seen in Table 4. Most notably for males, a relative improvement in the depression F_1-score of approximately 200% is seen for both development and test sets, when compared the gender independent and dependent systems. It should be noted that, the male eGeMAPS gender independent test set F_1-score is exceptionally poor.

The VL-Formants consistently outperform the eGeMAPS features, highlighting their suitability for capturing depression information (cf. Table 4). These results provide a strong evidence in support of our decision to perform gender dependent feature extraction and classification. In both the development and test sets, the VL-Features achieve the best possible best F_1-score of 1.00 for both the depressed and non-depressed female classes. This provides strong evidence that key depression information manifests in the formants of female speakers. The difference in performance of the VL-Formants between the genders is larger than expected but not completely unsurprising. Gender difference in formant difference have been reported for both emotional speech [17] and depression [8].

The early fusion of VL-Formants with eGeMAPS does not improve system performance over using VL-Formants alone (cf. Table 4). Given the strong performance of the VL-Formants, especially for females, this is not unexpected. We also tested late fusion of the different feature sets. However, the improvements gained did not outperform the early fusion set-up.

The results of our VL-Formant system is highly competitive when compared to the results published on the DAIC-WOZ corpus (under the conditions of AVEC-2016). Our system easily outperformed the challenge audio baseline which was set using the COVAREP feature representation and a SVM classifier [24]. COVAREP feature set mainly captures voice quality as well as prosodic characteristics of speech [36]. The VL formant depression F_1-scores of 0.75 and 0.54, for the development and test sets, achieved a relative improvement of 63% and 33% over the baselines F_1-score of 0.46 and and 0.41.

They also outperformed the development set depression F_1-score of 0.50 obtained using vocal tract correlation features [27], which are well known to capture depression information in speech [13, 27] (Note that, the test set scores were not given in [27]). The VL-Formants also outperformed the *i-vector* paradigm depression F_1-scores on the development and test sets (0.57 and 0.48, as presented in [28]). Impressively, they match performance with the *DepAudioNet* system presented in [29]. This system feeds spectral features into a network containing two Convolutional layers, a max pooling layer and a Long Short Term Memory (LSTM) layer. With this topology they reported the best F_1-score of 0.52 for both the development and test sets. The strong performance of the VL-formants in comparison with these state-of-the-art systems provides even more support of our decision to perform gender dependent feature extraction and classification.

7. Conclusions

Results presented in this paper indicate the suitability of gender dependant vowel-level formant features for classifying depression from speech. A key finding of our analysis is that the effects of depression may manifest differently in formant measures for male and females. Based on this finding, we extracted two sets of gender dependant vowel level formant features (VL-Formants) which showed promising performance improvement for classifying depression from speech. They outperformed, a range of state-of-the-art approaches including Vocal Tract Correlation features, i-vectors, and a deep neural network. Our results confirm two key findings presented in the literature: firstly, depression manifests at the phoneme level of speech [21]; and secondly, the effects of depression in speech can be captured by features which characterise speech motor control [7, 8].

As future work, we aim to verify these findings on other depression-speech databases. We also plan to explore techniques for performing accurate Automatic Speech Recognition (ASR) on speech affected by depression and complementing acoustic based classification with linguistic features derived from ASR generated transcripts.

8. Acknowledgements

The research leading to these results has received funding the European Union's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

9. References

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Med*, vol. 3, no. 11, pp. 2011–2030, 2006.

[2] M. Blais and L. Baer, "Understanding rating scales and assessment instruments," in *Handb. Clin. Rat. Scales Assess. Psychiatry Ment. Heal.*, ser. Current Clinical Psychiatry, L. Baer and M. Blais, Eds. New York, NY, USA: Humana Press, 2010, pp. 1–6.

[3] 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 Commun.*, vol. 71, pp. 1–49, 2015.

[4] B. N. Cuthbert and T. R. Insel, "Toward the future of psychiatric diagnosis: the seven pillars of RDoC," *BMC Med.*, vol. 11, no. 1, pp. 1–8, 2013.

[5] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image Vis. Comput.*, vol. 32, no. 10, pp. 641–647, Dec 2013.

[6] 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 *Proc. 4th ACM AVEC '14*. Orlando, FL, USA: ACM, 2014, pp. 3–10.

