Mitigating Confounding Factors in Depression Detection Using an Unsupervised Clustering Approach

Michelle Renee Morales
The Graduate Center, CUNY
365 5th Ave, New York, NY 10016
mmorales@gradcenter.cuny.edu

Rivka Levitan
Brooklyn College, CUNY
2900 Bedford Ave, Brooklyn, NY 11210
levitan@sci.brooklyn.cuny.edu

ABSTRACT

This work focuses on using speech as an objective marker for depression detection. One of the major challenges of this task is the presence of confounding factors, such as gender, age, emotion and personality. This work presents a technique to mitigate such factors by using a multi-step approach that performs unsupervised clustering prior to depression classification.

Author Keywords

Depression detection; speech; variability factors; unsupervised clustering

INTRODUCTION

Depression is a psychiatric mood disorder. At a global level, an estimated 350 million people of all ages suffer from depression [1]. In the United States, depression affects approximately 14.8 million adults, or about 6.7 percent of the U.S. population age 18 and older [2]. Depression is caused by a persons inability to cope with certain stressful events and situations. Depressed individuals often experience feelings of sadness or negativity, and have trouble coping with everyday responsibilities. Due to the variation in how depression presents itself within each person, it is difficult and time consuming to diagnose. Moreover, since diagnosis often relies on a clinician's assessment it is subjective. In addition, many undeserved regions have severe shortages of clinicians who can make this diagnosis. Many studies have documented the relationship between objective acoustical measures of voice and speech, and clinical subjective ratings of severity of depression. If this relationship is robust enough, voice acoustical analysis can serve as a powerful tool for depression detection to be used in complement with existing diagnostic measures, such as clinical interviews and self-reports.

The speech production system of a human is very complex. Lenneberg [12] estimated that more than 100 independently innervated muscles are coordinated in the tongue and mouth during speech. As a result of the systems complexity, speech is sensitive. Some have argued that slight physiological and cognitive changes produce acoustic changes in speech [15]. We will adopt the same hypothesis as others who pursue this line of inquiry and assume that depression produces cognitive and physiological changes that influence speech production, leading to a change in the acoustic quality of the speech produced. This change can then be measured and objectively evaluated. Speech is an attractive candidate for a diagnostic because it is cheap to collect and non-invasive. Moreover, it is objective. Speech processing research has investigated how we may automate the process of using speech as a diagnostic for depression. Although previous research has made great progress in understanding what models are most suitable for automatically predicting severity level of depression, there is a lack of exploration into dealing with sources of variability, which can significantly confound results.

In general, when eliciting speech as a marker for depression the following confounding factors complicate the task: biological traits such as gender, cultural traits such as dialect, and emotional signals such as fear and anger. These variability factors place a ceiling on the accuracy of a speech based system for depression detection. Given this potential limitation, it is important to research ways to mitigate these factors. This work presents an approach to deal with confounding factors by utilizing a two-layer architecture. To tease apart the traits/states of the speakers involved, we first perform unsupervised clustering using a K-means algorithm. We then perform depression detection on each of the clusters separately and find that clustering prior to classification can help boost performance.

RELATED WORK

Some work has investigated mitigating confounding factors, such as speaker characteristics, phonetic content, and recording setup variability. Cummins et al. [4] based their work on findings from emotion recognition research, hypothesizing that accurate selection of speech segments would provide maximal depressed/neutral speech discrimination [4]. They expected to find that voiced segments provided the most effective discrimination. In addition, they explored normalization techniques, such as mean and mean-variance normalization as well as feature warping, which attempts to reduce variation in data due to differences in speaker variability. They found that discriminating between voiced/voiceless speech segments was not critical to task. Mean and mean-variance normalization techniques were not reported due to their very poor performance and feature warping as a per-speaker fea-

ture space normalization technique offered little to no improvement.

In later work, Cummins et al. [6] provide an analysis of the Audio Visual Emotion recognition Challenge (AVEC) speech corpus [19]. They analyze the phonetic variability of the data by generating multiple sub-utterances per file. They then show that each sub-utterance differs vastly in phonetic content, by demonstrating that there exist a wide range of prediction scores for each file. Their analysis provided insight into the phonetic variability that exists across a depressed speech utterance.

