Addressing Variability in Speech when Recognizing Emotion and Mood In-the-Wild

by

John H. Gideon

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2019

Doctoral Committee:

Professor Emily Mower Provost, Chair
Professor Melvin McInnis
Professor Rada Mihalcea
Professor V.G. Vinod Vydiswaran

John H. Gideon

gideonjn@umich.edu

ORCID iD: 0000-0003-3945-3341

For my wife, Emily

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Emily Mower Provost, for her mentorship throughout my time at the University of Michigan. She has provided me with plenty of encouragement and opportunities to grow as a researcher and as an individual, and I am especially grateful. I would also like to thank Dr. Melvin McInnis, who has supported me since the beginning of my time working on the PRIORI project, and is always enthusiastic in talking about new developments and directions.

Many thanks to my other two committee members, Dr. Rada Mihalcea and Dr. V. G. Vinod Vydiswaran, who have provided useful feedback and advice on my work.

I am deeply grateful to the members of the Chai Lab – Yelin, Duc, June, Soheil, Didi, Zak, Mimansa, Katie, Matt, Amrit, and Zahi. They have always been happy to provide helpful guidance during our weekly lab lunches. In particular, I'd like to thank Soheil, Zak, and Katie, who have helped co-author papers featured in this dissertation. Thanks also to the bipolar research team in the Prechter Program at the University of Michigan Depression Center, who have been instrumental to the collection and annotation of the PRIORI dataset.

Thanks to my collaborators at Brown University and Butler Hospital, including Dr. Heather Schatten, who provided support and an interesting new direction for my work. I'd also like to thank the NSF, NIMH, the Heinz C Prechter Bipolar Research Fund, and the Richard Tam Foundation at the University of Michigan for their financial support.

Lastly, I'd like to thank my friends and family, who have motivated me every step

of the way. My friends in GradTONES helped me make the transition to Ann Arbor and made the stresses of graduate school much easier to handle. My parents and sister, Rachel, have provided me with love throughout my life and are always there to offer support. Above all, I'd like to thank my wife, Emily, for being my steadfast supporter in all things and always encouraging me to pursue my passions.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

ABSTRACT

Bipolar disorder is a chronic mental illness, affecting 4% of Americans, that is characterized by periodic mood changes ranging from severe depression to extreme compulsive highs. Both mania and depression profoundly impact the behavior of affected individuals, resulting in potentially devastating personal and social consequences. Bipolar disorder is managed clinically with regular interactions with care providers, who assess mood, energy levels, and the form and content of speech. Recent work has proposed smartphones for automatically monitoring mood using speech.

Much of the early work in speech-centered mood detection has been done in the laboratory or clinic and is not reflective of the variability found in real-world conversations and conditions. Outside of these settings, automatic mood detection is hard, as the recordings include environmental noise, differences in recording devices, and variations in subject speaking patterns. Without addressing these issues, it is difficult to move towards a passive mobile health system. My research works to address this variability present in speech so that such a system can be created, allowing for interventions to mitigate the life-changing effects of mood transitions.

However detecting mood directly from speech is difficult, as mood varies over the course of days or weeks, while speech fluctuates rapidly. To address this, my thesis explores how an intermediate step can be used to aid in this prediction. For example, one of the major symptoms of bipolar disorder is emotion dysregulation - changes in the way emotions are perceived and a lack of inhibition in their expression. My work has supported the relationship between automatically extracted emotion estimates

and mood. Because of this, my thesis explores how to mitigate the variability found when detecting emotion from speech. The remainder of my thesis is focused on employing these emotion-based features, as well as features based on language content, to real-world applications. This dissertation is divided into the following parts:

- **Part I**: I address the direct classification of mood from speech. This is accomplished by addressing variability due to recording device using preprocessing and multi-task learning. I then show how both subject-specific and population-general information can be combined to significantly improve mood detection.

- **Part II**: I explore the automatic detection of emotion from speech and how to control for the other factors of variability present in the speech signal. I use progressive networks as a method to augment emotion with other paralinguistic data including gender and speaker, as well as other datasets. Additionally, I introduce a novel domain generalization method for cross-corpus detection.

- **Part III**: I demonstrate real-world applications of speech mood monitoring using everyday conversations. I show how the previously introduced generalized model can predict emotion from the speech of individuals with suicidal ideation, demonstrating its effectiveness across domains. Furthermore, I use these predictions to distinguish individuals with suicidal thoughts from healthy controls. Lastly, I introduce a novel framework for intervention detection in individuals with bipolar disorder. I then create a natural speech mood monitoring system based on features derived from measures of emotion and automatic speech recognition (ASR) transcripts and show effective intervention detection.

I conclude this dissertation with the following future directions: (1) Extending my emotion generalization system to include multiple modalities and factors of variability; (2) Expanding natural speech mood monitoring by including more devices, exploring other data besides speech, and investigating mood rating causality.

CHAPTER I

Introduction

1.1 Problem Statement

Bipolar disorder is a severe, chronic mental illness that typically begins in early adulthood and is characterized by periodic and pathological mood changes ranging from extreme lows (depression) to extreme highs (mania) [20]. It is among the top 10 leading causes of disability in the United States [132] with up to 20% of people affected taking their own life [81]. Bipolar disorder has a core clinical expression pattern related to emotion, energy, and psychomotor activity that can be monitored to gauge the health and progress of the individual in treatment [20, 140]. Intense clinical monitoring is effective at mitigating the severity of mood episodes, but is unrealistic due to cost and the availability of skilled health care providers [15, 146].

The long-term goal of my research has been to create a system that is able to use passively recorded smartphone conversations for mania and depression detection. This will address the need for ongoing monitoring of bipolar disorder in a cost efficient manner and could be used to augment traditional clinical treatment.

However, mood recognition from speech is challenging, as there are many other sources of variability present in speech besides mood, obscuring its detection. These factors of variability include those due to the demographics of the speaker, environment, and recording device [21]. Methods are needed to address these other factors

of variability and produce a more robust representation of mood.

Furthermore, there is a relatively large temporal disconnect between mood and speech. Mood varies on the order of days or weeks, while speech recordings vary at a sub-second rate. An intermediate step is needed in order to improve mood classification. One potential candidate for this is emotion, as a hallmark symptom of bipolar disorder is emotional dysregulation [83]. Individuals with bipolar disorder may interpret neutral actions negatively, have reductions in emotional inhibitions, and inappropriate intense emotional reactions [41, 83, 89, 152]. Psychological research indicates that this is caused by a dysfunctional limbic system and prefrontal brain network [83]. This lack of emotional control can result in issues in social and professional settings.

Emotion can be automatically estimated from speech in order to determine an individual's level of emotional dysregulation. This would help to bridge the relative temporal gap by first detecting emotion from speech and then using those predictions as indicators of mood changes. However, speech emotion detection suffers from many of the same issues as speech mood detection, due to the other factors of variability present in speech. I argue that a real-world bipolar mood monitoring system is possible by both addressing the different types of speech variability and using emotion as a bridge for the speech-mood temporal disconnect.

1.2 Bipolar Disorder Overview

Bipolar disorder is classified into one of two categories - type I and type II [18]. The main distinguishing factor between the two is the severity of the mania or hypomania. Bipolar type II, involves what are called hypomanic states which last for times usually shorter than mania and do not cause extreme feelings of grandiosity or hallucinations. In contrast, patients with bipolar type I have had at least one episode of mania in their life, involving more intense and longer lasting symptoms, sometimes including

psychosis. Both diagnoses also require fluctuating mood states of depression, usually lasting at least 2 weeks [154]. Bipolar disorder not otherwise specified (BP-NOS) and cyclothymic disorder are two other types that are either outside the scope of the other diagnoses or are less severe. Another possible form of the disorder is rapid-cycling bipolar, which involves four or more episodes of mania, hypomania, depression, or mixed states in a year at some point in the life of an individual.

Individuals with depression experience increased feelings of guilt and low self-esteem, along with an overall loss of energy that can result in problems during work and social activities [154]. In addition, it can manifest itself through changes in appetite and sleep patterns and a slowing of psychomotor skills. Speech in individuals with depression can be slowed, slurred, and disorganized, with an increase difficulty in articulating. At its worst, individuals can experience strong thoughts of death or even attempt suicide.

On the opposite end of the spectrum, individuals with mania and hypomania experience heightened energy and self-esteem [154]. Additionally, the individual may have difficulty weighing risks and may engage in impulsive and dangerous activities, including issues managing money. They have racing thoughts, a lack of concentration, and feel little need for sleep. Their speech is often more frequent, quicker, louder, and difficult to interrupt. One issue with mania is that people often enjoy the feelings as a sort of "high" and can have little insight into the negative consequences of non-treatment during the episode [72].

The Hamilton Depression Rating Scale (HDRS/HamD) [90] and Young Mania Rating Scale (YMRS) [218] use clinical observations of individuals to quantify the severity of depression or mania and often used to gauge the progression of bipolar disorder. Both are performed in an interview setting and are used to retrospectively gauge the individual's mood over a prior period of time. They both involve a series of questions, related to the symptoms of mania and depression, as shown in Table 1.1.

HDRS Topics	YMRS Topics
Feelings of depression	Elevated mood
Feelings of Guilt	Increased motor activity
Thoughts of Suicide	Overactive sex drive
Insomnia	Decreased interest in sleep
Interest in work and activities	Irritability during interview
Slowness of thought or speech	Increased speaking rate
Agitation and anxiety	Incoherence of speech
Loss of appetite	Delusions
Lack of libido	Aggressive behavior
Weight loss	Physical appearance
Denial of illness	Denial of illness

Table 1.1: Topics covered during the HDRS and YMRS interviews.

