A Linguistically-Informed Fusion Approach for Multimodal Depression Detection

Michelle Renee Morales
Linguistics Department
The Graduate Center, CUNY
New York, NY 10016
mmorales@gradcenter.cuny.edu

Stefan Scherer
USC Institute for
Creative Technologies
Los Angeles, CA 90094
scherer@ict.usc.edu

Rivka Levitan
Computer Science Department
Brooklyn College, CUNY
Brooklyn, NY 11210
levitan@sci.brooklyn.cuny.edu

Abstract

Automated depression detection is inherently a multimodal problem. Therefore, it is critical that researchers investigate fusion techniques for multimodal design. This paper presents the first ever comprehensive study of fusion techniques for depression detection. In addition, we present novel linguistically-motivated fusion techniques, which we find outperform existing approaches.

1 Introduction

Depression is an extremely heterogeneous disorder that is difficult to diagnose. Given this difficulty, psychologists and linguists have investigated possible objective markers and have shown that depression influences how a person behaves and communicates, affecting facial expression, prosody, syntax, and semantics (Morales et al., 2017a). Given that depression affects both nonverbal and verbal behavior, an automated detection system should be multimodal. Initial studies on depression detection from multimodal features have shown performance gains can be achieved by combining information from various modalities (Morales and Levitan, 2016; Scherer et al., 2014). However, few studies have investigated fusion approaches for depression detection (Alghowinem et al., 2015). In this paper, we present a novel linguistically motivated approach to fusion: *syntax-informed fusion*. We compare this novel approach to early fusion and find it is able to outperform it. We also demonstrate that this approach overcomes some of the limitations of early fusion. Moreover, we test our approach's robustness by applying the same framework to generate a *visual-informed fusion* model. We find video-informed fusion also outperforms early fusion. In addition to presenting novel fusion techniques, we also evaluate existing approaches to fusion including early, late, and hybrid fusion. To the best of our knowledge, this work presents the first in-depth investigation of fusion techniques for depression detection. Lastly, we present interesting results to further support the relationship between depression and syntax.

2 Related Work

This work presents a multimodal detection system with a specific focus on the relationship between depression and syntax. This relationship motivates a novel approach to fusion. In contrast to a simple early fusion approach to combining modalities, a syntax-informed early fusion approach leverages the relationship between syntax and depression to help improve system performance. In this section, we first provide background on the relationship between depression and language, highlighting both the voice and syntax. In addition, we also evaluate a *video-informed fusion* approach which is motivated from the relationship between depression and facial activity as well as the relationship between facial behavior and speech production. Therefore, we also present related work on the relationship between visual information and depression. This is followed by a review of related work on multimodal fusion techniques that have been investigated for depression detection systems. In this section, we will only briefly cover relevant work, for a detailed review of multimodal depression detection systems see Morales et al. (2017a).

2.1 The Relationship between Depression and Language

Researchers have investigated the relationship between prosodic, articulatory, and acoustic features of speech and clinical ratings of depression (Cummins et al., 2015). In patients with depression, several changes in speech and voice have been noted, including changes in prosody (Blanken et al., 1993), speaking rate (?Stassen et al., 1998), speech pauses (Alpert et al., 2001), and voice quality (Scherer et al., 2013a).

In addition to voice and speech-based markers, researchers have also provided empirical support for the existence of a relationship between depression and syntax. Depressed individuals exhibit many syntactic patterns including an increased use of first person singular pronouns (Rude et al., 2004) and a decreased use of complex syntactic constructions, such as adverbial clauses (Zinken et al., 2010). The relationship between syntax and depression motivates our syntax-informed fusion approach.

2.2 The Relationship between Depression and Facial Activity

Similar to the relationship between language and depression, there also exists a body of research on the relationship between depression and facial activity. Depression affects individuals' facial expressions, including noted decreases in expressivity, eyebrow movements, and smiling (Cummins et al., 2015).