Cummins et al. [5] investigate acoustic volume proposing a novel GMM-based measure that is able to capture the decreasing spectral variability that is usually associated with depressed speech. Using this approach they are able to show that with increasing levels of depression the MFCC feature space narrows to become more tightly concentrated.

Some researchers have borrowed techniques from speech/speaker recognition, which have helped mitigate confounding forms of variability. Sturim et al. [16] were able to reduce the effects of speaker and intersession variability by using a Weiner Filtering Factor Analysis method to enhance a MFCC-based GMM system. In addition, they found that using a 2-class gender independent set up resulted in a reduction in Equal Error Rate of $\sim$21% and $\sim$29% for the male and female systems when compared to one single model for both genders. Sturim et al.'s findings demonstrate the influence gender differences have on a system.

Other interesting approaches have performed analyses of the relationships that exist between different symptoms of depression and different prosodic and acoustic features. Some have even found significantly stronger correlations between their measures on individual items on the HAMD[1], such as low mood when compared to the total HAMD score [17, 13, 9].

The literature discussed above serves to demonstrate the variability inherent in depressed speech as well as highlight the importance in dealing with this variability. Overall, previous work has aimed to mitigate confounding factors, by exploring different normalization techniques, statistical models, and architectures. This work builds upon previous work by introducing a multi-layer architecture, which involves unsupervised clustering. This technique is also borrowed from work in speaker identification. Clustering has proven to be a successful technique in segmenting speakers without any prior knowledge of the identities or the number of speakers [11, 10]. Here, clustering provides a way to tease apart different sources of variability prior to depression classification.

METHODS

This section describes the methods of this work.

Data

The data used is the 2014 AVEC corpus [18], which is a subset of the 2013 AVEC corpus [19].

[1]Hamilton Rating Scale for Depression

The AVEC 2014 corpus consists of recordings of 2 different human-computer interaction tasks. Both tasks were Power Point guided. Each of the tasks are supplied as separate recordings. In total, the corpus includes 300 videos; the duration ranges from 6 seconds to 4 minutes. There was a total of 84 subjects, the mean age of subjects was 31.5 years, with a standard deviation of 12.3 years, and a range of 18 to 63 years. The two tasks include a read task and a spontaneous speech task, described below:

- Participants read aloud an excerpt of the fable "Die Sonne und der Wind" (The North Wind and the Sun). (German)
- Participants respond to one of a number of questions such as: What is your favorite dish or discuss a sad childhood memory. (German)

Each recording is labeled for severity of depression. Depression severity is determined using the Beck Depression Inventory-II (BDI-II) [3]. The BDI-II contains 21 questions. Each item of the BDI-II is a forced-choice question scored on a discrete scale with values ranging from 0 to 3. Final BDI-II scores range from 0 63 (0-13 no or minimal depression, 14-19 mild depression, 20-28 moderate depression, 29-63 severe depression). The AVEC 2014 corpus is already partitioned for training and development data sets. For the training and development corpora respectively the average BDI-II is 15.0 and 15.6 (with standard deviations of 12.3 and 12.0). Each of the partitions contains 92 audio files.

Features

The feature set used was borrowed from the AVEC 2014 baseline [18]. The set consists of 2268 features extracted using the OpenSmile toolkit [7]. The features are composed of 32 energy and spectral related low-level descriptors (LLD) x 42 functionals, 6 voicing related LLD x 32 functionals, 32 delta coefficients of the energy/spectral LLD x 19 functionals, 6 delta coefficients of the voicing related LLD x 19 functionals, and 10 voiced/unvoiced durational features. In addition, the LLD set covers a standard range of commonly used features in audio signal analysis and emotion recognition. Features were extracted over overlapping short fixed length segments of 20 seconds which are shifted forward at a rate of one second [18].

