Enhancing speech-based depression detection through gender dependent vowel-level formant features

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

[1] Chair of Complex & Intelligent Systems, University of Passau, Germany
[2] Department of Computing, Imperial College London, U.K.
nicholas.cummins@uni-passau.de

Abstract. Depression has been consistently linked with alterations in speech motor control characterised by changes in formant dynamics. However, potential differences in the manifestation of depression between male and female speech have not been fully realised or explored. This paper considers speech-based depression classification using gender dependant features and classifiers. Presented key observations reveal gender differences in the effect of depression on vowel-level formant features. Considering this observation, we also show that a small set of handcrafted gender dependent formant features can outperform acoustic-only based features (on two state-of-the-art acoustic features sets) when performing two-class (depressed and non-depressed) classification.

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

1 Introduction

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 is a highly active and growing area of research [1]. In this regard, possible key markers of depression are changes in paralinguisitic cues [1]. Formant features, representing the dominant components in the speech spectrum and capturing information on the resonance properties of the vocal tract, are one such feature [2–4]. They are strongly linked to changes in speech motor control associated with depression [2].

Whilst there is evidence for differentiation in depression symptoms between men and women (e. g., appetite and weight [5]), possible potential acoustic differences have received very little attention. Investigations into the similarity and differences between speech affected by depression or fatigue suggest that the effect of depression on formant features may differ between the genders [3]. This result is supported, in part, by studies which show the usefulness of performing gender dependent classification when using formant and spectral features [4]. Such results are not unexpected; formant distributions should differ between genders due to physiological differences and variations in emotionality [6, 7].

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. The presented results are generated from a subset of *The Distress Analysis Interview Corpus – Wizard of Oz* (DAIC-WOZ) database [8], containing speech from 12 males and 16 females clinically diagnosed with depression, as assessed by the widely used *Patient Health Questionnaire* (PHQ-8) self-assessed depression questionnaire. As a control group, speech from 67 males and 47 females without clinical depression is also provided.

2 Vowel-Level Formant Analysis

In the first stage of our evaluation, we automatically estimated the phoneme boundaries. These were determined, using *forced alignment* provided by *HTK*[3]. 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 a 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*.

Upon automatic extraction of phoneme borders, we estimate the average of the *first formant* ($F1$) and the *second formant* ($F2$) values for each vowel instance. Formant contours were extracted via the Burg algorithm using *PRAAT* [9]. The following setup was used: the maximum number of formants tracked was five, the maximum frequency of the highest formant = 6 k Hz, the effective duration of the analysis window = 25 ms, the time step between two consecutive analysis frames = 10 ms, and the amount of pre-emphasis = 50 Hz.

As can be seen in Figure 1, the vowel-level mean values for the $F1$ and $F2$ 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 [7]. 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.

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 ARPAbet non-stressed phoneme set) individually. This resulted in $2 \times 15 = 30$ pairs of mean and standard deviations for average $F1$ and $F2$ values extracted. The random variables which represent 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 indicative of the effects of depression in speech, were selected using the z-test; these vowels are underlined in Figure 1.

