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

Cross-Cultural Depression Recognition from Vocal Biomarkers

Sharifa Alghowinem[1,5], *Roland Goecke*[2,1], *Julien Epps*[3], *Michael Wagner*[2,1], *Jeffrey Cohn*[4],

[1]Australian National University, Research School of Computer Science, Australia
[2]University of Canberra, Human-Centred Technology Research Centre, Australia
[3]University of New South Wales, Australia
[4]University of Pittsburgh, USA
[5]Prince Sultan University, College of Computer & Information Sciences, Saudi Arabia

sghowinem@psu.edu.sa, roland.goecke@ieee.org, j.epps@unsw.edu.au,
michael.wagner@canberra.edu.au, jeffcohn@pitt.edu

Abstract

No studies have investigated cross-cultural and cross-language characteristics of depressed speech. We investigated the generalisability of a vocal biomarker-based approach to depression detection in clinical interviews recorded in three countries (Australia, the USA and Germany), two languages (German and English) and different accents (Australian and American). Several approaches to training and testing within and between datasets were evaluated. Using the same experimental protocol separately within each dataset, (cross-classification) accuracy was high. Combining datasets, high accuracy was high again and consistent across language, recording environment, and culture. Training and testing between datasets, however, attenuated accuracy. These finding emphasize the importance of heterogeneous training sets for robust depression detection.

1. Introduction

Clinical depression is estimated to be the leading cause of suffering and disability worldwide in 2020 [1]. One of the difficulties of diagnosing depression is that it depends on subjects' report and clinical opinion, which risks a range of subjective biases. Objective assessment could be obtained with the utilisation of developments in affective sensing technology. Ultimately, we want to develop an objective affective sensing system that supports clinicians in their diagnosis and monitoring of clinical depression across different countries.

Studies in depression detection have mainly investigated single datasets. When using a single dataset, many intervening variables are kept constant, such as recording settings and environment. Therefore, when using different corpora to generalise a depression recognition system, results might not be comparable due to these variables. In particular, speech analysis is extremely affected by recording environment, such as varying room acoustics, and different microphone types and distances [2]. Moreover, different languages and accents of the speakers, number and labels of the classes, and collecting procedure could introduce other variabilities when using multiple datasets. Therefore, when different datasets are used, several equalisation and normalisation methods have to be considered to control for variability between the datasets.

This study investigates generalising an approach to detect depression from verbal biomarkers in a cross-cultural context. We use three different depression datasets from different countries and languages, where we attempt to control for their differences. The approach extracts and normalises functional speech features to reduce the effect of different recording characteristics. For the generalisation investigation, we evaluate the depression classification from each of the three corpora individually, and in different combinations. When applied on individual datasets, we hypothesise that the experimental approach is data-independent. When using different combinations of datasets, we hypothesise that training over varied samples and reducing model overfitting to the training set will increase the generalisability of the approach across the datasets. Studies analysing depressed speech have mainly been utilising one dataset of one culture with one language. In this study, we evaluate a cross-cultural, cross-language, and cross-corpora depression detection approach from verbal biomarkers.

2. Background

Recently, a few studies investigated developing a system that automatically detects depression from either audio or video input, or multimodal input. Speech analysis has been investigated in several studies using speech prosody (e.g. [3]) and speech style (e.g. [4]). Several studies have found distinguishable prosody features such as pitch, loudness, energy, formants, jitter, shimmer and HNR (Harmonic-Noise-Ratio) [3].

In general affect studies, cross-corpus generalisation is a very young research area. To the best of our knowledge, only a few studies have investigated method robustness on different environments [2, 5, 6]. Speech in particular is immensely affected by the recording environment, due to varying room acoustics and different types of and distance to the microphones [2]. In general, due to dataset differences (e.g. class labels and numbers, recording conditions and procedure), generalising a system to a new dataset results in lower accuracy than for the original data (e.g. [2]). Thus, normalisation methods have to be applied to eliminate recording environment differences [2].