Clustering

Using the above mentioned feature set, unsupervised clustering was performed using the K-means clustering algorithm in Weka [8]. Clustering can be defined as the unsupervised classification of patterns into groups. The resulting groups or clusters should ideally exhibit the following characteristics: (1) homogeneity within the clusters, and (2) heterogeneity between clusters. Several algorithms require certain parameters for clustering, such as the number of clusters. For K-means clustering, the number of clusters k must be specified. Since the data set is relatively small only small values of k are explored (k=2 up to k=5). We hypothesized that low values of k would capture the most basic forms of variation, such as a gender, where higher values of k would capture more complex forms of variation. Subsequently, the clusters established were used to train different models based on each

Number of Clusters (k = 2 to k = 5)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
k = 2	9.28 t-45 T-48	9.94 t-47 T-44	—	—	—
k = 3	17.01 t-36 T-29	10.57 t-35 T-34	7.49 t-21 T-29	—	—
k = 4	26.14 t-21 T-21	8.73 t-21 T-25	8.82 t-17 T-18	10.31 t-33 T-28	—
k = 5	23.85 t-21 T-26	10.87 t-3 T-15	8.82 t-17 T-18	10.17 t-41 T-18	5.85 t-10 T-15

Table 1. MAE Results using different values of k, t represents the number of training instances and T represents the number of test instances.

cluster. During test time, each feature vector is compared to all existing cluster centroids by computing the euclidean distance between the 2 vectors. The cluster centroid represents the average across all the points in the cluster. The closest cluster to the new feature vector in question is then chosen as the model that will be used during classification. So for example for a given cluster centroid p and a feature vector q we calculate the distance between the two using the formula below.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_n - q_n}$$

We calculate the distance for every cluster centroid and the closest centroid is marked as the cluster that will be used during training for that specific feature vector.

Evaluation

Since depression is measured on a severity scale this task represents a single regression problem. The learning algorithm we employ is SMO regression (default parameters) in Weka [8]. We choose to adopt the evaluation metric of the AVEC 2014 Depression Sub-challenge: mean absolute error (MAE) [18]. Lastly, we use as our baseline, an MAE of 10.26. This MAE score is achieved by evaluating our feature set on the test data without clustering (92 instances in train and test). Mean absolute error can be defined, at a basic level, as the absolute average difference between the actual labels and the predictions made. In most work involving the AVEC 2014 corpus, Root Mean Square Error is also reported, here we choose to only report MAE. We choose to adopt this baseline and not the challenge or challenge participants' baselines because our feature set is only extracted from the audio signal. Since the AVEC 2014 corpus also includes video, many systems chose to incorporate features from that signal. For this reason, a direct comparison can not be easily made between our system and those existing systems.

RESULTS

The results of the clustering experiments are given in the table above. The hypothesis we tested was whether or not clustering could provide a way to tease apart different sources of variability prior to depression classification. The scores achieved with clustering can be compared to our baseline: an MAE of 10.26.

Label	Homogeneity	Completeness
task	.0584	.0581
gender	.0002	.0002
depression level	.0068	.0065

Table 2. Metrics for cluster assignment. Depression level represents a discrete label of low versus high.

It can be noted that clustering in many cases does help boost performance . When the number of clusters is small we see a uniform improvement of lower MAE across clusters; cluster 1 and cluster 2 for k=2 achieve a MAE of 9.28 and 9.94 respectively. For each of the values of k we see improvements in some clusters but not in all. In some cases, we see substantially worst performance. Due to data size, it is possible that when the values of k increase performance worsens due to the small number of training instances. Data size definitely presents a limitation to this approach.

In order to make claims about what traits or labels the clusters may possibly be representing, we use two metrics to evaluate the clusters performance. Given the knowledge of the ground truth class assignments of the data, it is possible to define some intuitive metrics. Specifically some [14] have defined the following two desirable objectives for any cluster assignment:

1. homogeneity: each cluster contains only members of a single class.
2. completeness: all members of a given class are assigned to the same cluster

Scikit-learn's implementation of the above metrics is used. The labels considered are task (read speech vs. spontatneous speech), gender (male vs. female), and depression level (low/none vs. high). Levels of depression are determined by using the clinical suggestion attached to the Beck Depression Inventory rating scale, which suggests that an individual should seek out professional help when receiving a score of 17 and above. Therefore, any participants rated 17 and above are considered to be in the high group and the rest in the low/none group.