In addition, there also exists an interesting relationship between video and audio, e.g. the *McGurk effect*. McGurk and MacDonald (1976) were the first to report a previously unrecognized influence of vision upon speech perception. In their study, they showed participants a video of a young woman speaking, where she repeated utterances of the syllable [ba] which had been dubbed on to lip movements for the syllable [ga]. Participants reported hearing [da]. Then with the reverse dubbing process, a majority reported hearing [bagba] or [gaba]. However, when participants listened to only the sound of the video or when they watched the unprocessed video, they reported the syllables accurately as repetitions of [ba] or [ga]. These findings had important implications for the understanding of speech perception, specifically that visual information a person gets from seeing a person speak changes the way they hear the sound.

These interesting relationships —between the face and voice as well as facial expressions and depression —motivate our video-informed fusion approach.

2.3 Existing Fusion Approaches

In recent years, researchers have begun to investigate multimodal features for depression detection systems (Morales et al., 2017b). However, it is a fairly new research interest and as a result only a few studies have compared techniques for fusing features from different modalities (Alghowinem et al., 2015). In the few studies that have investigated fusion techniques, the canonical fusion techniques have been considered, including early, late, and hybrid fusion. In the **early fusion** approach, features are integrated immediately after they are generated through simple concatenation of feature vectors. In the **late fusion** approach integration occurs after each of the modalities have made a decision. In the **hybrid fusion** approach outputs from early fusion and individual unimodal predictors are combined (Baltrusaitis et al., 2017).

Researchers have found early fusion, although simple, to be a successful technique to combine modalities for depression, noting improvements over unimodal systems (Alghowinem et al., 2015; Morales and Levitan, 2016; Morales et al., 2017b; Scherer et al., 2013b). However, a drawback of the early fusion approach is the high dimensionality of the combined feature vector. Given that drawback, Joshi et al. (2013) considered early fusion as well as early fusion followed by Principal Component Analysis (PCA), where 98% of the variance was kept. They found that training a depression detection model on this reduced dimensionality feature set led to improved performance of the system over simple early fusion.

Researchers have also investigated late and hybrid fusion. In Alghowinem et al. (2015) a hybrid fusion approach was investigated, which involved concatenating results from individual modalities to the the early fusion feature vector. A majority voting method was used. They evaluated how hybrid fusion and early fusion approaches compare to unimodal approaches. They found that in most cases their early and hybrid fusion models outperformed the unimodal models. Moreover, hybrid fusion models tended to outperform early fusion. Late fusion approaches have also been investigated by some (Joshi et al., 2013; Meng et al., 2013). For example, Meng et al. (2013) used a late fusion approach that trained a separate model from each modality and combined decisions using the weighted sum rule. They found that combining visual and vocal features at the decision level resulted in further system improvement for depression detection.

Although, in this work, we focus on fusion approaches for depression detection, there exist various studies investigating fusion for other machine learning tasks. Researchers have also proposed new approaches to fusion which differ from the canonical approaches. In particular, deep learning approaches to fusion appear to be particularly promising. For example, Mendels et al. (2017) presented a single hybrid deep model with both acoustic and lexical features trained jointly and found that this approach to fusion achieved state-of-the-art results for deception detection. However, deep learning is not currently a good approach for depression detection, since labeled corpora are not very large and interpretable models are important.

3 Dataset

In this work, we use the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ; Gratch et al., 2014). The corpus is multimodal (video, audio, and transcripts) and is comprised of video interviews between participants and an animated virtual interviewer called Ellie, which is controlled by a human interviewer in another room.

Interview participants were drawn from the Greater Los Angeles metropolitan area and included two distinct populations: (1) the general public and (2) veterans of the U.S. armed forces. Participants were coded for depression, Posttraumatic Stress Disorder (PTSD), and anxiety based on accepted psychiatric questionnaires. All participants were fluent English speakers and all interviews were conducted in English. The DAIC-WOZ interviews ranged from 5 to 20 minutes.