Cross-corpus generalisation in depression detection is particularly challenging to investigate. Acquiring and sharing depression datasets involves ethical, clinical, and legal procedures. Differences in recording environment, recording procedure and depression evaluation add another challenge in a cross-corpus generalisation study. To the best of our knowledge, three studies that used more than one dataset in their analysis of depression speech are [7, 8, 9]. In these studies, preprocessing and normalisation procedures were performed to reduce the possible differences in recordings. In [7, 8], the authors raised the concerns of differences affecting the analysis results. Moreover, in [7, 8], each database was treated as a different class. The separation

http://dx.doi.org/10.21437/Interspeech.2016-1339

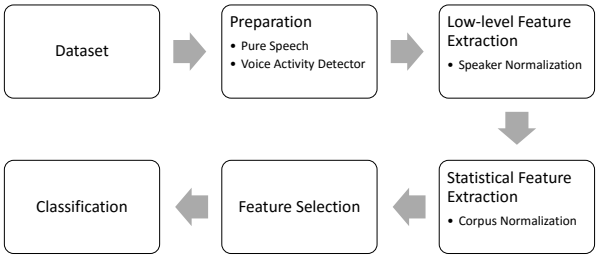

Figure 1: Approach to classify severe depressed from low-/non-depressed subjects

of each dataset as one class could result in classifying the differences in their recording environment characteristics, not the actual class label. In [9], two depression datasets were used to analyse the variability of acoustic space of depression speech, where no classification was performed. The two datasets were not combined, but the analysis results of the individual datasets were compared [9].

Previously, we investigated *nonverbal behaviour* (eye and head movements) for depression detection in a cross-cultural context using three depression datasets [10]. The nonverbal analysis and classification were able to generalise on cross-corpus experiments and achieved similar performance on all three datasets. Here, we analyse and generalise *verbal* biomarkers to detect depression from three different datasets.

3. Method

In this section, a brief overview of the datasets and the approach used to investigate the cross-dataset generalisation of depression detection is given. Table 1 summarises the datasets and Figure 1 shows the general design and individual components.

3.1. Depression Datasets

For the purpose of cross-corpus generalisation, three depression datasets were used: Black Dog Institute depression dataset (BlackDog) [4], University of Pittsburgh depression dataset (Pitt) [11], and Audio/Visual Emotion Challenge depression dataset (AVEC) [12]. The specifications and differences of these datasets are summarised in Table 1.

As can be seen in Table 1, the three datasets differ in various characteristics that could affect the generalisation investigation. Therefore, we control for these differences by (cf. [10]):

- equalising depression measurement scores,
- forming a control/low-depression class and a severe depression class for each database (i.e. we categorise control and low depressed subjects as one group when combining datasets),
- selecting spontaneous speech from each dataset collection procedure,
- selecting one session per subject when multiple sessions are available, and
- extracting statistical functional features to overcome the variations in speech durations between each subject in each dataset.

Moreover, the audio channel in particular is more vulnerable to the recording environment than the video channel (e.g. microphone distance, background noise, sampling rate). Therefore, to reduce the effect of the differences in recording environments, we equalise the sampling rate and normalise the extracted features (as detailed in Section 3.3).

3.2. Audio Preparation and Pre-processing

In order to extract speech features accurately, the subjects' speech should be isolated from other sounds such as noises or other speakers such as the interviewer. In order to extract pure subject speech accurately a manual annotation approach was used for BlackDog [4] and a manual transcription was used for Pitt [11]. Furthermore, speech annotation and speaker separation could be performed automatically using advanced speaker diarisation techniques (automatic feature extraction was not the focus of this study). As the AVEC dataset involves only one speaker in each recording, no separation is required to extract pure subject speech. Once the pure subject speech is extracted, it passes through several speech pre-processing steps such as framing, windowing, and segmenting. For all speech signals, the acoustic features are extracted with frame size set to 25ms at a shift of 10ms and using a Hamming window. In this work, voice activity detection (VAD) is applied to the pure subject speech segments using Praat software [14].

3.3. Feature Extraction and Normalisation

Speech features can be acquired from both uttered words (linguistic) and acoustic cues (para-linguistic). However, linguistic features, including sementics, word choices, sentence structure etc., are not in the scope of this research, since it would conflict with the generalisation goal of this work as we investigate both English and German languages. Prosody features were extracted from sounding segments, which can also be categorised into two branches: low-level descriptors (LLD) and statistical functionals. To extract low-level features, the publicly available "openSMILE" software was used [15]. The most common features in the depression detection literature from the fields of psychology and affective computing were extracted as follows: the fundamental frequency (F0), energy, intensity, loudness, jitter, shimmer, Harmonic-to-Noise-Ratio (HNR), voice probability and quality, first three formants, and MFCCs. The first (Δ) and second ($\Delta\Delta$) derivatives of each LLD feature were also extracted. The total number of extracted features were 84 per frame. For each subject, the extracted low-level features were normalised using Z-score normalisation to reduce recording setting differences between subjects. This method of speaker normalisation is widely used in speaker recognition tasks to reduce the variations in speech signal [2].