Each question contributes a certain number of points to a running total, depending on their answer. This final total is called the individual's HDRS or YMRS score and is a continuous measure of depression or mania, respectively.

1.3 Sources of Variability in Mood and Emotion Recognition

The detection of mood and emotion from speech is difficult, as there are many other sources of variability co-occurring with affect. This section explores different types of speech variability and gives an overviews of methods to address each.

Subject Differences: Every individual has their own particular way of expressing affect. Physiological differences, such as variations in the the vocal tract and sex can impact the dynamics of how speech is produced. Furthermore, psychological differences, including gender expression, regional accents, dialects, and different languages can impact the manifestation of affect in speech. These differences can make the detection of mood and emotion across individuals difficult, as available affect datasets tend to include smaller numbers of participants [14]. Affect recognition systems can attempt to alleviate this by explicitly removing subject differences to form a more generalized representation of affect. Techniques include feature normalization

[185, 224], sample selection [187], decision fusion [188], and incorporating auxiliary tasks [116, 161]. However, as more data becomes available for each individual over time, it may be instead better to personalize the system to participants.

Linguistic Information: Another source of variability is the content of the speech. The affect of an individual can effect both the acoustics of their speech and the language used. These acoustic changes can vary across the different types of phonemes - the smallest unit of linguistic information. Because of this, affect recognition can be improved by compensating for the lexical variability [137]. Other work has demonstrated effective mood detection by instead relying solely on the speech content using transcripts [138]. However, these approaches only give half of the full perspective. Recent work has demonstrated that a system considering both the acoustic and linguistic facets of speech is more effective than either source alone and indicates that both are needed [221].

Style: The underlying manner in which speech is prompted can also have an impact on the way affect is manifested. Acted or improvisational speech can exaggerate emotions, compared with natural speech. Prior work has shown that the perceived naturalness of speech can be considered as an additional task to improve the recognition of emotion [116]. Additionally, datasets can contain speech read in a monologue or conversations between multiple participants. In order to create an affect classifier that works effectively across multiple datasets, it will be necessary to account for these differences.

Recording Environment and Device: The ability to recognize emotion and mood can also be diminished by variability in the recording of the speech. Background noise and other conversations can obscure the subject's speech, making affect recognition challenging. Furthermore, the limitations of the recording device can introduce noise and will effect the quality of the recording. For example, phone recordings are

limited to an 8 kHz sampling rate. Autoencoders have been used to find a compressed representation of speech that removes noise [53, 54]. However, these methods typically rely on artificially introducing noise or relying on paired training example. This is often not practical for many forms of real-world environmental variability. To address this, adversarial methods have been recently introduced as a way to remove the variability due to different recording conditions, without requiring paired samples [1]. However, these methods can themselves introduce noise or have issues converging to a consistent representation [1, 110].

1.4 Contributions

My research explores the automatic recognition of emotion and mood from speech. This is often complicated in real-world circumstances, due to the many other confounding factors present in speech. Much of my work addresses this variability for either mood or emotion (Parts I and II, respectively). The remainder of my work employs these and other techniques to develop applications for mood monitoring using natural, unstructured speech (Part III). The contributions from my research are as follows:

- Addressing Variability in Mood Recognition [75, 114]
  - I demonstrated how a combination of signal processing and multi-task learning could mitigate the variability of detecting mood from speech recorded on different devices. These techniques produced significantly better performance compared with simply concatenating the device data or building specialist systems.
  - I investigated the trade-off between subject-specific and population-general information for the detection of depression from Bipolar speech collected

in-the-wild. This fusion produced an overall better system than either source of information alone.

- Addressing Variability in Emotion Recognition [74, 76]
  - I conducted the first investigation into progressive neural networks for automatic speech emotion recognition. I found that progressive nets could augment the emotion recognition task with speaker, gender, and additional datasets to improve over a system solely using emotion.
  - I introduced a new approach for more generalized representation of emotion for cross-corpus testing. While previous related methods have had issues converging, these newly introduced methods follow a "meet in the middle" paradigm for consistent convergence. I then demonstrated that these methods significantly improve on traditional methods when both laboratory and in-the-wild data are combined.
- Applications [73, 77]
  - I demonstrated that the previously found generalized model could be used to effectively classify emotion in a new domain - the natural speech of individuals with suicidal thoughts. While most previous natural speech emotion work required some amount of outside annotation, this work solely relied upon the self-ratings of participants. I then examined these in-the-wild predictions of speech emotion and found that decreased emotion variability was indicative of suicidal ideation.
  - I coordinated an outcome-based annotation of bipolar mood that identifies the need for clinical interventions. I introduced a novel framing of bipolar mood intervention in the context of anomaly detection. I then used this technique, combined with a neural network, to detect anomalous mood using both emotion and transcript-based features over natural speech.

These works demonstrate the feasibility of emotion and mood detection using natural speech by either controlling for factors of variability or adapting to subject data. The combination and extension of these works, as described in Chapter XI, will allow for tracking mood entirely from natural conversations and greatly assist in the way that mental health care is managed. Furthermore, the domain generalization method introduced in Chapter VIII and the anomaly detection framework from Chapter X have potential applications to other domains outside of speech affect monitoring.

1.5 Outline of Dissertation

This dissertation is outlined as follows. Chapter II provides background and related works. Chapter III discusses the PRIORI dataset, which is used for all mood experiments. Chapter IV describes each of the three emotion datasets used in this work.

Part I explores how to address variability in mood recognition and also contains two chapters. Chapter V covers my work on classifying mood when dealing with different types of recording devices. Chapter VI shows my investigation into leveraging both individual and cohort information for mood.

Part II is focused on addressing variability in emotion recognition and is divided into two chapters. Chapter VII describes my work using progressive neural networks to augment emotion classification with gender and speaker information and additional datasets. Chapter VIII details my work on domain generalization for emotion to improve cross-corpus recognition.

Part III explores two real-world speech monitoring application for emotion and mood. Chapter IX covers my work using the previously introduced domain generalization model to detect emotion from natural speech within a new domain – individuals with suicidal ideation. Chapter X investigates the importance of adapting a bipolar mood monitoring system to each individual and demonstrates the detection of needed

interventions using conversational speech.

Finally, Chapter XI summarizes the main findings in this dissertation and presents possible future directions.

CHAPTER II

Related Works

2.1 Methods for Speech Mood Recognition

There has been an extensive amount of speech research into the effects of depression on the psychomotor systems. Depressed speech has been described as lacking energy, monotonic, slowed, and disorganized [18]. The perception of this type of speech has been characterized by changes in the fundamental frequency, amplitude, pitch, energy, shimmer, jitter, zero crossing rate (ZCR) and harmonic to noise ratio (HNR) in the acoustics of the voice to objectively measure depression [12, 43, 48, 68, 87, 144, 150, 151, 169, 206]. Additionally, statistics on the speech and silence times as well as more complicated rhythmic features such as total vocalization time, pause variability, speaking rate, and the amounts of short pauses have proven effective [35, 70, 150, 151]. Another set of features that have been proposed are formants, which are resonant frequencies resulting from the shape of the vocal tract. Statistics of the first three formants and their coordination have also been demonstrated as a method of detecting depression severity [48, 68, 144, 211]. Additionally, estimates of the glottal source signal have been used as biomarkers successfully [144, 182]. Deviations from the baseline vowel pronunciations have also shown to correlate with depression in vowel space analysis [181]. Several methods have shown that the frequency domain may carry important information with power spectral density analysis and MFCC

statistics and their coordination as features [12, 48, 49, 68, 211]. The use of GMMs, SVMs, ANNs, and Hierarchical Fuzzy Signatures (HFS) have been used to classify depressed mood [11].

Manic speech is often characterized by being quicker, more frequent, louder, less coherent, and more difficult to interrupt (known as pressure of speech) [18]. Work has gone into determining the intensity of mania based on these rhythmic features and coherence such as the amounts of short pauses in speech [70, 200]. Additionally, some prior research implies that an increased pitch could be a sign of a hypomanic mood state [87, 206]. Huang et al. differentiated BPD from unipolar depression by detecting the presence of manic speech using an attention-based Long Short-Term Memory (LSTM) classifier [102]. However, much of the work related to mania has been at the level of differentiating it from other mental health illnesses, rather than detecting its severity. These methods usually examine speech coherence, utterance length, and total speech deviance [100, 166, 174, 199]. Because these methods do not compare with a healthy control group, it is uncertain whether or not these features will translate into mania severity determination.

2.2 Mobile Healthcare

The advent of the smartphone as a common personal device has greatly increased the everyday access many individuals have to healthcare [135]. As of 2013, over half of American adults owned smartphones [193]. Now with the use of various app stores, people have access to thousands of apps to aid in maintaining their health [135]. Of particular interest are apps that aid in the treatment of different mental health conditions. For example, the *Mobile Assessment and Treatment for Schizophrenia* (MATS) pilot study investigated the effect of text message interventions on schizophrenia [82]. The participants received and responded to multiple choice questions and most saw improved drug adherence, social interaction, and reduced auditory hallucinations.

Another study that developed the application *StressSense* found that is was possible to accurately measure the amount of stress during everyday conversational situations through audio analysis [134]. Some of the explored features were pitch, spectral centroid, speaking rate, MFCCs, and the ratio of high frequencies. The app *MoodHacker* has been developed to help individuals with depression through self tracking of mood and suggestions including meditation and physical activity. A clinical study using the app found significant effects on individuals' depressed symptoms and decreased work absences [23].