[7] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Commun.*, vol. 75, pp. 27–49, 2015.

[8] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 59–73, Jan 2016.

[9] R. D. Kent and Y. J. Kim, "Toward an acoustic typology of motor speech disorders," *Clin. Linguist. Phon.*, vol. 17, no. 6, pp. 427–445, Jan 2003.

[10] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech," *J. Speech. Lang. Hear. Res.*, vol. 53, no. 1, pp. 114–25, Feb 2010.

[11] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *J. Psychiatr. Res.*, vol. 27, no. 3, pp. 309–319, 1993.

[12] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Automatic modelling of depressed speech: Relevant features and relevance of gender," in *Proc of INTERSPEECH*. Singapore: ISCA, 2014, pp. 1248–1252.

[13] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proc. of the 3rd ACM AVEC'13*. Barcelona, Spain: ACM, 2013, pp. 41–48.

[14] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents; speech during family interactions," *Biomed. Eng. IEEE Trans.*, vol. 58, no. 3, pp. 574–586, 2011.

[15] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "From joyous to clinically depressed: Mood detection using spontaneous speech," in *25th Int. FLAIRS Conf.* Marco Island, FL, USA: AAAI, 2012, pp. 141–146.

[16] A. M. Kring and A. H. Gordon, "Sex differences in emotion: expression, experience, and physiology," *J. Pers. Soc. Psychol.*, vol. 74, no. 3, pp. 686–703, 1998.

[17] B. Vlasenko, D. Prylipko, D. Philippou-Hübner, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions," in *Proc. of INTERSPEECH*. Florence, Italy: ISCA, 2011, pp. 1577–1580.

[18] M. A. Young, W. A. Scheftner, J. Fawcett, and G. L. Klerman, "Gender differences in the clinical features of unipolar major depressive disorder." *J. Nerv. Ment. Dis.*, vol. 178, no. 3, pp. 200–203, 1990.

[19] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Audio, Speech*, vol. 13, no. 2, pp. 293–303, March 2005.

[20] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proc. of the 12th European Sig. Proc. Conf.* IEEE, Sept 2004, pp. 341–344.

[21] A. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, pp. 1–18, 2011.

[22] B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Comput. Speech & Lang.*, vol. 28, no. 2, pp. 483–500, 2014.

[23] B. Vlasenko and A. Wendemuth, "Determining the smallest emotional unit for level of arousal classification," in *Proc.of ACII 2013*, Geneva, Switzerland, 2013, pp. 734–739.

[24] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 - Depression, Mood, and Emotion recognition workshop and challenge," in *Proc. 6th ACM AVEC'16*. Amsterdam, The Netherlands: ACM, 2016, pp. 3–10.

[25] D. DeVault, K. Georgila, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, and L.-P. Morency, "Verbal indicators of psychological distress in interactive dialogue with a virtual human," in *Proc. SigDial*. Metz, France: ACL, 2013, pp. 193–202.

[26] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disord.*, vol. 114, no. 1-3, pp. 163 – 173, 2009.

[27] 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," in *Proc. 6th ACM AVEC'16*. Amsterdam, The Netherlands: ACM, 2016, pp. 11–18.

[28] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proc. 6th ACM AVEC'16*. Amsterdam, The Netherlands: ACM, 2016, pp. 43–50.

[29] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proc. 6th ACM AVEC'16*. Amsterdam, The Netherlands: ACM, 2016, pp. 35–42.

[30] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, *The HTK Book (for HTK Version 3.4)*, Cambridge Uni. Engineering Depart., 2009.

[31] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2002.

[32] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr 2016.

[33] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015: the first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. 5th ACM AVEC'15*. Brisbane, QLD, Australia: ACM, 2015, pp. 3–8.

[34] B. Stasak, J. Epps, N. Cummins, and R. Goecke, "An investigation of emotional speech in depression classification," in *Proc. of INTERSPEECH*. San Francsico, CA, USA: ISCA, 2016, pp. 485–489.

[35] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Aug 2008.

[36] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP – a collaborative voice analysis repository for speech technologies," in *Proc. of ICASSP*. Florence, Italy: IEEE, 2014, pp. 960–964.