The interview started with neutral questions, which were designed to build rapport and make the participant comfortable. The interview then progressed into more targeted questions about symptoms and events related to depression and PTSD. Lastly, the interview ended with a 'cool-down' phase, which ensured that participants would not leave the interview in a distressed state. The depression label provided includes a PHQ–8[1] score (scale from 0 to 24) as well as a binary depression class label, i.e., score $>=$ 10.

4 Features

In this work we use the OpenMM[2] pipeline to extract multimodal features (Morales et al., 2017b), which uses Covarep (Degottex et al., 2014) and Parsey McParseface (Andor et al., 2016) to extract voice and syntax features.

4.1 Voice

In order to extract features from the voice, OpenMM employs Covarep (Degottex et al., 2014). The audio features extracted include prosodic, voice quality, and spectral features. Prosodic features include Fundamental frequency (F_0) and voicing boundaries (VUV). Covarep voice quality features include Normalised amplitude quotient (NAQ), quasi open quotient (QOQ), the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), spectral tilt/slope of wavelet responses (peakslope), and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd). Spectral features include Mel cepstral coefficients (MCEP0-24), harmonic model and phase distortion mean (HMPDM0-24) and deviations (HMPDD0-12). Lastly, Covarep includes a creak feature which is derived through a creaky voice detection algorithm.

4.2 Syntax

In order to generate syntactic features OpenMM employs Google's state-of-the-art pre-trained tagger: Parsey McParseface (Andor et al., 2016). For each sentence S, the tagger outputs POS tags. In this work, we make use of 17 POS tags, which are outlined in Table **??** of the Appendix.

4.3 Visual

The visual features we consider are Action Units (AUs), which were extracted from the DAIC-WOZ corpus as part of the baseline system for the AVEC 2017 challenge (Ringeval et al., 2017). AUs represent the fundamental actions of individual muscles or groups of muscles. It is a commonly used tool and has become standard to systematically categorize physical expressions, which has proven very useful for psychologists. A detailed list of the facial AUs we consider are given in Table **??** of the Appendix. Each AU receives a presence score, between -5 and 5, which measures how present that feature is for a given frame of video.

[1] http://patienteducation.stanford.edu/research/phq.pdf

[2] https://github.com/michellemorales/OpenMM

5 Fusion Approaches

5.1 Early Fusion

In our early fusion approach, features are extracted from each modality and then concatenated to generate a single feature vector. Visual and acoustic features are extracted at the frame level while POS tags are extracted at the sentence level. Therefore, the modalities do not align automatically. In order to handle these differences, we first compute statistics (mean, median, standard deviation, maximum, and minimum) across frames/sentences. This results in 370 acoustic features (74 acoustic features × 5 statistical functionals), 100 visual features (20 visual × 5 statistical functionals), and 85 syntactic features (17 syntactic features × 5 statistical functionals). We then fuse the feature vectors to achieve one multimodal feature vector, $features_{early}$.

$$features_{early} = \overset{\text{Acoustic}}{\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_i \end{bmatrix}} + \overset{\text{Syntax}}{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_i \end{bmatrix}}$$

5.2 Informed Early Fusion

5.2.1 Syntax-informed Early Fusion

We compare early fusion to our proposed approach. Our approach leverages syntactic information to target more informative aspects of the speech signal. Given the relationship between depression and syntax, we hypothesize that this approach will help lead to improvements in system performance. First, we align the audio file and transcript file. In order to perform alignment, we use the tool *gentle*[3], which is a forced-aligner built on Kaldi. We then tag each sentence and retrieve the timestamp information for each POS tag. For each POS tag span we extract acoustic features for that time span.

$$features_{mm} = \begin{array}{cc} & \begin{matrix} y_0 & \cdots & y_i \end{matrix} \\ \begin{matrix} x_0 \\ x_1 \\ \vdots \\ x_i \end{matrix} & \begin{pmatrix} \bar{x}_0 & \cdots & \bar{x}_0 \\ \bar{x}_1 & \cdots & \bar{x}_1 \\ \vdots & \vdots & \vdots \\ \bar{x}_i & \cdots & \bar{x}_i \end{pmatrix} \end{array}$$

In other words, we are specifically extracting features at the POS level and we are continuously updating our audio features each time we come across a POS tag. For example, each time we see a VERB we use its timestamp information to extract mean F_0 from that specific window and we do this continuously, updating our F_0 value every time we come across a VERB. In the end, we have a mean F_0 value across all VERBs, ADJs, NOUNs, etc., as shown in $features_{mm}$. This representation is different from early fusion in that it conditions the audio features on POS information, providing a representation that does not simply add features from each modality, but instead aims to jointly represent them.