Several apps have also been created to help individuals with bipolar disorder. For example, apps such as *Mood 24/7* and *T2 Mood Tracker* provide individuals a method of self-assessment [3]. *Moodscope*, another mood tracking app, monitors a user's basic interactions with their phone to improve the app's estimation of mood in addition to the self-recorded log [128]. Some of these monitored behaviors include the use of email, SMS, web, and certain apps, as well as phone call contacts and GPS. It also provides the ability to share mood ratings with family and friends to gain support.

Recent work has used smartphones to record speech and automatically monitor bipolar disorder. Huang et al. focused on the detection of depression by asking subjects to record their speech using an app in natural environments [105]. Despite differences in noise characteristics due to device and environment, they were able to detect depressed speech using both short utterances [105] and landmark bigrams [104]. Pan et al. performed similar experiments for mania by using an app to record a patient's phone call with a psychiatrist [159, 223]. The call was conducted in a noise-suppressed laboratory environment, with the call being conversational in tone.

2.3 MONARCA

Of particular interest is the *MONARCA* project, which aims to monitor bipolar disorder using a system of sensors from wearable devices to smartphone apps [84,

156]. In particular, they have explored the benefit of passively monitoring phone speech for analysis and this section goes into further detail on these works.

Grünerbl et al. investigated the automatic recognition of changes in mood state, rather than detecting individual moods [84]. Six total subjects were enrolled for 12 weeks each and used an app to passively record their phone conversations. Subjects experienced between 1 to 3 state changes, resulting in 17 total state changes across all subjects. The ground truth was attained from in-person interviews that occurred every three weeks. Additionally, varying amounts of over the phone assessments were performed, resulting in between 5 and 9 clinical assessments per subject. A set of knowledge-based features were extracted over each day's speech and a Naive Bayes classifier was used to model mood state changes. The experiments were run on a per-speaker basis and did not involve speaker independent testing. The result of their speech-only classifier was 70% accuracy for within-subject recognition.

Faurholt-Jepson et al. improved on this work by demonstrating speaker independent detection of bipolar mood from natural speech [65]. The dataset used in the analysis was expanded to a total of 28 participants. Each participant was enrolled for 12 weeks, given the MONARCA app, and asked to participate in the same assessments as in [84]. The openSMILE *emolarge* set, including 6,552 knowledge-based features, was used to extract speech-based statistics over all days with recordings. These features were then modeled using speaker independent cross validation and a random forest classifier. Their speech-only classifier attained an accuracy of 0.68 when distinguishing between depressive and euthymic states and 0.74 when differentiating mania and euthymic states.

The MONARCA project has clinically tested the effectiveness of their app, using a combination of speech-based features, as well as other sensor data [63, 64]. These other data include measures of phone usage, social activity, physical activity, and mobility. Interventions ranged from calling the patient to give over-the-phone advice

to contacting emergency services. Patients using the app reported improved quality of life, reduced perceived stress, and a reduction in manic episodes. However, those in the intervention group had a higher risk of depressive episodes. This indicates that further work is needed to investigate both when and how interventions are conducted.

2.4 Linking Emotion and Mood

The relationship between emotion and mood has been recently gaining attention. Stasak et al. investigated the utility of using emotion to detect depressed speech [197], using the AVEC 2014 dataset [204]. However, these data were collected in a controlled environment, potentially limiting their use "in the wild". Carrillo et al. identified a relationship between emotional intensity and mood in the context of bipolar disorder [37]. However, they relied upon transcribed interviews, rather than on acoustics directly. Further work is necessitated using the acoustics of real-world conversations.

2.5 Methods for Speech Emotion Recognition

Much of the earlier work in emotion recognition followed in the footsteps of automatic speech recognition (ASR) and used generative models, such as Hidden Markov Models (HMMs) [129, 184] and Guassian Mixture Models (GMMs) [101]. Other work took a discriminative route, and trained Support Vector Machines (SVMs) over utterance-level features [101, 129, 158]. These features were often knowledge based and have contributed to the formulation of feature sets, such as eGeMAPS [61] and emobase [60].

More recent speech emotion research has focused on employing deep learning. Initial work into deep learning used deep neural networks (DNNs) over the same utterance-level features used in prior work [91]. However, deep learning is especially

suited to learning deep representations over low-level features. For example, convolutional neural networks, or CNNs, have recognized speech emotion using Mel filter banks (MFBs) [7, 106, 220]. Trigeorgis et al. demonstrated that it was possible to forgo features all together and successfully recognized speech emotion using recurrent LSTM layers over the raw audio signal [202].

However, most of this work was trained and tested using a single dataset. Such methods fail when unseen data are introduced. This can be due to differences in recording conditions, microphone quality, elicitation strategy (acted versus natural), and the distribution of labels [185]. Additionally, the demographics of subjects may widely vary between datasets [14].

One of the earliest works in cross-corpus emotion by Schuller et al. examined how acoustic and annotation differences can result in decreased performance [185]. They explored different techniques of feature normalization and found speaker-based z-normalization to work best. Additionally, they demonstrated how differences in selected sub-groups of emotions can cause large discrepancies in performance. This indicated the importance of carefully selecting annotations across all datasets in a multi-corpus experiment. Zhang et al. addressed the problem of dataset label mismatch by creating a knowledge-based mapping between classes [224]. They further explored feature normalization for utterance-level features and found that within-corpus normalization with unlabelled data boosted performance. Additional work by Schuller et al. explored how selecting only the most prototypical examples when training cross-dataset systems can improve activation classification [187]. This suggests that the most exemplar samples within datasets may also be those samples most consistently represented across datasets. Later work demonstrated how fusing the outputs of expert systems trained on individual datasets can outperform classifiers of the agglomerated data [188]. However, this performance difference depended heavily on the selected model.

Further work by researchers began exploring more complex methods of adapting features and models for more robust testing. Hassan et al. explored how previous transfer learning methods, including Kernel Mean Matching (KMM), Unconstrained Least-Squares Importance Fitting (uLSIF), and the Kullback-Leibler Importance Estimation Procedure (KLIEP), could be used to compensate for dataset differences [94]. Song et al. investigated how dimensionality reduction algorithms could be used to form a more generalized emotion feature space [195]. They found that Locality Preserving Projections (LPP), introduced in [96], resulted in the best classification performance. Abdelwahab et al. explored variations of SVMs for supervised adaptation with small amounts of target domain data [2]. Using just 9% of the data, they were able to use Adaptive SVMs and Incremental SVMs to significantly improve cross-corpus performance compared with no data.

Training deep networks over multiple tasks simultaneously has been shown to improve performance when considering cross-corpus emotion. Parthasarathy et al. explored jointly predicting activation, valence, and dominance with a DNN, considering one as the primary task and the others as auxiliary [161]. This method significantly increased cross-corpus performance compared to a single task system, especially for models with large layer sizes. Kim et al. investigated whether adding additional non-emotion tasks, such as gender and the naturalness of the expression, would improve cross-corpus performance [116]. They achieved better or comparable performance when compared with systems not incorporating the additional tasks.

CHAPTER III

PRIORI

3.1 Introduction

The long-term goal of my research is to be able to monitor the mood changes of individuals with bipolar disorder. This could eventually allow for interventions to attempt to mitigate the sometimes life-altering consequences of severe episodes. Variations in speaking patterns have been shown to be indicative of bipolar mood [154]. Smartphone conversations provide a realistic method of naturally capturing speech without disrupting individuals' everyday activities. However, there are no publicly available datasets of phone conversations of individuals with bipolar disorder, due to their sensitive nature. Because of this, it was necessary to undertake our own dataset collection - the PRIORI (Predicting Individual Outcomes for Rapid Intervention) bipolar mood dataset.

3.2 Data Collection

The PRIORI Dataset is an ongoing collection of smartphone conversational data (reviewed and approved by the Institutional Review Board of the University of Michigan, HUM00052163) [75, 111, 114, 115]. The participants are recruited from the HC Prechter Longitudinal Study of Bipolar Disorder at the University of Michigan [121].

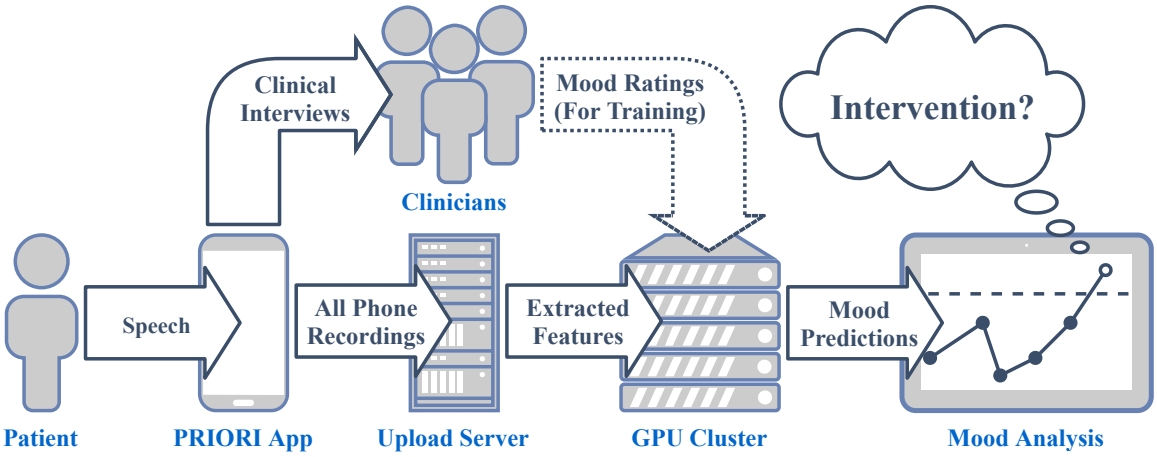


Figure 3.1: The PRIORI app and mood prediction pipeline.