Scores are bounded between 0 and 1 (1 being the best). For this evaluation, only the k=2 clustered is considered. The scores for each of the metrics are given in the table above. As

the results show, neither of the labels received high marks. Relative to the results given, the clusters seem to be capturing task above any other trait, suggesting that this factor should be separated out prior to classification. Consequently, when task is addressed and 2 models are trained (one for each task), the MAE results improve to 7.6 and 9.71 respectively for the spontaneous and read speech task. These results are consistent with previous findings that suggest task differences should be considered [18]. These metrics, to some extent, support the claim that clustering is capable of capturing variation and differences, across task, and potentially across speaker traits. Important to note, we are not very concerned with whether the clusters capture specific labels; they may be capturing any combination of factors that may affect depression detection—gender, age, accent, class, emotion, etc. We are more concerned with whether clustering helps the task of depression detection.

CONCLUSION

There are many challenges to depression classification. This work focused on addressing one issue: variability factors. We presented an approach based on unsupervised clustering that resulted in slight gains to performance as measured by MAE. We found that by clustering prior to classification we are able to mitigate factors, such as speech task. However, as the value of k increased it seems that data size presented as an issue. Future work should consider upsampling and data augmentation techniques to help overcome this limitation. In addition, this work only used a standard feature set with no feature analysis or development. Since cluster assignment is based solely on the features employed, feature development should be explored to improve upon current work. Lastly, only one simple clustering algorithm was explored: K-means. More sophisticated algorithms may help leverage performance gains without the draw backs of the simple approach taken here. Future work will focus on these issues to improve upon existing work.

REFERENCES

1. Depression factsheet, World Health Association, 2015.
2. Facts & Statistics, Anxiety and Depression, Association of America, ADAA, 2015.
3. Beck, A. T., Ward, C., Mendelson, M., et al. Beck depression inventory (bdi). *Arch Gen Psychiatry 4*, 6 (1961), 561–571.
4. Cummins, N., Epps, J., Breakspear, M., and Goecke, R. An investigation of depressed speech detection: Features and normalization. In *Interspeech* (2011), 2997–3000.
5. Cummins, N., Epps, J., Sethu, V., and Krajewski, J. Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE (2014), 970–974.
6. Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., and Epps, J. Diagnosis of depression by behavioural signals: a multimodal approach. In *3rd ACM international workshop on Audio/visual emotion challenge Proc.*, ACM (2013), 11–20.
7. Eyben, F., Weninger, F., Gross, F., and Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *21st ACM international conference on Multimedia Proc.*, ACM (2013), 835–838.
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*, 1 (2009), 10–18.
9. Horwitz, R., Quatieri, T. F., Helfer, B. S., Yu, B., Williamson, J. R., and Mundt, J. On the relative importance of vocal source, system, and prosody in human depression. In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, IEEE (2013), 1–6.
10. Hu, Y., Wu, D., and Nucci, A. Fuzzy-clustering-based decision tree approach for large population speaker identification. *Audio, Speech, and Language Processing, IEEE Transactions on 21*, 4 (2013), 762–774.
11. Kinnunen, T., KILPELÄINEN, T., and FrÄnti, P. Comparison of clustering algorithms in speaker identification. *dim 1* (2011), 2.
12. Lenneberg, E. H., Chomsky, N., and Marx, O. *Biological foundations of language*, vol. 68. Wiley New York, 1967.
13. Quatieri, T. F., and Malyska, N. Vocal-source biomarkers for depression: A link to psychomotor activity. In *Interspeech* (2012).
14. Rosenberg, A., and Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, vol. 7 (2007), 410–420.
15. Scherer, K. R. Vocal affect expression: a review and a model for future research. *Psychological bulletin 99*, 2 (1986), 143.
16. Sturim, D. E., Torres-Carrasquillo, P. A., Quatieri, T. F., Malyska, N., and McCree, A. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Interspeech* (2011), 2981–2984.
17. Trevino, A. C., Quatieri, T. F., and Malyska, N. Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing 2011*, 1 (2011), 1–18.
18. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. Avec 2014: 3d dimensional affect and depression recognition challenge. In *4th Audio/Visual Emotion Challenge Proc.*, ACM (2014), 3–10.
19. Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *3rd ACM international workshop on Audio/visual emotion challenge Proc.*, ACM (2013), 3–10.