5.2.2 Video-informed Early Fusion

In order to test the robustness of our novel fusion approach —*informed early fusion* —we perform additional experiments using other modalities. The relationship between a person's facial behavior and speech production, motivates our *video-informed fusion* approach. Similar to our syntax-informed approach, where we target POS tags' time frames to identify more informative aspects of the speech signal, we also target aspects of the speech signal using visual information. We hypothesize that targeting informative aspects of the speech signal using visual cues will help boost system performance when compared with a simple early fusion system.

Similar to syntax-informed fusion, this representation conditions the audio features on AU information. For each frame of video, we identify the AU with the highest presence (value between -5 and 5). Therefore, we assume only one AU can occur per frame. For the AU with the highest presence, we extract acoustic features across that span of time. For each AU, we then aggregate its acoustic features across the entire video. In the end, we have a mean value for each acoustic feature across all AUs.

5.3 Late Fusion

We explore two types of late fusion approaches: (1) voting and (2) ensemble. In our voting approach, we train separate classification models for each modality. Each unimodal system makes a classification prediction, depressed or not depressed. We then take the majority vote as our ultimate prediction. We also consider an ensemble approach. In our ensemble late fusion approach, we again train separate classification models for each modality. The models' predictions are then

[3] https://github.com/lowerquality/gentle

Modality	Fusion Type	Precision	Recall	F1-score
A	–	0.34	0.70	0.45
S	–	0.21	0.96	0.35
V	–	0.16	0.52	0.25
A + S	E	0.34	0.70	0.45
A + S	I	0.40	0.69	**0.49**
A + S	E + I	0.36	0.62	0.44
A + V	E	0.37	0.70	0.48
A + V	I	0.36	0.77	**0.49**
A + V	E + I	0.34	0.74	0.46

Table 1: Results for 5-fold cross-validation using SVM. Results reported for the audio (A), syntax (S), video (V), and fusion (A + S) approaches. Fusion types include early (E), syntax-informed (I), and both (E + I).

used as features to train a new classification system. The predictions from the newly trained classification system are then used as the final prediction.

5.4 Hybrid Fusion

In our hybrid fusion approach, outputs from early fusion and individual unimodal predictors are combined. Therefore, we train separate classification models for each modality. We then take the predictions from each unimodal system and concatenate it with the early fused feature vectors. These new feature vectors (early fusion + unimodal predictors) are then used to train a new model to make the ultimate prediction.

6 Results

6.1 Binary Classification Experiments

In order to evaluate our approach, we conduct a series of participant-level binary classification experiments. We train both unimodal and multimodal models. Our *early + syntax-informed fusion* model combines both the early fusion and syntax-informed fusion feature sets, by early fusion, i.e. simple concatenation. Using scikit-learn[4] we train a Support Vector Machine (SVM) for classification, (linear kernel, C = 0.1). We conduct 5-fold cross-validation on 136 participant interviews (depressed = 26, non-depressed = 110). During cross-validation, each fold is speaker independent and drawn at random. Given the skewness of the dataset, we set the SVM model's class weight parameter to 'balanced', which automatically adjusts the weights of the model inversely proportional to the class frequencies in the data, helping adjust for the class imbalance. Given the possibility of sparse feature values and the differences in dimensionality across feature sets, we also perform feature selection. We use scikit-learn's *Select K-Best* feature selection approach, which computes the ANOVA F-value across features and identifies the K most significant features. We set K to 20 and evaluate each feature set's best set. We report our findings in Table **??**. We report precision, recall, and F1-score for the depressed class. We choose to report these values instead of the average values across both classes because the depressed class label is the harder class to detect. As a result, the non-depressed class usually reports very high scores which tend to inflate the average score. If we can increase the performance of the depressed class, it can be assumed that the overall performance will go up as a result.