The inclusion criteria are: bipolar disorder type I or II, no medical or neurological disease, and no active history of substance abuse. Participants are enrolled for six to twelve months, for an average of 32±16 weeks. In total, there are currently 51 patients and 9 healthy controls in the dataset - 18 males and 42 females. Five of the participants are African-American, three are Asian, two are American Indian or Alaska Natives, 48 are White or Caucasian, and two have more than one race. Two of the participants are in their 20's, 11 are in their 30's, 19 are in their 40's, 16 are in their 50's, and 11 are in their 60's.

All participants are provided an Android smartphone (Samsung Galaxy S3, S4, or S5) with the secure recording application (*PRIORI app*) installed. The app runs in the background and turns on whenever a phone call is made, recording only the participant's side of the dialog. The speech is encoded as 8 kHz wav files, encrypted in real-time, stored on the phone, and then uploaded to a HIPAA-compliant server, as depicted in Figure 3.1. The participants were asked to use the smartphone as their primary device throughout their time on the study. The data include phone calls collected only when an individual was not using the speaker phone, to ensure that other people besides the participants are not recorded.

Participant mood is assessed weekly, over the phone, by a member of the study

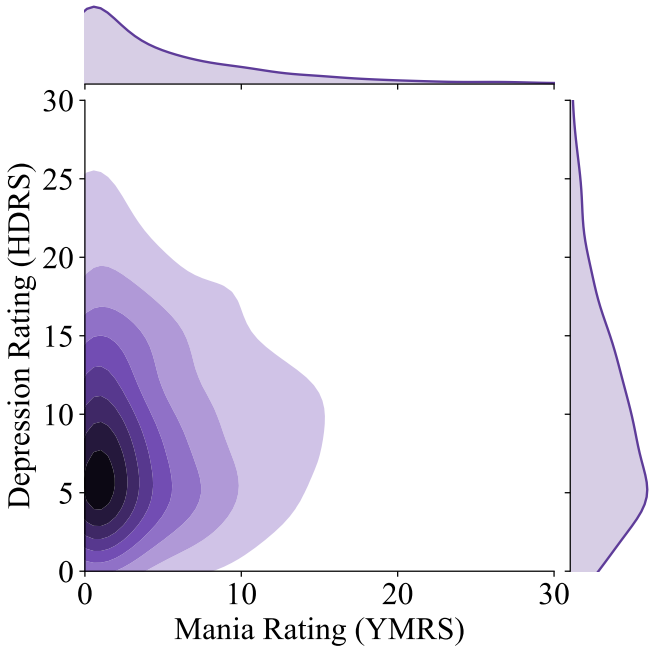


Figure 3.2: The distribution of HDRS and YMRS ratings in PRIORI.

team using the Hamilton Depression Rating Scale (HDRS) [90] and the Young Mania Rating Scale (YMRS) [218]. These calls are referred to as *assessment calls* and are structured in an interview format. The dataset includes 1,268 recorded weekly clinical assessments, out of 1,516 total assessments performed. This discrepancy is due to participants calling in on a device other than the study phone, as well as other recording issues. Of the recorded weekly assessments, 23 were transcribed for algorithm development. All other calls outside the clinical context that make up a participant's everyday conversations are referred to as *personal calls*. The dataset includes 51,970 phone calls totalling 3,997 hours of recordings. We manually transcribe a subset of 25 hours of speech for feature development.

The HDRS and YMRS scales are continuous measures of mood, ranging from a score of 0 (not symptomatic) to 36 (highly symptomatic). Figure 3.2 depicts the distribution of assessment ratings combinations from the 51 patients in the dataset. In our earlier mood experiments, we treated the prediction problem as classification, binning the HDRS and YMRS into categories of symptomatic (depressed or manic,

Mood	HDRS	YMRS	Number	# Per Subject
Euthymic	$\leq$6	$\leq$6	612	11.3 ± 7.8
Manic	<10	$\geq$10	125	3.9 ± 3.9
Depressed	$\geq$10	<10	416	9.5 ± 8.1
Excluded	Else	Else	363	8.1 ± 6.1

Table 3.1: Mood state categories defined by HDRS and YMRS measures, including the number of total assessments and the mean and standard deviation of assessments per subject.

respectively) and asymptomatic (euthymic). A call is labelled *euthymic* if it has a score of six or less on both the HDRS and YMRS scales; *manic* if the score is ten or greater on the YMRS and less than ten on the HDRS; and *depressed* if the score is ten or greater on the HDRS and less than ten on the YMRS. All other assessments are excluded from the classification experiments (Table 3.1). Experiments in Chapter X instead use the continuous mood ratings and do not have the same restriction.

The large standard deviations seen in Table 3.1 demonstrate the widely varying amounts of mood episodes between individuals with bipolar disorder. Additionally, some individuals have disparities among the proportions of times spent in each mood. For example, one participant experienced 27 weeks of euthymia and two weeks of mania. This demonstrates the need for mood monitoring systems that learn from and adapt to individuals, as explored in Chapter X.

3.3 Clinical Annotation

We created a new annotated subset of the PRIORI data, called the PRIORI Annotated Mood dataset (PRAM) to better understand when and why interventions are needed. We define *interventions* as changes in the treatment plan, such as emergency room visits, hospitalizations, or drug modifications to negative mood events. This definition corresponds to the concept of *necessary clinical medication adjustments* – an important measure of illness stability as reflected in the number of alterations in care considered necessary [24].

Clinical Summary: The University of Michigan electronic health record system was accessed to review relevant health information available for each subject during their time in PRIORI. This included all clinical notes and laboratory results from the Michigan Medical system, as well as other health systems that shared their records. Each subject's data was summarized to include relevant information about subject mood, clinical condition, or recent changes in either.

Flagging Application: We developed an application to review the clinical data for each week of participation in the PRIORI study and identify decision-making points for interventions (Figure 3.3). Annotation is divided into weeks based upon the date of each YMRS and HDRS retrospective interview. The ratings are displayed for each week, as well as the suicide sub-score from the interview and day-of mood ratings, if available. The application allows annotators to browse all mood ratings and clinical summaries up to and including the day of the assessment, but prevents seeing future information. This ensures that annotators can only make decisions based on the data that would have been available at the time when a decision to intervene was made. The application allows the annotators to flag or not flag each week for intervention and then rate their confidence on a 1-3 scale (weak to strong). Annotators can indicate whether the intervention is urgent (needed within 24 hours) or a non-urgent follow-up. Once the annotation for a given week has been submitted, the application advances to the next week and previous submissions cannot be modified.

Annotation: Four clinically trained members (3 PhD and 1 MD) of the bipolar research team annotated all available clinical and research data. Each session consisted of a group of at least two annotators, who came to a consensus on the need for intervention. Groups were given summaries of clinical data before each subject's enrollment in PRIORI to establish a baseline summary of their medical history. Annotators were asked to complete subjects' entire annotations in one session to maintain consistency. Presently, 26 subjects totalling 555 weeks (71 with flagged interventions)

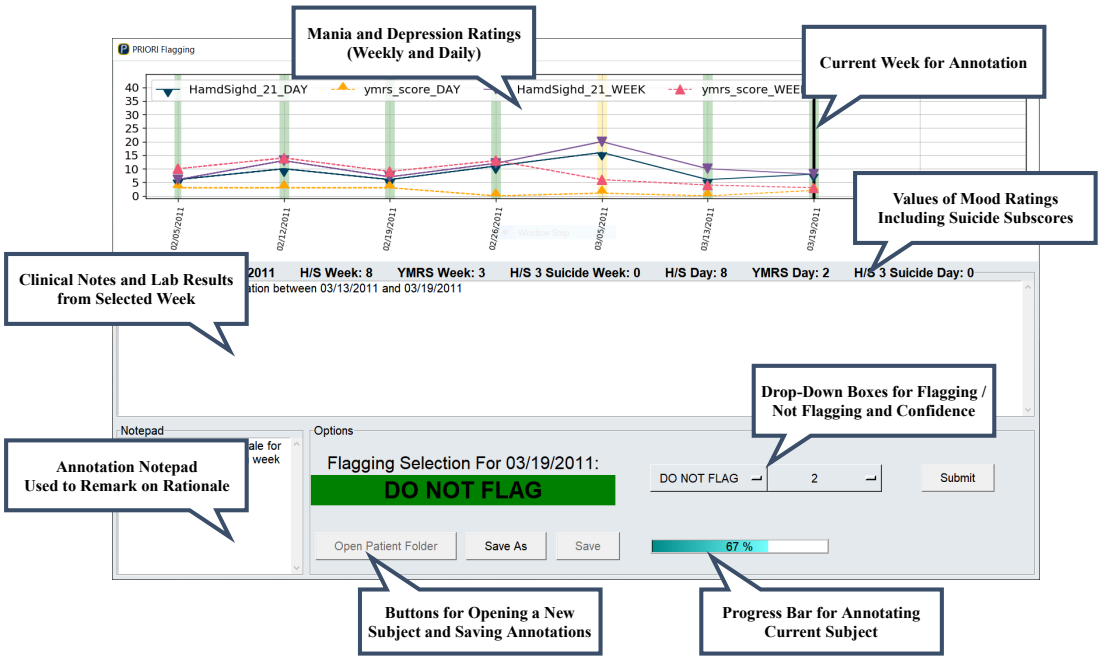


Figure 3.3: The application used to read clinical data and flag interventions for each subject. The top of the application provides current and prior week-retrospective and day-of YMRS and HDRS ratings. The middle displays any clinical notes or lab results since the prior week. The user is able to click dates in the above graph to view notes for other weeks. The bottom is used to mark whether or not to flag a week for intervention and the confidence of the rating (1-3). The application advances to the next week upon submission and entries cannot be modified afterwards.

have been annotated. Annotators were asked to provide reasons underlying the flagging to learn contributing factors that led to intervention recommendations. Common rationale used to flag for interventions include:

- High YMRS or HDRS, compared with personal baseline
- Lack of improvement from previous weeks
- Severity of clinical symptoms, e.g., suicidal thoughts

The PRAM dataset is further investigated in Chapter X to determine when interventions are warranted.