[3] `http://htk.eng.cam.ac.uk/`

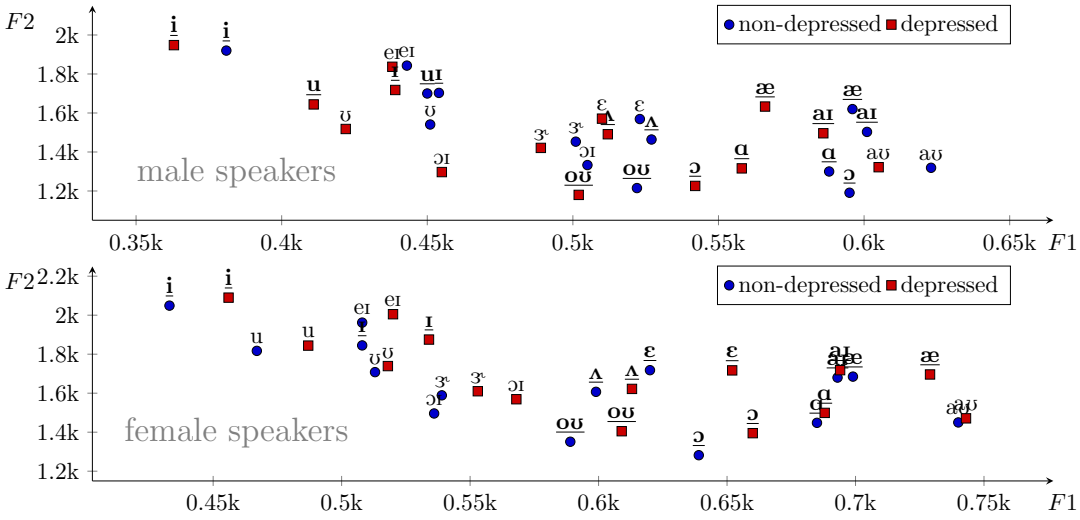

Fig. 1. Positions of average $F1$ and $F2$ values of English vowels in $F1/F2$ [Hz/Hz] space in the training and development partitions of the DAIC-WOZ depression corpus. Abbreviations: $F1$ – first formant, $F2$ – second formant. The formants values for indicative vowels selected by our analysis are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

3 Classification Experiments

3.1 Set-Up

We compared the efficacy of our extracted *vowel-level formant features* (VL-Formants) for classifying speech affected by depression with two commonly used audio feature sets: the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [10], and the COVAREP feature set [11]. Both feature representations are extracted at the *turn-level*. The VL-Formants are extracted on a *participant-level* as per Section 2. With the turn-level eGeMAPS and COVAREP features, we simply concatenate the per participant features with the matching turn-level features and treat this new feature representation as an *enhanced* turn-level feature.

All classifications are performed using the *Liblinear* package [12], with the cost parameter tuned separately for each experiment via a grid search. As the number of depressed participants is very low, we perform a series of speaker independent 4-fold participant-based cross-validation tests (make-up of folds available upon request), noting we use 4 as it is a divisor of the total number of depressed participants available.

All results are reported in terms of F_1*-score* for both the *depressed* and *non-depressed* classes. Results for the eGeMAPs and COVAREP features performed in both a *Gender Independent* (classifier trained with instances from both genders) or *Gender Dependent* (classifier trained with only instances from the target gender) scenario. As the VL-Formants representations differ for each gender (cf. Section 2), the results for this feature set and for feature fusion are reported for the *Gender Independent* scenario only.

Table 1. Results for depression classification using either eGeMAPS, COVAREP, our gender dependent VL-Formants, and early fusion combination thereof. Performance is given in terms of F_1-score for *depressed* (*not-depressed*) classes. Scores are the average F_1-score of four-fold cross validation on the training and development partitions of the DAIC-WOZ Corpus. Gender dependent: GD. Note, gender independent (GI) testing is not performed on the VL-Formants as extracted on a per gender basis

F_1	eGeMaps		COVAREP		VL-Formants	VL-Formants & COVAREP	VL-Formants & eGeMaps	
	GI	*GD*	*GI*	*GD*	*GD*	*GD*	*GD*	
Male	.07 (.67)	.13 (.46)	.17 (.51)	.08 (.91)	.37 (.81)	.42 (.87)	**.49**	(.85)
Female	.19 (.45)	.23 (.44)	.33 (.21)	.28 (.20)	.55 (.86)	.52 (.62)	**.55**	(.80)
Overall	.15 (.68)	.36 (.71)	.26 (.41)	.28 (.69)	.38 (.73)	.45 (.75)	**.63**	(.89)

3.2 Results

Results from our classification analysis indicate that performance gains can be found by performing gender dependent depression classification while using eGeMAPS features (cf. Table 1). These results provide support for our decision to extract the VL-formant features on a per-gender basis. The advantages of gender dependent classifiers when using COVAREP are not as obvious; the male gender dependent results are in particular poor (cf. Table 1). The weaker performance of the COVAREP in the gender dependent setting is not unexpected; the voice quality features in COVAREP have been shown to be gender independent in relation to detecting depression [13].