We find that the novel syntax-informed fusion approach performs best, with an F1-score of 0.49. We believe this approach is able to leverage syntactic information to target more informative aspects of the speech signal resulting in higher performing models. By conditioning acoustic models on syntactic information this approach com-

[4] http://scikit-learn.org/

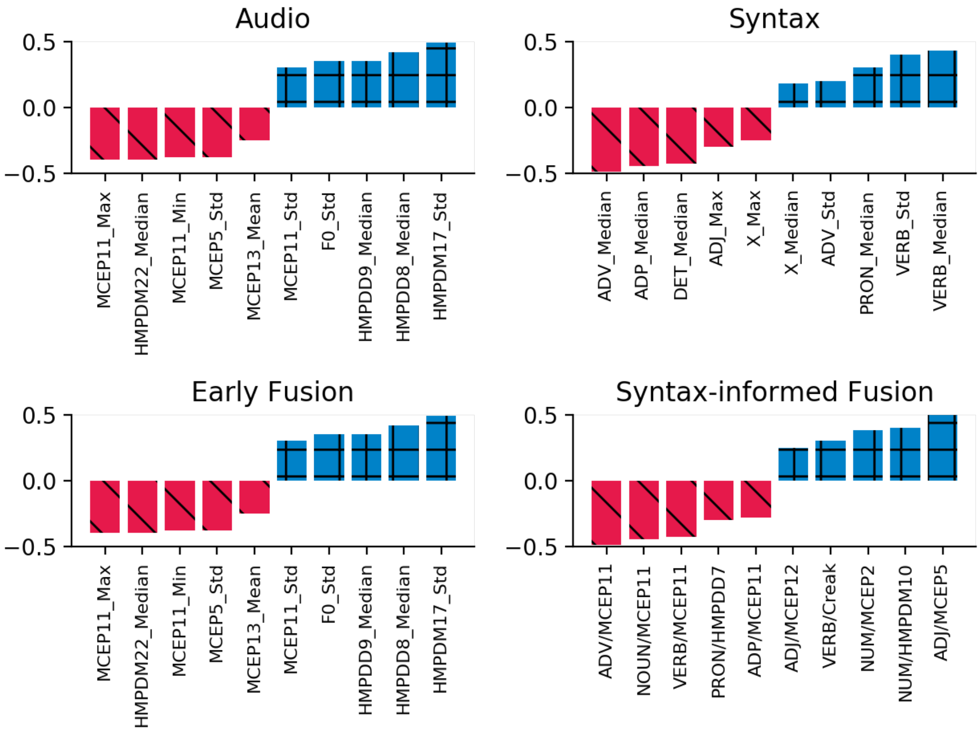

Figure 1: Illustration of linear kernel SVM's coefficient weights by class. Blue checkered bars represent the positive or depressed class. Red striped bars represent negative or healthy class.

bines information from both modalities in a way a human clinician might. Syntax-informed fusion substantially outperforms early fusion in precision and F1-score. In recall, performance is similar for both approaches. In addition, the syntax-informed method surfaces novel multimodal features. For example, creak is not a useful feature in the early fusion or the acoustic model. However, when we consider verb creak we find it extremely useful. This is demonstrated in Figure 1. To better understand each model, we inspect the coefficient weights of the SVM models. Using the weight coefficients from the models, we plot the top 5 most important features by class in Figure 1. The absolute size of the coefficients in relation to each other can be used to determine feature importance for the depression detection task.

If we consider the audio and early fusion models in Figure 1, we find that both models weight the same features highly. Although the early fusion model also includes the set of syntax features, it still prefers the same five features as the audio-only model. Since early fusion is simply concatenating the audio and syntax feature vectors it is understandable to find similar features performing well. These results show the promise of these specific audio features, which include spectral and prosodic (F_0) features. These results support previous work that showed spectral and prosodic features were useful for detecting depression (Cummins et al., 2015).