CHAPTER IV

Emotion Datasets

4.1 IEMOCAP

The "Interactive Emotional Dyadic MOtion Capture Database" (IEMOCAP) was created to explore the relationship between emotion, gestures, and speech. Ten actors (five male and five female) were recorded over five sessions. Each session consisted of a male and a female performing given either a series of scripts or improvisational scenarios. During the session, motion capture markers were attached to just one of the actors at a time. Once all scripts and improvisations were performed, the other actor was given the motion capture markers and the whole process was repeated. The audio was recorded using two high quality shotgun microphones at a 48 kHz sampling rate and later downsampled to 16 kHz.

The data were segmented by speaker turn, resulting in 10,039 total utterances (5,255 scripted turns, 4,784 improvised turns). Segments were then annotated for emotion, including valence and activation on a 1 to 5 scale. Discrete emotions were also annotated, including happiness, anger, sadness, frustration, excitement, disgust, fear, surprise, and neutral. Between two and four annotations were performed per utterance. Further information about the IEMOCAP dataset can be found in [31].

4.2 MSP-Improv

The MSP-Improv dataset aims to capture more naturalistic emotion from improvised scenarios, while also partially controlling for lexical content. The collection involved a total of twelve actors (six male and six female). Like IEMOCAP, the dataset is split into six sessions, each including interactions between one male actor and one female actor. Each actor wore a collar microphone to record speech at 48 kHz (later downsampled to 44.1 kHz).

MSP-Improv controls for lexical content by including specific "target sentences" with fixed lexical content that can be embedded into different emotional scenarios (i.e., angry, happy, sad, neutral). In each pair, one of the actors was tasked with ensuring that the target sentence was spoken in each scenario. Once all target sentences and scenarios were recorded, the actors switched roles and the second actor assumed this responsibility. Using this method, the researchers were able to control for lexical content, while still allowing for more natural emotion expression.

The data was divided into 652 target sentences, 4,381 improvised turns (the remainder of the improvised scenario, excluding the target sentence), 2,785 natural interactions (interactions between the actors in between recordings of the scenarios), and 620 read sentences (emotional readings of the target sentences). This totaled 8,438 utterances over 8.9 hours. These utterances were then annotated for emotion using crowd-sourcing on Amazon Mechanical Turk. Valence and activation were rated on a scale from 1 to 5, as well as the discrete emotions of happiness, anger, sadness, and neutral. There is a minimum of five annotators per utterance up to a maximum of 50 (median of 5). Please refer to [34] for additional information about the MSP-Improv dataset.

4.3 PRIORI Emotion

The PRIORI Emotion Dataset is an affect-annotated subset of the larger PRIORI bipolar mood dataset, described in Chapter III. The goal of the annotation is to allow for the exploration of the relationship between mood state and emotion expression, as there were no other natural smartphone conversational speech datasets annotated in this manner. The PRIORI Emotion Dataset contains manual valence/activation annotations of both assessment and personal calls. We use a dimensional labeling strategy [28], motivated by the concept of *core affect* [175]. This construct provides a de-contextualized manner of considering emotion expression.

The PRIORI Emotion Dataset includes natural conversational speech from 12 subjects, seven females and five males, totaling 11,337 calls (928 hours). The selected subjects are between 24 and 63 years old. We selected the subjects based on three factors: (1) bipolar disorder diagnosis, which allows us to examine the link between emotion and bipolar mood (future work will focus on healthy controls); (2) used Samsung S5, which provides microphone consistency, lack of which was identified as a challenge in our prior work [75]; (3) provided informed consent for annotation of personal calls, which allows us to generate ground-truth emotion labels.

We then annotate a subset of these data using: (1) segmentation, (2) segment selection, (3) segment inspection, and (4) segment annotation. We explain each in the following sections.

Segmentation: Calls are segmented using a noise-robust method by Sadjadi and Hansen [178] which compensates for variations in background noise. Their algorithm extracts five representations of speech likelihood including: harmonicity, clarity, prediction gain, periodicity, and perceptral spectral flux. Principal Component Analysis (PCA) is performed to combine them into a single signal by taking the largest eigenvalue. We extend this approach by first converting this signal into contiguous speech segments. We then smooth the signal with a Hanning window of 25ms and normalize

it by subtracting by the 5th percentile over the call and dividing by the standard deviation to ensure comparability between calls. Segments of 25ms are created whenever this signal exceeds a 1.8 threshold. This forms a set of overlapping segments which are merged, removing any silence less than 700ms. We determined these parameters by validating over the transcribed assessments. Segments longer than 2s are further divided into subsegments of 2s with 1s overlap. Constant segment size is used to ensure that variations in features are only due to variations in rhythm [75].

Segment Selection: We identified a subset of segments for manual annotation from the assessment and personal calls. Our first filter was for segment length, to increase the likelihood that segments contained sufficient data to assess, but were not so long that the emotion would vary over the course of the segment. We exclude segments shorter than three seconds and longer than 30 seconds. Next, we sampled from both personal calls and assessment calls. Assessment calls are important because they are the only calls that are directly associated with mood labels. Personal calls are important because they contain natural unstructured speech. Therefore, we sampled from both to ensure a diversity of examples. For each assessment call, we select up to ten random segments. For personal calls, we sample as a function of proximity in time to assessment calls, preferring those that occurred closer to the assessments. We select 1,200 segments randomly considering the weight of $max(4 - d, 1)$, where d is the number of days between the call and its future assessment day. Calls on the day of assessment receive a weight of four (these are most closely linked to the HDRS/YMRS score). Other calls receive a weight that reduces linearly up to 3 days before assessment day, calls outside this range have a weight of one. This results in 17,237 segments, 2,837 and 14,400 segments from the assessment and personal calls, respectively.

Segment Inspection: We manually inspected each segment prior to annotation and removed those that were deemed inappropriate for the annotation task, if: (1)

background noise dominates the speech signal, (2) speech content of the segment lasts less than two seconds, (3) subject is not talking to the phone (e.g., talking to someone else in the room), (4) emotion clearly varies over the course of the segment, and (5) segment contains identifiable information (e.g., name, address, phone number, etc.). This results in **13,611** segments (**25.20** hours), 2,209 and 11,402 segments from the assessment and personal calls, respectively.

Segment Annotation: We annotated the activation and valence of the 13,611 speech segments using the established pictorial manikins method across a 9-point Likert scale (1: very low to 9: very high) [28]. There were 11 annotators (7 female, 4 male) aged between 21 and 34 and native speakers of English.

We conducted a training session for each annotator, including a training video and manuscript, to introduce the annotation software and provide annotation examples. In the training session, annotators were asked to consider two important points:

1. Although challenging, we asked that annotators to only consider the acoustic characteristics of the recordings, not the lexical content. They were asked to avoid letting speech content "color" their activation and valence labels.

2. We asked that annotators consider the subject-specificity of emotion expression. When approaching a new subject, annotators were asked to spend some time listening to a few segments without assigning a rating in order to get a better sense of what that person's baseline sounds like.

We further supported the assessment of subject-dependent emotion patterns by providing *individual context* for each participant. The annotation software randomly selected a participant and presented all segments of that participant, in random order, to the annotator before moving on to the next participant's segments. In this way, annotators can consider participant-specific features to define emotion labels more accurately.

