Crosslinguistic Multimodal Feature and Fusion Analysis for Automatic Detection of Depression

Michelle Renee Morales

The Graduate Center, City University of New York, USA

mmorales@gradcenter.cuny.edu

Abstract

Depression is a serious illness that affects millions of people globally. In recent years, the task of automatic depression detection from speech has gained popularity but, several challenges remain, including which features provide the best discrimination between classes or depression levels. Thus far, most research has focused on extracting features from the speech or video signal. However, the speech production system is complex and depression has been shown to affect many linguistic properties, including phonetics, semantics, and syntax. Therefore, we argue that researchers should look beyond the acoustic properties of speech by building features that capture syntactic structure and semantic content. This work provides an in-depth analysis of features for depression detection. Using various corpora, this project evaluates how features extracted from different signals (video, audio, text) compare to one another. We also experiment with various techniques for how best to fuse features from different modalities. Additionally, we explore crosslinguistic detection to determine whether language-specific characteristics have an effect. Lastly, we present a novel multimodal crosslinguistic representation.

Index Terms: Depression detection, feature development, feature fusion, multimodal, crosslinguistic

1. Introduction

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 [1]. Due to the variation in how depression presents itself within each person, it is difficult and time consuming to diagnose. Since diagnosis often relies on a clinician's assessment, it is also subjective. Moreover, many under-served areas have severe shortages of clinicians who can make this diagnosis. Numerous 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, which could be used to complement existing diagnostic measures.

There exist many challenges to detecting and modelling depression. Key among these is developing novel clinically-motivated or neuroscience-motivated features, i.e. features that capture characteristics specific to depression. This work will focus on this challenge, by exploring feature development for depression detection and investigating how best to build a detection system that encompasses features from multiple linguistic levels of analyses. The aim of this exploration is to discover which features provide the best discrimination between levels of severity of depression levels.

If we consider the process of information flow during speech production, we note that the entire set of operations by which a speaker transforms ideas into acoustic output is enormously complex [2, 3]. We know that the stages of semantics and syntax affect speech [3]. We also know that a multitude of psychological symptoms can be assessed by analysis of language behavior [4]. Specifically, for individuals with depression, many linguistic variables have been shown to be affected, including prosody [5, 6, 7, 8], syntax [9], and semantics [10, 11]. In addition to verbal behavior, we also know that non-verbal behaviors are also affected by depression, including facial expression, gesture, and posture [12, 13]. Therefore, we hypothesize that the best depression detection system will include features from as many facets of communication as possible. However, these facets represent different modalities and combining them is no simple task. Therefore, one aim of this work is to explore different techniques for multimodal learning. Additionally, whenever features are extracted directly from a linguistic signal it is important to consider whether there exist crosslinguistic effects. Therefore, this work will investigate how features perform across multiple corpora in different languages.

2. Research Goals

Completion of the research goals in this project is envisioned through three components.

2.1. Feature Development and Evaluation

This work will develop and evaluate features generated from three different signals: (1) video, (2) audio, and (3) text. ***Video features*** are included to help capture an individual's non-verbal behavior. Visual features will be extracted using OpenFace. OpenFace is an open source tool capable of facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation [14]. ***Audio features*** are included to help target specific verbal behavior associated with depression. Acoustic features will be extracted using openSMILE [15]. ***Text features*** are included to help target specific language use. Features derived from the transcripts will include both syntactic and semantic features. Semantic features will be generated by calculating word category frequency using the Linguistic Inquiry and Word Count tool [16]. In addition, sequential word embeddings will also be used to represent the semantic content of each utterance [17]. Each utterance will also be parsed using Google's parser [18]. Syntactic features will then be generated using the parser's output. One of the obstacles to using text-based features is the time consuming nature of transcription. An additional step that is not required for video or audio. In order to mitigate this issue, we will explore the use of automatic speech recognition (ASR). Consequently, text features will be gener-

ated from both manual transcripts and ASR output to determine whether or not ASR can be used successfully. All feature sets will then be evaluated in isolation to determine which set or sets of features provide the best discrimination between classes.

2.2. Multimodal Learning Techniques

We hypothesize that the best depression detection system will include features from as many aspects of communication as possible, i.e. a multimodal approach. We intend to explore different techniques for combining features from different modalities. Fusion techniques for multimodal features have been extensively explored [19, 20, 21]. This work will evaluate three existing techniques: early fusion, late fusion, and late fusion with voting. In early fusion, features from all modalities are fused (concatenated) in the feature space before supervised learning occurs, whereas in late fusion, classes/scores are learned directly from unimodal features, then these scores are integrated to learn depression labels [22]. In late fusion with voting, unimodal features are used to train the models and each model votes, with each vote weighted by confidence. The class/score with the highest vote is the final prediction[20].