However, these findings also highlight the limitation of early fusion. The intention behind early fusion is to have access to multiple modalities that observe the same phenomenon to allow for more robust predictions, allowing for complementary information from each modality. Something not visible in individual modalities may appear when using multiple modalities. However, in early fusion, we can not guarantee that information from both modalities is considered. For example, if we inspect the feature set for early fusion we find that no syntax features appear; this could be attributed to the strength of the audio features as well as the difference in dimensionality size between the audio and syntax sets; the audio feature set is almost 5 times larger than the syntax set.

The syntax-informed fusion model is promising because it does not possess the same limitation as early fusion; with syntax-informed fusion we can guarantee that information from both modalities is considered. This could also be considered

Modality/Features	Fusion Type	Precision	Recall	F1-score
A + S	Early	0.34	0.70	0.45
A + S	Informed	0.40	0.69	0.49
A + S	Late - ensemble	0.36	0.78	0.49
A + S	Hybrid - informed	0.36	0.78	0.49
A + S	Hybrid - early	0.34	0.74	0.46
A + V	Early	0.37	0.70	0.48
A + V	Informed	0.36	0.77	0.49
A + V	Late - ensemble	0.36	0.78	0.49
A + V	Hybrid - informed	0.36	0.78	0.49
A + V	Hybrid - early	0.50	0.74	0.35
A + S + V	Early	0.37	0.70	0.48
A + S + V	Late - vote	0.50	0.17	0.25
A + S + V	Late - ensemble	0.36	0.78	0.49

Table 2: Results for fusion experiments using SVM.Results for fusion approaches including features from audio (A), syntax (S), and video (V).

a drawback of syntax-informed fusion, in circumstances where one would like to be agnostic regarding the value of each modality. However, in a task for which multiple modalities are known to be important and interconnected, such as depression detection, it is valuable to represent them jointly. The syntax-informed fusion model in Figure 1 demonstrates that syntax-informed fusion is able to capture important information from both modalities. We find the best features used to distinguish between classes are spectral features that span the production of pronouns, verbs, and adverbials. In other words, the best syntax-informed features represent a fused multimodal representation of the best features from each unimodal domain.

We also find further support of the relationship between depression and syntax. From the syntax-only model, we find pronouns (PRON) to be useful in identifying the depressed class, which supports previous findings that pronoun use can help identify depression (Rude et al., 2004). In addition, we find the POS tag category X (other) to be useful in distinguishing between classes. After manually inspecting the transcripts, we find the X POS tag is often assigned to filler words such as *uh, um, mm*. These results suggest filler words can be helpful in identifying depression. Lastly, we find adverbials (ADV) to be useful in distinguishing between classes. These results are especially interesting because Zinken et al. (2010) argued that adverbial clauses could help predict the improvement of depression symptoms. To the best of our knowledge, these results are the first to show support that adverbial clauses could also help predict depression.

We find similar results for video-informed fusion. Video-informed fusion outperforms early fusion in recall and F1-score. Similar to syntax-informed fusion we find that video-informed fused features are able to jointly capture the most informative features from each individual modality. For example, we find the best performing acoustic features and AUs from the unimodal systems to appear together in the video-informed system [5].

6.2 Fusion Experiments

In addition to evaluating how well our novel approach compares to early fusion, we also evaluate other types of fusion such as late and hybrid fusion. These series of experiments follow the

[5] Full charts of the the video-informed coefficient weights can be viewed in Figure 2 of the Appendix

same configuration as our first series of experiments: 5 fold cross-validation using SVM (linear kernel, C = 0.1, class weights balanced). We evaluate each method of fusion —early, informed, late (vote/ensemble), and hybrid (early/informed) —and report our results in Table ??.