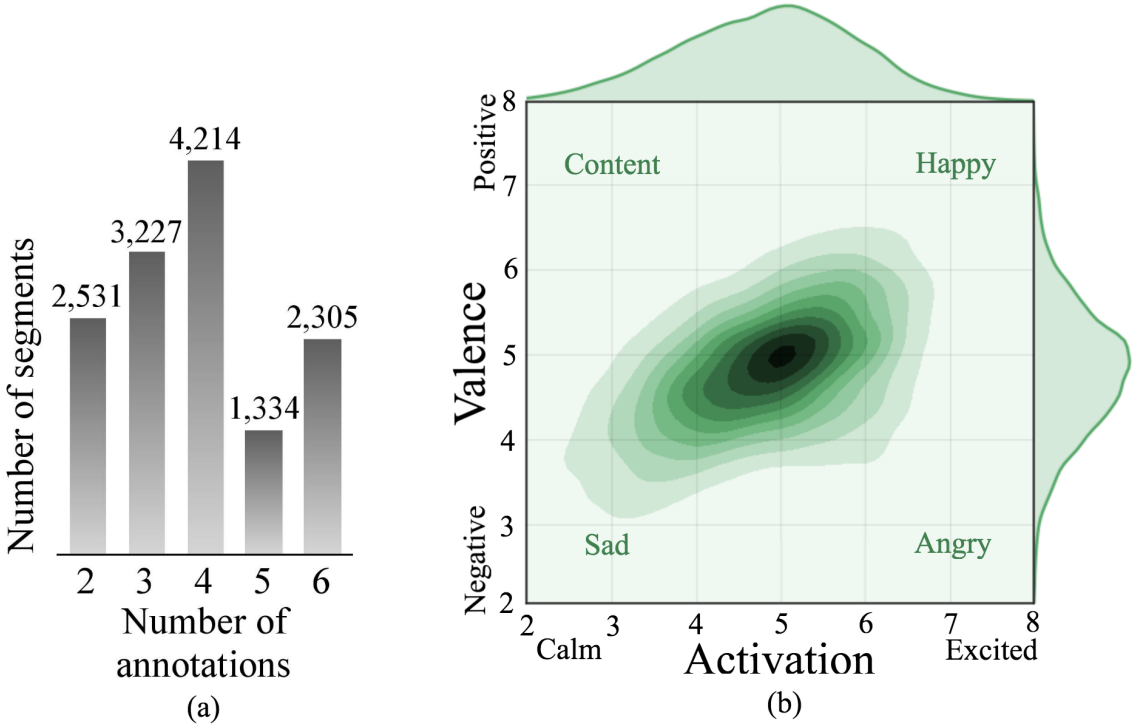


Figure 4.1: (a) Distribution of the number of labels annotated for the segments. (b) Distribution of the activation and valence ratings in the PRIORI Emotion Dataset. Categorical labels are provided only as reference points for the four quadrants.

We collected between two and six labels for each segment (**3.83** $\pm$ **1.31** labels per segment). Figure 4.1 shows the distribution of the number of annotations for each segment. See Figure 4.1 for a distribution of the activation and valence labels defined by the annotators. We found that the activation and valence values are significantly correlated with a PCC of 0.46 ($p < 0.01$).

4.4 EMASS

The Ecological Measurement of Affect, Speech, and Suicide (EMASS) Dataset is a collection of natural smartphone speech and momentary self-ratings. The collection is ongoing, and the current snapshot includes 43 individuals, each enrolled for eight weeks. Participants were divided into four groups - healthy controls (HC), psychiatric controls (PC), and individuals that have experienced recent suicidal ideation (SI) or

	All	HC	PC	SI	SA
Subjects	43	19	7	12	4
Calls	4078	1780	761	1208	295
Hours	402	239	51	93	15

Table 4.1: The amounts of data from different groups of subjects, including healthy controls (HC), psychiatric controls (PC), and individuals hospitalized for suicidal ideation (SI) and attempts (SA). Subjects must contain at least five calls to be included.

suicide attempts (SA). Individuals in the SI and SA groups were admitted to the hospital for thoughts or behavior related to suicide. Individuals in the PC group were admitted to the hospital for reasons other than suicide (e.g., substance use). All groups, with the exception of HC, were enrolled in the study during their psychiatric admission. Immediately following discharge, they were given a smartphone with the PRIORI app, which securely records their end of phone conversations [111]. These recordings are then encrypted and uploaded to our server for automatic analysis. Table 4.1 shows the number of calls collected for each subject group, for a total of 4,078 calls over 402 hours.

In addition to PRIORI, the mEMA app by ilumivu was installed on the smartphone, which presented participants with three surveys throughout the day at random times. They were also asked to initiate surveys if they experienced suicidal ideation or behavior. In these surveys, participants were asked to report on their current affect, using items from the Positive and Negative Affect Schedule (PANAS-X) [210]. Affect was rated on a five point Likert Scale for 11 different categories, which are divided into three groups, based on previous work [17]. The three groups are: **(1) Positive Emotion** - Confident, Excited, Happy; **(2) Negative Emotion** - Sad, Guilty, Worried, Shame, Hopeless; **(3) Anger/Irritability** - Anger at Others, Anger at Self, and Irritable. There are 3,359 surveys included in the dataset snapshot used in this thesis.

Part I

Addressing Variability in Mood Recognition

CHAPTER V

Mood Recognition: Device Variability

5.1 Introduction

While others have explored automatic bipolar mood recognition from speech with some success, the experiments are often constrained in some way that doesn't reflect real-world conditions. Experiments may be performed in the laboratory or only use one type of recording device. In this chapter, I present an investigation into automatic speech analysis using mobile phone conversations as a way to predict mood, as well as the complications that arise from the diversity of real world recordings on different types of devices.

Research has demonstrated that speech patterns are affected by mood and contribute to accurate clinical assessments [154]. For example, both the Hamilton Depression Rating Scale (HDRS) [90] and Young Mania Rating Scale (YMRS) [218] use clinical observations of speech to determine the severity of depression or mania [90, 218]. There is an opportunity to discover how speech cues can be automatically processed to augment objective measures available in clinical assessments. Mobile phones provide an effective platform for naturally monitoring these speech cues and have shown promise for bipolar disorder [111, 128, 156]. However, changes in recording quality between different types of phones can severely decrease the predictive capabilities of a system. These include clipping, loudness, and background noise.

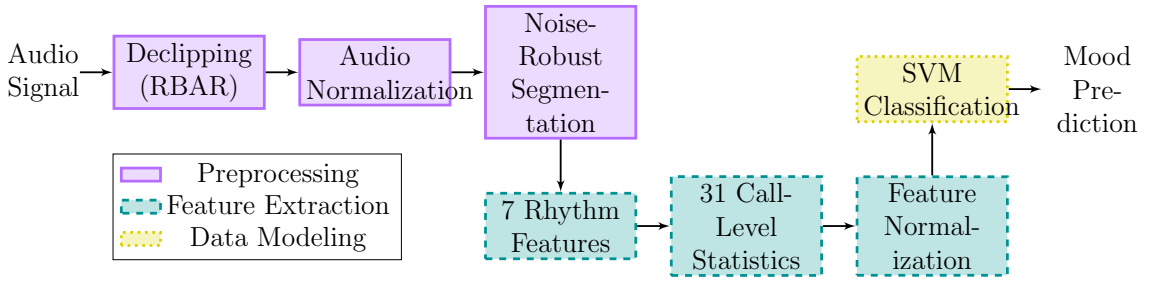


Figure 5.1: Audio pipeline divided into three stages of preprocessing (Section 5.3), feature extraction (Section 5.4), and data modeling (Section 5.5).

Much mood speech research has focused on identifying features for recognizing depression. Among these, are pitch, energy, rhythm, and formants [48, 68, 70, 87, 169, 206]. Short pauses and increased pitch have been correlated with mania [70, 87, 166, 174, 206]. However, much of the work in identifying mania from speech has focused on differentiating it from schizophrenia and cannot be directly applied [100, 199]. Many mood related studies collected their speech from controlled environments [48, 87, 206] or used a single type of recording device [85, 128, 156] and do not necessarily reflect the noise and microphone quality present in real world recordings. As such, their models would be difficult to apply to a widely distributed mobile health system.

In this chapter, I focus on one of the challenges associated with real-world distributed mood recognition: variability in recording. I examine the differences between the two phones used in this study and analyze preprocessing and modeling methods that allow us to build models of mood across the database as a whole. These methods include declipping [93], noise-robust segmentation [178], feature normalization [48], and multi-task learning [58]. I provide evidence that mood-related changes in speech are captured in this model using the structured assessment calls captured from different phone types. Please see Figure 5.1 for a system overview.

The novelty of my approach is the investigation into acoustic variations caused by recording with different types of phones and the preprocessing and modeling changes

Mood	Total	# Per Subject	% Per Subject
Euthymic	275	7.9±7.7	30%
Manic	107	3.1±4.0	12%
Depressed	247	7.1±7.5	28%
Mixed	95	2.7±3.6	13%
Excluded	175	5.0±4.7	17%

Table 5.1: Distribution of assessment classes of mood used in this chapter's experiments. The total number of observations of each mood class is given. The mean and standard deviation of observations for each class per subject is shown, along with the average percentage of each.

necessary to detect mood under these conditions. My results suggest that this pipeline of methods including preprocessing, feature extraction, and data modeling can effectively increase the performance of these types of mixed device systems. The results show a significant increase in performance from AUCs of 0.57±0.25 and 0.64±0.14 for manic and depressed, respectively, to 0.72±0.20 and 0.75±0.14, highlighting the importance of proper processing of acoustic data from multiple sources.

5.2 Data

The experiments in this chapter use an earlier snapshot of the PRIORI Bipolar Dataset (explained in Chapter III). The snapshot contains 37 participants who have made 34,830 calls over 2,436 hours. Each participant has been on the study for an average of 29.2 weeks with a standard deviation of 16.4 weeks. Additionally, there have been 780 recorded weekly clinical assessments. Only these structured calls are used in this study. The distribution of assessment classes is shown in Table 5.1.

The Samsung Galaxy series of phones, including the S3, S4, and S5 are used by participants. Only two of the participants were given S4s and their data are excluded from this study. The distribution of subjects with S3s and S5s can be seen in Table 5.2. The two models of phone include model-specific microphones and processing. One of the effects of this recording and processing is clipping. Clipping occurs most

Phone	#Subjects	#Assess.	%Clipped	RMS	SNR_{dB}
S3	18	456	2.74%	0.397	21.2
S5	17	287	0.02%	0.066	25.1
Both	35	743	1.69%	0.269	23.1

Table 5.2: Differences in data amounts and acoustics between the Galaxy S3 and S5. The percent clipped assessments (Assess.) and the mean percent of samples per call clipped are shown. Root mean square (RMS) values are calculated to show the loudness for each device microphone. Signal to noise ratio (SNR) is calculated as the relative power in the speech verses silence regions in decibels (dB).

often in the S3, with an average of 2.74% of speech samples at maximum range. This sensitivity is also demonstrated by the average root mean square value of 0.397 for the S3. Additionally, the noise is much more pronounced, as seen in the lower signal to noise ratio of 21.2 dB for the S3.