2.3. Crosslinguistic Effects and Novel Representation

The data in this project will include corpora in two different languages: English and German. The Distress Analysis Interview Corpus (DAIC) [23] is a multimodal collection of semi-structured clinical interviews in English. The corpus was designed to simulate standard protocols for identifying people at risk for post-traumatic stress disorder (PTSD) and major depression. The corpus contains four types of interviews: (1) face-to-face interviews between participants and a human interviewer, (2) teleconference interviews, conducted by a human interviewer over a teleconferencing system, (3) Wizard-of-Oz interviews, conducted by an animated virtual interviewer, controlled by a human interviewer in another room and (4) automated interviews, where participants are interviewed by a fully automated operating agent. The German corpus used in this work is the 2014 Audio-Visual Emotion recognition Challenge corpus (AVEC) [24]. The corpus includes videos of individuals responding to a question as well as individuals reading a story aloud. Both the AVEC and DAIC corpora are labeled for depression level. This work will evaluate how different sets of features perform across languages. We hypothesize that features from certain modalities will be more robust across languages. For example, facial expression may be robust in capturing differences between labels whereas verb choice is not. We will evaluate how crosslinguistic classification performs, e.g. can we train on English and predict on German? This would be especially useful since data size is often a problem for this task. Intuitively, video and audio provide signals which capture aspects of emotion and behavior irrespective of language. However, text is more strongly tied to language. Therefore, it is a signal, which may not be as robust as the others for crosslinguistic detection. Consequently, we will explore how to use information from other modalities to make a more robust representation for text-based features. Word lists and dictionaries, which can be useful to detect depression by targeting word use, are time intensive to create and do not exist for every language. Word embeddings have been used successfully in multiple languages and it has been shown that semantically similar words in different languages tend to appear in the same semantic space [25]. However, building informative word embeddings still requires a large amount of data in each language. Moreover, word embeddings are limited by the size and uniqueness of the data they are trained on. Therefore, unseen words tend to cause problems. Although useful for monolingual depression detection, word embeddings may prove not to be as robust for crosslinguistic depression detection. Therefore, we will experiment with domain adaptation of word embeddings. In addition, we propose a novel technique. First, all modalities will be force aligned. Then, instead of relying on specific lexical information, we will abstract away to each word's syntactic representation (i.e., part-of-speech (POS) tag). Using the Universal tag-set will allow us to easily extend this approach to any language. After abstracting away, we can then use information from the other modalities to enrich the POS tag with information about the nature of that word. For example, given a sentence this approach would include three components, shown below:

(1)	I	feel	awful.
(2)	PRN	VERB	ADJ
(3)	+neutral +pitch	+neutral +pitch	+eyes down +pitch

This approach can then be further extended to include an additional stage which would include dependency structure information, leading to a better understanding about the nature of the relationships in a given sentence. By using information from the other modalities we can learn information about the types of verbs/adjectives/etc. being used. Acoustic features may provide information about whether or not a word is a positively or negatively valenced word. Facial expressions during use of pronouns/proper nouns could give insight into how an individual feels about the self and others. This representation provides a way to fuse information from different modalities while also providing a text-based representation that can be used for any language. Our crosslinguistic evaluation will provide important insight into how to develop systems that perform well on multiple languages.

3. Contributions

The main contribution of this work is a crosslinguistic evaluation of features and fusion techniques for depression detection. In addition, we explore a novel multimodal crosslinguistic representation for language. Participation in the Interspeech 2016 Doctoral Consortium would allow me to learn how to communicate the challenges and techniques of this research proposal to a wider audience. Learning how to frame my research effectively with proper perspective will be invaluable as I prepare for my thesis proposal. I welcome the opportunity to receive valuable feedback.

4. Acknowledgments

This work is being carried out under the supervision of Rivka Levitan, Brooklyn College, City University of New York, with additional guidance from Martin Chodorow, Hunter College, City University of New York, and Stefan Scherer, USC Institute for Creative Technologies.

5. References

[1] ADAA, "Facts & Statistics, Anxiety and Depression, Association of America, ADAA," 2015. [Online]. Available: http://www.adaa.org/about-adaa/press-room/facts-statistics

[2] E. H. Lenneberg, N. Chomsky, and O. Marx, *Biological foundations of language*. Wiley New York, 1967, vol. 68.

[3] W. E. Cooper and J. Paccia-Cooper, *Syntax and speech*. Harvard University Press, 1980, no. 3.

[4] G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall, and C.-W. Wallesch, *Linguistic disorders and pathologies: an international handbook*. Walter de Gruyter, 1993, vol. 8.

[5] A. C. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–18, 2011.

[6] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biological Psychiatry*, vol. 72, no. 7, pp. 580–587, Oct. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006322312002636

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

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

[9] J. Zinken, K. Zinken, J. C. Wilson, L. Butler, and T. Skinner, "Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression," *Psychiatry research*, vol. 179, no. 2, pp. 181–186, 2010.

[10] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[11] T. E. Oxman, S. D. Rosenberg, P. P. Schnurr, and G. J. Tucker, "Diagnostic classification through content analysis of patients speech," *American Journal of Psychiatry*, vol. 145, no. 4, pp. 464–468, 1988.

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

[13] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[14] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[15] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *21st ACM international conference on Multimedia Proc.* ACM, 2013, pp. 835–838.

[16] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Linguistic inquiry and word count: Liwc [computer software]," *Austin, TX: liwc. net*, 2007.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[18] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally normalized transition-based neural networks," *arXiv preprint arXiv:1603.06042*, 2016.

[19] S. Walter, S. Scherer, M. Schels, M. Glodek, D. Hrabal, M. Schmidt, R. Böck, K. Limbrecht, H. C. Traue, and F. Schwenker, "Multimodal emotion classification in naturalistic user behavior," in *International Conference on Human-Computer Interaction*. Springer, 2011, pp. 603–611.

[20] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu, "Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 481–486.

[21] M. Chatterjee, S. Park, L.-P. Morency, and S. Scherer, "Combining two perspectives on classifying multimodal data for recognizing speaker traits," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 7–14.

[22] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[23] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*, 2014, pp. 3123–3128.

[24] 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 *4th Audio/Visual Emotion Challenge Proc.* ACM, 2014, pp. 3–10.

[25] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–43.