To calculate functional features, several statistical measures were applied to the normalised low-level features. The statistical functionals include mean, minimum, maximum, and range, which gives a total of 504 functional features. For each dataset, the extracted functional features were normalised using Min-Max normalisation before usage in combination with other corpora (corpus normalisation). These 504 normalised functional features form one observation for each subject. That is, the LLD features are not used in the classification process, but used in the extraction of the functional features.

3.4. Classification and Evaluation

For classification, we used Support Vector Machine (SVM) classifiers, applied in a binary (i.e. severe depressed vs. low-/non-depressed) subject-independent scenario. LibSVM [16]

Table 1: Summary of the three datasets specification used in this research

Dataset	BlackDog	Pitt	AVEC
Language	English (Australian)	English (American)	German
Classification	Severely Depressed/Healthy Control	Severe/Low depression	Severe/Low depression
Number of subjects per class	30	19	16
Males-Females	30-30	14-24	9-23
Procedure	open ended questions interview	HRSD clinical interview	human-computer interaction experiment (story telling)
Symptom severity measure	QIDS-SR	HRSD	BDI
Mean score (range)	Severe:19 (14-26) / Healthy Control	Severe:22.4 (17-35) / Low:2.9 (1-7)	Severe:35.9 (30-45) / Low:0.6 (0-3)
Equivalent QIDS-SR Score [13]	Severe:19 (14-26) / Healthy Control	Severe:17 (13-26) / Low:2 (1-5)	Severe:20 (16-22) / Low:1 (0-2)
Total Duration (minutes)	509	355.9	33.2
Pure subject speech (minutes)	119.3	92.0	23.9
Hardware	1 camera + 1 microphone	4 cameras + 2 microphones	1 web camera + 1 microphone
Audio sampling rate	44100 Hz	48000 Hz	44100 Hz
Audio preparation	manual labeling	manual transcription	none (no human interviewer)

was used for SVM implementation using a radial basis function (RBF) kernel. To increase the accuracy of SVMs, the cost and gamma parameters were optimised via a wide range grid search for the best parameters. The performance of the approach was measured in terms of Average Recall (AR), which considers the correct recognition in both groups (severe depressed vs. low-/non-depressed). To shed more light onto cross-corpora generalisation, we investigate two methods for selecting the training and testing sets for classification:

Leave-one-subject-out cross-validation (LOSO): This method was used in the classification of individual datasets, as well as the combinations of the datasets, without any overlap between training and testing data. For individual datasets classification, this method is beneficial to mitigate the limitations of the relatively small number of subjects. When using different dataset combinations, this method is beneficial to train the classifiers on varied observations. This method could overcome overfitting the classifier model on the training set and, therefore, assist in generalising to different observations in each leave-one-out cross-validation turn.

Separate train-test dataset: In this method, one or two datasets (from BlackDog, Pitt and/or AVEC) were used for training and then the remaining dataset(s) for testing. We applied this method to investigate the generalisability of the depression detection method to unseen dataset characteristics. However, this method could suffer from overfitting to the training set and might not generalise to the completely different testing set(s).

3.5. Feature Selection

Feature selection techniques select a subset of features to reduce irrelevant features that could affect the classification performance using statistical function methods, filter methods, search strategies, etc. Moreover, simple statistical tests such as a t-test could be used to evaluate the significance of individual features for selection, especially for two-class classification (e.g. [17]). In this work, a simple T-test threshold was used on the extracted 504 normalised functional features to perform feature reduction, which is suitable for the binary classification of this work (depressed vs. low-/non-depressed). That is, only features that showed a statistically significant difference between the means of the two classes were selected. The T-tests were obtained as a two-sample two-tailed T-test, assuming unequal variances with significance $p = 0.05$. Features that exceeded a statistical threshold set in advance by a t-value corresponding to an uncorrected p-value of 0.05 ($p < 0.05$) (named ETF) were selected for the classification problem. With leave-one-subject-out cross-validation, features that exceeded the t-statistic in the training turn were selected on the testing data. Similarly, in the train-test classification method, the features that exceeded the t-statistic in the training set were selected on the testing set. Out of the extracted 504 functional features, the average number of functional features selected was 64 features ($\sigma = 25$).