5.3 Preprocessing

The two phones used in this study have different acoustic properties. The S3, compared to the S5, has more clipping, higher volume, and a sensitivity to background noise. Because of this, it is necessary to carefully preprocess the data before feature extraction using declipping, audio normalization, and noise-robust segmentation in order to make calls from different devices more comparable.

Declipping: The declipping algorithm *Regularlized Blind Amplitude Reconstruction* (RBAR) [93] was used to approximate the original signal. This is a closed form solution approximation of an algorithm called *Constrained Blind Amplitude Reconstruction* (CBAR) [92]. Each algorithm extrapolates the clipped sections of audio beyond their original values, while minimizing the second derivative of the signal, and have been shown to improve the performance of automatic speech recognition [92, 93]. Both algorithms ignore unclipped regions, beneficial for audio recordings that have variable amounts of clipping, as seen in Table 5.2.

Audio Normalization: The audio signal is scaled by dividing by the maximum

absolute value. This ensures that the signal ranges from -1 to 1, which is necessary after running declipping, as it extrapolates the signal beyond these bounds. It also ensures that the loudness between the two phone types, as seen in Table 5.2, is more comparable.

Segmentation: Each call is segmented using an extension of Sadjadi and Hansen's algorithm [178], which is robust to variation in noise. This is necessary, given the differences in SNR between the phones (Table 5.2). The algorithm extracts five signals representative of speech likelihood, including: harmonicity, clarity, prediction gain, periodicity, and perceptual spectral flux. These are then combined using principal component analysis (PCA). The final signal is the largest eigenvalue. It is smoothed by a Hanning window of 25ms and normalized by subtracting by the 5th percentile over the call and dividing by the standard deviation. This ensures that signals from different calls all share a similar silence baseline. Segments of 25ms are created wherever the combo signal exceeds a 1.8 threshold. Overlapping segments are merged and silences less than 700ms are removed. These parameters were found by validating over the transcribed assessments for segment alignment. Segments are further divided into subsegments of 2s with 1s overlap. Segments less than 2s are discarded. Constant window sizes are used to ensure that variations in the features are not caused by changes in segment size [201]. The full segmentation process is shown in Figure 5.2.

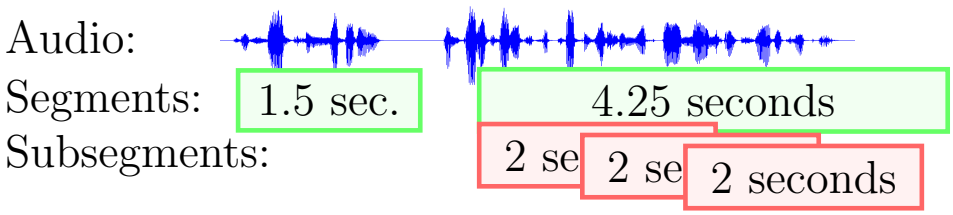


Figure 5.2: Segments of speech are found. Segments of 2 seconds or longer are divided into subsegments of 2 seconds in 1 second steps.

5.4 Feature Extraction

Rhythm Features: Individuals in manic or depressed episodes exhibit changes in the rhythm of their speech [81]. Rhythm features are calculated for each subsegment by first extracting the voicing envelope. The envelope is used to calculate the spectral power ratio and spectral centroid. The envelope is decomposed into two intrinsic mode functions (IMF) using empirical mode decomposition [103]. Tilsen and Arvaniti [201] empirically demonstrated that the extracted IMFs are reflective of syllable- and word-level fluctuations. The IMFs are used to extract five segment-level features: the power ratio between the two IMFs and the mean and standard deviation of the instantaneous frequencies associated with each IMF.

Call-Level Statistics: The seven rhythm features are transformed into call-level features by taking the mean, standard deviation, skewness, kurtosis, minimum, maximum, range, and 1^{st}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 99^{th} percentiles of the subsegment measures. Additionally, the differences between the 50^{th} and 25^{th}, 75^{th} and 50^{th}, 75^{th} and 25^{th}, 90^{th} and 10^{th}, and 99^{th} and 1^{st} percentiles are included. This set is augmented with the percentage of the call that is above 10%, 25%, 50%, 75%, and 90% of the range. Finally, the call-level feature trend is captured by fitting a linear regression model to the features extracted over each segment (R^2, mean error, and mean squared error). This results in a total of 217 features.

Feature Normalization: Call-level features are Z-normalized either (1) globally, using the mean and standard deviation of all training data, or (2) by subject, using the mean and standard deviation of each subject's own data. Previous research has shown that normalization by subject can reduce the disparity between subject feature distributions caused by speaker differences and aid in the detection of mood [48]. This method may also help reduce some of the differences in subject feature distributions due to differences in phones.

5.5 Data Modeling

The classification goal is to identify if a given call is (1) from a manic or euthymic episode or (2) from a depressed or euthymic episode. Subjects are only included in analysis if they have at least six total assessments in order to ensure enough data to process features by subject. Additionally, subjects must contain at least two euthymic calls and two manic/depressed calls. This ensures that there is enough data to measure test performance. With these restrictions, 15 subjects are used when considering mania (12 S3s and 3 S5s) and 18 subjects are used when considering depression (11 S3s and 7 S5s).

Support Vector Machines (SVM) [45] are used to classify the speech. SVMs learn a decision boundary between two classes of data with an explicit goal of identifying a boundary that maximally separates the two classes. The classifiers are implemented using both linear and radial basis function (RBF) kernels. Euthymic samples are given a weight equal to the number of manic/depressed samples divided by the number of euthymic samples. Manic/depressed samples are given a weight of one. This ensures that there is no bias towards the mood with more samples by increasing the penalty for misclassification of minority labels. Multi-task SVMs [58] are also used for certain experiments. This algorithm weights the kernel function using a parameter *rho* in order to decrease the importance of data from a different task. In this case, the task is considered to be the phone type. On one extreme, rho can be selected to behave as a single-task SVM and consider the tasks to be equal. On the other extreme, the selected rho can consider the tasks to be completely independent.

The models are trained using leave-one-subject-out cross-validation, ensuring that there is no overlap between the speakers used to train and test the system. The model parameters include: kernel type (RBF vs. linear), gamma (RBF only), number of features with respect to a ranked list, cost parameter (C), and rho (multi-task only). The parameter combination is chosen to optimize leave-one-training-subject-

out cross-validation, where the contribution of each training subject is proportional to his/her amount of data.

Features are ranked using a heuristic of Weighted Information Gain (WIG). The heuristic was chosen due to the observed subject-specific label imbalance, which may result in the identification of features that are tied to subject identity, rather than mood. This can result in a classifier learning to associate all instances with a single mood state from a biased subject. WIG allows for each sample to be ascribed an importance that ensures both classes contribute equally from each subject. This is implemented using the weighted entropy functions described in [192]. Each sample is given a weight equal to the total number of samples in its subject divided by the number of occurrences of its label in its subject. This ensures that minority and majority samples are given equal weight over each subject, while subjects are given weight proportional to their number of samples.

The performance was measured using Area Under the Receiver Operating Characteristic Curve (AUC). AUC assesses the ability of a system to correctly rank pairs of instances from opposing classes. It has a chance rating of 0.5 and ideal rating of 1.

5.6 Results and Discussion

In this section I demonstrate the ability to differentiate between euthymic and symptomatic moods, despite using two types of mobile phones with different acoustics. The results are presented in Table 5.3. In addition to reporting the combined test performance of both phone types, results are broken down into individual types. However, all phones from both types are always used to train models. A paired t-test with a significance of 0.05 is used to compare results to baseline performance and a significant difference is marked with an asterisk and bolded.

Baseline Performance: The baseline system uses global normalization and does not include declipping. The results in Table 5.3a show an AUC of 0.64±0.14 for

Model	Manic AUC	Depressed AUC
S3	0.52±0.22	0.66±0.17
S5	0.78±0.31	0.62±0.09
Both	0.57±0.25	0.64±0.14

(a) No Declipping and Global Normalization (Baseline)

Model	Manic AUC	Depressed AUC
S3	0.68±0.16	0.62±0.14
S5	0.79±0.21	0.69±0.18
Both	**0.70±0.17***	0.65±0.15

(b) RBAR Declipping and Global Normalization

Model	Manic AUC	Depressed AUC
S3	0.73±0.22	0.74±0.10
S5	0.79±0.37	0.80±0.21
Both	**0.74±0.24***	**0.77±0.15***

(c) No Speech Segmentation (Silence Included)

Model	Manic AUC	Depressed AUC
S3	0.66±0.15	0.73±0.15
S5	0.71±0.35	0.78±0.10
Both	**0.67±0.19***	**0.75±0.14***

(d) No Declipping and Subject Normalization

Model	Manic AUC	Depressed AUC
S3	0.67±0.20	0.67±0.21
S5	0.72±0.41	0.65±0.11
Both	**0.68±0.23***	0.66±0.18

(e) Multi-Task SVM Using Baseline Preprocessing

Model	Manic AUC	Depressed AUC
S3	0.71±0.19	0.66±0.14
S5	0.78±0.23	0.79±0.13
Both	**0.72±0.20***	0.71±0.15

(f) Multi-Task SVM Using Best Preprocessing

Table 5.3: Classification results using various methods. **Bolded*** AUCs denote results significantly better than baseline (paired t-test, p=0.05).

depressed and a near chance performance of 0.57±0.25 AUC for manic. However, the three S5s performed better than the S3s in the manic test with 0.78±0.31 AUC. This could indicate that even though the S5 only makes up 20% of the phones, its higher quality recordings allow for it to perform well in testing. Alternately, the speaker population that makes up those subjects using the S5s could be more homogeneous. The S5 continues to outperform the S3 in the rest of the manic experiments.