As mentioned previously, in regards to early fusion methods, the *informed fusion* approaches outperform simple early fusion. When we compare the syntax and video-informed fusion techniques with other approaches, such as late and hybrid fusion, we do not find differences between the systems. When we evaluate systems that use all three modalities (A + S + V), we find a late ensemble approach performs best. We also find that late fusion techniques which rely on voting perform the worst. We believe these results can be attributed to the low performing unimodal video system, as demonstrated in Table ??. This finding highlights a weakness of the late fusion (voting) approach. Since it weighs the prediction from each system equally, this can lead to poor performance when one of the unimodal systems is weak.

7 Conclusion

In this work, we present a novel approach to early fusion: *informed fusion*. The syntax-informed fusion approach is able to leverage syntactic information to target more informative aspects of the speech signal. We find that syntax-informed fusion approach outperforms early fusion. Given some of the limitations to early fusion, we believe syntax-informed fusion is a promising alternative dependent on the classification task. In addition, we evaluate this approach's robustness by evaluating the technique with other modalities. Specifically, we evaluate video-informed fusion and confirm our findings that *informed fusion* outperforms early fusion. We also confirm previous findings that spectral features and prosodic features are useful in identifying depression. In addition, we present further support for the relationship between syntax and depression. Specifically we find pronouns, adverbials, and fillers to be useful in identifying individuals with depression. Lastly, we perform an in-depth investigation of fusion techniques and find that *informed*, late, and hybrid approaches perform comparably. To the best of our knowledge, this work represents the most comprehensive empirical study of fusion techniques for multimodal depression detection. However, this analysis is conducted on one dataset. Future work will consider extending this study to include many of the publicly-available existing datasets.

References

Sharifa Alghowinem, Roland Goecke, Jeffrey F Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear. 2015. Cross-cultural detection of depression from nonverbal behaviour. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE.

Murray Alpert, Enrique R Pouget, and Raul R Silva. 2001. Reflections of depression in acoustic measures of the patients speech. *Journal of Affective Disorders*, 66(1):59–69.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406.

Gerhard Blanken, Jürgen Dittmann, Hannelore Grimm, John C Marshall, and Claus-W Wallesch. 1993. *Linguistic disorders and pathologies: an international handbook*, volume 8. Walter de Gruyter.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964. IEEE.

Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear. 2013. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces*, 7(3):217–228.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. 264:746–748.

Gideon Mendels, Sarah Ita Levitan, Kai-Zhan Lee, and Julia Hirschberg. 2017. Hybrid acoustic-lexical deep learning approach for deception detection.

Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed AI-Shuraifi, and Yunhong Wang. 2013. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 21–30. ACM.

Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 136–143. IEEE.

Michelle Renee Morales, Stefan Scherer, and Rivka Levitan. 2017a. A Cross-modal Review of Indicators for Depression Detection Systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–12, Vancouver, Canada. Association for Computational Linguistics.

Michelle Renee Morales, Stefan Scherer, and Rivka Levitan. 2017b. OpenMM: An Open-Source Multimodal Feature Extraction Tool. In *Proceedings of Interspeech 2017*, pages 3354–3358, Stockholm, Sweden. ISCA.

Fabien Ringeval, Bjorn W. Schuller, Michel F. Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmi, and Maja Pantic. 2017. Avec 2017 real-life depression, and aect recognition workshop and challenge.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. 2013a. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Interspeech*, pages 847–851.

Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Louis-Philippe Morency, et al. 2014. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658.

Stefan Scherer, Giota Stratou, and Louis-Philippe Morency. 2013b. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 135–140. ACM.

H. H Stassen, S Kuny, and D Hell. 1998. The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology*, 8(4):303–310.

Jörg Zinken, Katarzyna Zinken, J Clare Wilson, Lisa Butler, and Timothy Skinner. 2010. Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression. *Psychiatry research*, 179(2):181–186.