4. Results

The results of applying the approach on the **three datasets individually** are presented in Figure 2 (the first three bars). Using statistical functional features from speech, classification results were similar in the three datasets individually. Speech features performed best in the AVEC dataset (97%), and equally in the BlackDog and Pitt datasets (82%). The high performance in the classification results in the three datasets might be due to the clear distinction between severely depressed and low-/non-depressed participants, as the gap between depression scores for the two groups is very wide, see Table 1. Such differences in depression severity might have a distinct effect on the patients' vocal cords. Vocal cord dysfunction has been associated with multiple psychological conditions, including major depression [9]. Moreover, the high performance achieved by speech features on the individual datasets suggests that the experimental protocol used has the ability to detect depression regardless of the dataset characteristics. This finding implies that the approach is data-independent and, therefore, has generalisability aspect.

Following the success of applying the approach to the datasets individually, we attempt to apply it to different combinations of the three datasets. We acknowledge the differences in the datasets, which could have a large effect on the classification results when combining different datasets. However, for the purpose of the generalisability investigation, we evaluate depression classification results of different training and testing dataset combinations. This is performed in two methods: (1) Leave-one-subject-out cross-validation, and (2) train-test classification, as described in Section 3.4. We hypothesise that LOSO cross-validation results in a higher performance than the train-test method, because it trains over varied samples, which reduces model overfitting to the training set. The results of these two methods are presented in Figure 2 (bar #4-13).

Using different **combinations of the three datasets**, the approach was applied to validate its generalisability in **LOSO cross-validation**. The results are presented in Figure 2 (bar #4-7). In general, the classification results on dataset combinations in LOSO were considerably high (on average at 79%AR), even with the dataset differences. We believe that this is due to the classifier learning from varied observations from each dataset, which therefore reduces the effect of overfitting the model to

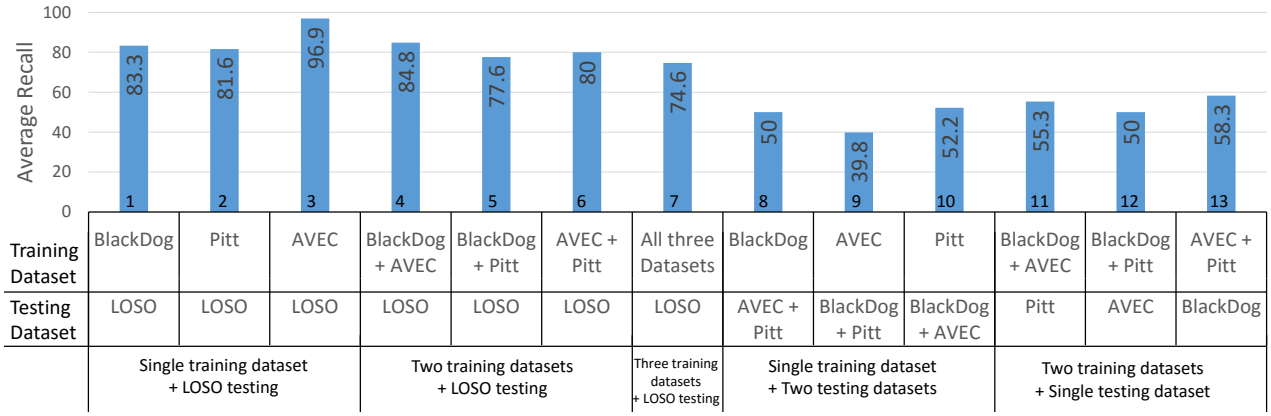

Figure 2: Average recall of classification results of individual datasets and dataset combinations.

specific observation conditions.

Classification results (bar #4-7) show no improvement compared with their individual datasets classification results (bar #1-3) in Figure 2, with only one exception. This exception is when combining BlackDog with AVEC datasets, where the result had a slight improvement (1.5% absolute) over the individual BlackDog classification results. Given the differences between the datasets, a reduction or at least no improvement from classification results when different datasets combinations are used was expected. Nevertheless, the classification results from the combined datasets in leave-one-out cross-validation were statistically above chance level. We believe that the classification result was not affected by the language differences, as when the AVEC dataset combined with the BlackDog or Pitt datasets resulted in a high performance despite their different languages. The same applies to signal quality differences.