VL-Formants perform the strongest out of the feature sets tested (cf. Table 1), highlighting their suitability for capturing depression information. The results provide a strong evidence in support of our decision to perform gender dependent feature extraction and classification. As indicated in Table 1, the early fusion of VL-Formants with the other two feature sets improves the overall F_1 scores for depression when compared to the individual feature set alone. The biggest gain was obtained when combining VL-Formants and eGeMAPS. The early fusion of all feature sets did not result in any further improvements in system accuracy (result not given). We also tested late fusion of the different feature sets; however, the improvements gained did not outperform the early fusion set-up.

4 Conclusions

This paper investigated the effects of depression on formant dynamics analysed on a per gender basis. Our analysis indicates that, indeed, 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 which showed promising performance improvement for classifying depression from speech. This result matches with two key results presented in the literature: firstly, depression manifests at the phoneme level of speech [14]; and secondly, the effects of depression in speech can be captured by features which characterise speech motor control [2, 15]. In future work, we aim to verify these findings on other depression-speech databases.

5 Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and IMI RADAR-CNS under grant agreement No. 115902.

References

1. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.: A Review of Depression and Suicide Risk Assessment using Speech Analysis. Speech Comm. 71, 1–49 (2015).
2. Scherer, S., Lucas, G.M., Gratch, J., Rizzo, A.S., Morency, L-P.: Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews. IEEE Trans. Affect. Comput. 7, 59–73 (2016).
3. Hönig, F., Batliner, A., Nöth, E., Schnieder, S., Krajewski, J.: Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender. Proc. of INTERSPEECH. p. 1248–1252, ISCA, Singapore (2014).
4. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G.: From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech. Proc. of FLAIRS. p. 141–146, AAAI, Marco Island, FL, USA (2012).
5. Young, M.A., Scheftner, W.A., Fawcett, J., and Klerman, G.L.: Gender differences in the clinical features of unipolar major depressive disorder. J. of Nerv. and Ment. Dis. 178, 3, 200–203 (1990).
6. Kring, A.M., Gordon A.H.: Sex differences in emotion: expression, experience, and physiology. J. Pers. Soc. Psychol 74, 3, 686–703 (1998).
7. Vlasenko, B., Prylipko, D., Philippou-Hübner, D. Wendemuth, A.: Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. Proc. of INTERSPEECH. p 1577–1580, ISCA, Florence, Italy (2011).
8. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres, M.T., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge. Proc. 6th ACM Int. Work. Audio/Visual Emot. Chall. p 3–10, ACM, Amsterdam, Netherlands (2016).
9. Boersma, P., Weenink, D.S.: Praat, a system for doing phonetics by computer. Glot Int. 5, 9/10, 341–345 (2002).
10. Eyben, F. and Scherer, K. R. and Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans. Affect. Comput. 7, 190–202 (2016).
11. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP - A collaborative voice analysis repository for speech technologies. In: Proc. of ICASSP. p 960–964, IEEE, Florence, Italy (2014).
12. Rong-En, F., Chang, K-W., Hsieh, C-J., Wang, X-R., Lin, C-J.: LIBLINEAR: A Library for Large Linear Classification. J. Mach. Learn. Res. 9, 1871–1874 (2008).
13. Scherer, S., Stratou, G., Gratch, J., Morency, L-P.: Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD. Proc. of INTERSPEECH. p 847–851, ISCA, Lyon, France (2013).
14. Trevino, A., Quatieri, T., Malyska, N.: Phonologically-based biomarkers for major depressive disorder. EURASIP J. Adv. Sig. Proc. 2011, 1–18 (2011).
15. Cummins, N., Sethu, V., Epps, J., Schnieder, S., Krajewski, J.: Analysis of acoustic space variability in speech affected by depression. Speech Comm. 75, 27–49 (2015).