A Appendix

POS Tag	Description
ADJ	Adjectives
ADV	Adverbs
ADP	Adpositions
AUX	Auxiliaries
CONJ	Conjunctions
DET	Determiners
INTJ	Interjections
NOUN	Nouns
NUM	Cardinal numbers
PPRON	Proper nouns
PRON	Pronouns
PRT	Particles or other functions words
PUNCT	Punctuation
SCONJ	Subordinating conjunctions
SYM	Symbols
VERB	Verbs
X	Other

Table 3: Description of POS tags.

Action Unit	Description
1	Inner brow raise
2	Outer brow raise
4	Brow lowerer
5	Upper lid raiser
6	Check raiser
7	Lid tightener
9	Nose wrinkler
10	Upper lip raiser
12	Lip corner puller
14	Dimpler
15	Lip corner depressor
17	Chin raiser
18	Lip puckerer
20	Lip strecher
23	Lip tightener
24	Lip pressor
25	Lips part
26	Jaw drop
28	Lip suck
43	Eyes closed

Table 4: Description of facial AUs.

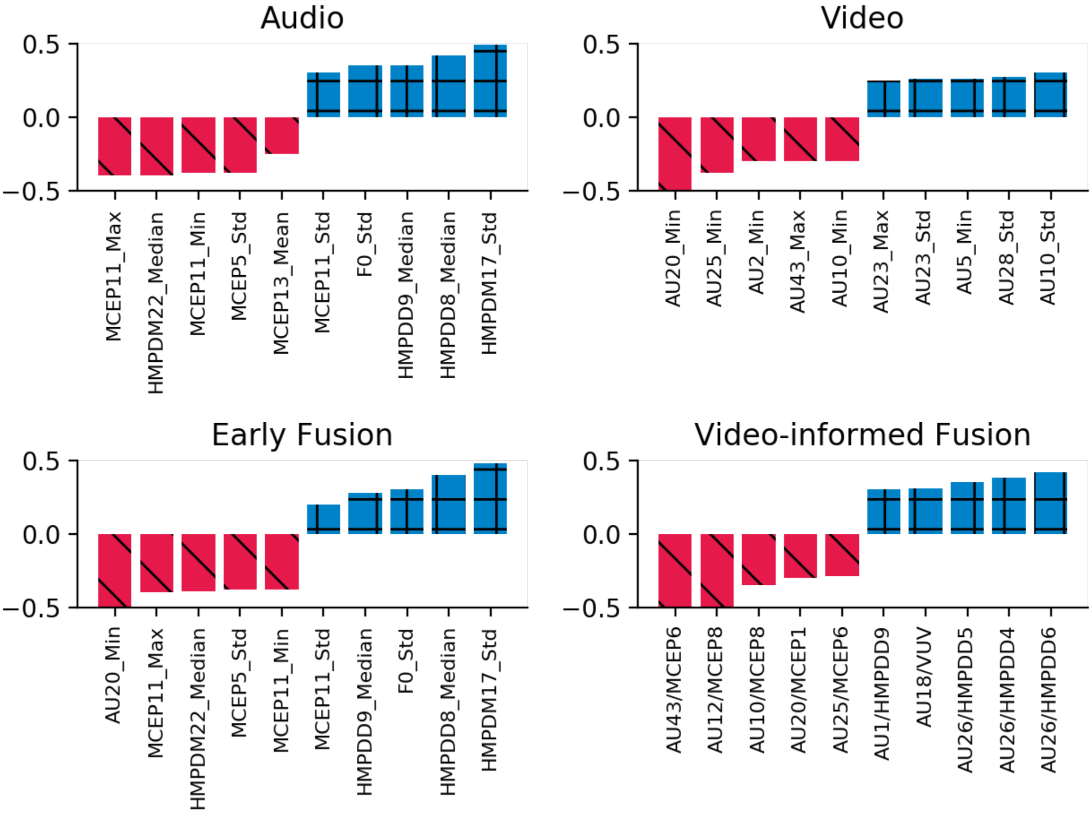

Figure 2: Illustration of linear kernel SVM's coefficient weights by class. Blue checkered bars represent the positive or depressed class. Red striped bars represent negative or healthy class.