Beside using the LOSO cross-validation method, the **train-test method of dataset combinations** for generalisation investigation was also used. In this method, one or two datasets were used for training and the remaining dataset(s) for testing. The classification results of generalisation using the train-test method are illustrated in Figure 2 (bar #8-13). In general, the classification results when using train-test method are mostly at or lower than chance level. That is expected as, unlike in the leave-one-out crossvalidation method with dataset combinations, the classifier on the train-test method is trained on observations of the training dataset(s). Therefore, the created classification model contains certain characteristics of the training dataset(s), which risks overfitting to the training dataset(s). The overfitting issue reduces the classifier's ability to generalise to separate and different dataset observations (unseen data). Moreover, data mismatch could also be the reason behind the drop of the classification results in the train-test dataset method. However, data mismatch did not have the same effect in the LOSO in dataset combinations, where a strong depression recognition was obtained. Therefore, future work could investigate different normalization techniques which may improve performance of the cross-corpora experiments.

In our cross-cultural depression recognition investigation, the finding of this study on verbal biomarkers is in line with our previous study on nonverbal behaviour [10]. The generalisability of the approach could be caused by the normalised statistical features, that reduced the dataset differences. It could also be due to the similarity of the depression symptoms in Western societies (American, Australian and German) [18]. Nevertheless, these findings shed more light onto the need for observations from varied dataset characteristics for a generalised system to detect depression when the approach is exported to the real clinical environment.

5. Conclusions and Future Work

The work presented here aimed at a generalised and objective diagnostic aid to support clinicians in diagnosing depression. To the best of our knowledge, depressed speech analysis has mainly been investigated using one dataset with one language, and no studies exist on cross-cultural and cross-language characteristics of depressed speech. Using verbal biomarkers, we investigated the generalisability of an approach to detect depression cross-culturally from Australia, the USA and Germany. We hypothesised that if the experimental approach is data-independent, it could generalise to different dataset combinations, as long as the classifier is trained over varied samples to reduce model overfitting to the training set. The results confirmed our hypotheses of the generalisability of the approach to training and testing within and between datasets, even with the several differences between the datasets. Using LOSO cross-validation, both individual and different dataset combinations gave considerably high classification results in detecting severe depression. The high and comparable classification results from the individual datasets implies a data-independent aspect of the experimental design. On the other hand, the strong classification results with different combinations of the three datasets implies that the differences in language, recording environment and culture did not influence the ability of the approach to recognise depression and thus demonstrates the generalisability of the approach. When using one or two datasets for training and the rest for testing, the depression recognition drops. This finding emphasises the need for observations from varied dataset characteristics , when the approach is exported to the real clinical environment, to design a generalised depression detection system. Future work should extend the analysis to include datasets from non-western cultures, e.g. Arabic cultures, to further investigate cultural differences in depression expression.

6. Acknowledgements

Funded in part by ARC grant DPI 30101094.

7. References

[1] W. World Health Organization, *The world health report 2003: shaping the future*. World Health Organization, 2003.

[2] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, July 2010.

[3] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Characterising depressed speech for classification." in *INTERSPEECH*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 2534–2538.

[4] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "From joyous to clinically depressed: Mood detection using spontaneous speech." in *FLAIRS Conference*, G. M. Youngblood and P. M. McCarthy, Eds. AAAI Press, 2012, pp. 141–146.

[5] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *INTERSPEECH*, 2011, pp. 1553–1556.

[6] I. Lefter, L. Rothkrantz, P. Wiggers, and D. van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horak, I. Kopeek, and K. Pala, Eds. Springer Berlin Heidelberg, 2010, vol. 6231, pp. 353–360.

[7] D. France, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, July 2000.

[8] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, Sept 2004.

[9] 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, no. C, pp. 27–49, Dec. 2015.

[10] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1, May 2015, pp. 1–8.

[11] Y. Yang, C. Fairbairn, and J. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, April 2013.

[12] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: ACM, 2013, pp. 3–10.

[13] Depression Scores Conversion. (Online) Inventory of Depressive Symptomatology (IDS) & Quick Inventory of Depressive Symptomatology (QIDS).

[14] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2009.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462.

[16] C. C. Chang and C. J. Lin, "Libsvm: a library for svm," *2006-03-04]. http:// www. csic. ntu. edu. tw/rcjlin/papers/lib. svm*, 2001.

[17] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[18] K. Singer, "Depressive disorders from a transcultural perspective," *Social Science & Medicine (1967)*, vol. 9, no. 6, pp. 289 – 301, 1975.