Evaluation of Declipping: Table 5.3b shows the results of declipping when using global normalization. While the performance of the depressed tests remain mostly unaffected, the manic test increases significantly to an AUC of 0.70±0.17. This is due to the improvement in the S3, where larger amounts of clipping occurred, as seen in Table 5.2. I hypothesize that the stronger improvement in manic tests, compared with depressed tests, is due to the fact that manic S3 calls have significantly more

clipping than euthymic and depressed S3 calls (unpaired t-test, p=0.05). The percent of clipping in euthymic, manic, and depressed S3 calls are 2.73±1.25%, 3.21±1.13%, and 2.41±1.07%, respectively.

Evaluation of Segmentation: The effect of segmentation was studied by eliminating the algorithm described in Section 5.3. Instead, the 2 second subsegments were taken over the entire call - silences included. It performed the best of all tests with significant increases from the baselines for both moods (Table 5.3c). However, I hypothesize that this is actually due to the rhythm features indirectly capturing information about the assessment structure. For example, an individual who is euthymic would have more silence due to their brief interview answers. This highlights one of the potential pitfalls to avoid when working with structured calls to train a model to recognize acoustic aspects of mood. For this reason, it is necessary to use accurate segmentation to avoid these misleading results.

Evaluation of Feature Normalization: Normalization by subject significantly increased the performance of both manic and depressed tests from baseline, as shown in Table 5.3d. This method has the ability to correct for different feature distributions among speakers, as explained in [48]. These results demonstrate that this correction can also benefit systems with variable recording devices of different quality.

Multi-task SVM Analysis: The use of a multi-task SVM can also control for the variability in device types by giving lower weight to data from different phone types and higher weight to data from the same phone types. Table 5.3e shows a significant improvement in manic from baseline by selecting a low value for rho and treating data from across different phone types as less informative. Depression does not see much improvement, as a high rho value is selected, indicating that the data is already comparable without preprocessing. This gives further evidence to the reason preprocessing works well for manic speech but has little effect on depressed speech. Another multi-task experiment was run using the preprocessing methods that

worked best for each mood - RBAR declipping and subject normalization for manic and subject normalization for depressed. These results can be seen in Table 5.3f, with the highest manic AUC of 0.72±0.20, which is significantly better than baseline.

5.7 Conclusion

This chapter presents methods to improve the comparability of data collected from across devices of different acoustics. This is essential for any mobile health system using speech that aims to be widely distributed, as the prospect of varying audio quality is unavoidable. My results demonstrate that through certain preprocessing, feature extraction, and data modeling techniques it is possible to mitigate the effects of differing amounts of clipping, loudness, and noise. This is best shown by the increase in performance from the baseline AUCs of 0.57±0.25 for manic and 0.64±0.14 for depressed to the significantly higher AUCs of 0.72±0.20 and 0.75±0.14, respectively. This excludes the results without segmentation, as those features capture the structure of the mood interview instead of the characteristics of the speech. There was not a comprehensive solution for both mood types, which indicates the need for careful consideration of all steps along any pipeline.

CHAPTER VI

Mood Recognition: Individual Versus Cohort

6.1 Introduction

Mood recognition systems need to control for other factors when considering real-world data. While Chapter V explored the impact of recording device differences, this chapter instead examines the contribution of individual subject characteristics. I again use speech gathered from mobile phone conversations during the *PRIORI* project (Chapter III) to predict depression in patients with bipolar disorder. The proposed techniques exploit both *population-general and subject-specific* knowledge.

Speech is modulated by the mood of an individual [154]. In particular, the Hamilton Depression Rating Scale (HDRS) [90], lists speech retardation as one of the indicators of depression. Previous work showed that it is possible to augment the clinical diagnosis of depression with objective rating by automatic detection from speech [47, 87, 206]. This could be used to help to better target care to those most in need and help in areas with scarce resources [15]. However, variations in individual symptoms make the adoption of such a system difficult.

Many computational models have been proposed to predict depression from speech [11, 47, 85, 87, 98, 133, 142, 189, 206]. These models are normally trained to capture common patterns in cohorts due to limitations in the size of available datasets. For example, [11] explores the performance of common speaker-independent classifiers

for detection of depression. Additionally, [98] compares Guassian Mixture Model (GMM) and Support Vector Machine (SVM) classifiers using only formant frequencies and their dynamics. Subject-independent systems using i-vectors have recently been proposed for depression detection [47, 133, 142, 189]. One important issue of using i-vectors for depression detection is the limited dataset available [46]. [47] proposes an oversampling approach to increase the number of available utterances.

There are a limited number of systems that leverage speaker-specific information in their classifiers. The work reported in [85] incorporates speech patterns along with other sensor modalities collected from smartphone devices to predict mood using a subject-specific classifier. Additionally, the work by Vanello's group [87, 206] found that jitter, pitch, and pitch contours were effective indicators of mood from a subject-specific perspective. However, in order to detect mood effectively on a large scale it is necessary to both understand population level indicators in addition to subject-specific variations.

The PRIORI database is a longitudinal collection of cellphone speech data from individuals with bipolar disorder (Chapter III). It contains a considerable amount of data for each participant. This enables us to capture subject-specific as well as population-general aspects of speech. In this chapter, I use an SVM trained with rhythm to characterize population-general speech, as in Chapter V. I propose the use of *i-vectors* [52] extracted over *Mel-Frequency Cepstral Coeffiients (MFCC)* to capture mood variation at the subject-level using a *speaker-dependent SVM*. I leverage the large unlabeled subset of the data to circumvent the sparsity problems that often accompany i-vector extraction. Further, I employ the *Within-Class Covariance Normalization (WCCN)* technique [95] on the total variability subspace to alleviate the undesirable effect of mobile phone channels.

The main novelty of my approach is the fusion of a subject-specific system using unlabeled personal calls with a population-general system for the detection of de-

Mood	Total	# Per Subject	% Per Subject
Euthymic	306	7.1±6.5	36%
Depressed	266	6.2±6.8	27%
Excluded	361	8.4±7.0	37%

Table 6.1: *Distribution of mood in the assessments. Shown are the total number of observations, the mean and standard deviation of subject observations, and the mean percentage of each.*

pression from Bipolar speech. I compared this fusion with the baseline system from Chapter V, which modeled rhythm features in a population-general manner. My results showed significant improvement from the baseline with the Unweighted Average Recall (UAR) increasing from 0.66±0.11 to 0.73±0.09 and the Area Under the receiver operating Curve (AUC) increasing from 0.69±0.15 to 0.78±0.12. This shows the importance of using both cohort and subject-specific knowledge when modeling mood.

6.2 Data

This chapter uses a later snapshot of the PRIORI dataset than the prior chapter, and contains data from 43 participants with an average collection duration of 21.2±14.2 weeks per subject, including 39,445 calls and over 2,880 hours of speech. All experiments, including the baseline, are performed on the dataset snapshot at this time. This chapter focuses on depression and as such, all manic and mixed speech is excluded. See Table 6.1 for the data distribution.

6.3 Features

6.3.1 Rhythm Features

The rhythm features are calculated for each subsegment using an algorithm by Tilsen and Arvaniti [201]. The audio envelope is extracted and the spectral power ratio and centroid are found. The first two Intrinsic Mode Functions (IMF) are then

extracted using empirical mode decomposition [103]. The power ratio between the IMFs, as well as the mean and standard deviation of their instantaneous frequencies are found. This forms a total of seven segment-level statistics that have shown to be related to syllable- and word-level rhythm [201]. A total of 31 statistics are used to form the call level feature vector of 217 dimensions. These include mean, standard deviation, skewness, kurtosis, minimum, maximum, and range. Additionally, I calculated various percentiles and percentile ranges, the percentage of the call above thresholds of the range, and linear regression coefficients and error. Please refer to Chapter V for more information.

6.3.2 i-vectors

Subject-specific mood variation is captured using i-vector representation. Recent works have demonstrated the efficacy of this technique for predicting depression over a cohort [47, 133, 142, 189].

i-vector Formulation: Speech contains many sources of variability, including identity [52], age [25], gender [25], and critically, mood [47]. These variations can be captured using the i-vector technique. Its underlying assumption is that factors of variation lie in a low-dimensional subspace spanned by the columns of the total variability matrix T, a low-rank rectangular matrix. An arbitrary speech instance, u, can be represented by a *GMM mean supervector*, $M(u)$, which is modeled by:

$$M(u) = m + Tw(u) \tag{6.1}$$

where m is the supervector constructed from the *Universal Background Model (UBM)* trained using all *personal call* data, $w(u) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an utterance-dependent identity vector (i-vector) [52]. The total variability matrix is trained using an *Expectation Maximization (EM)* algorithm introduced in [112].

Acoustic Features: 19 MFCCs and log energy are extracted using a 25ms Ham-

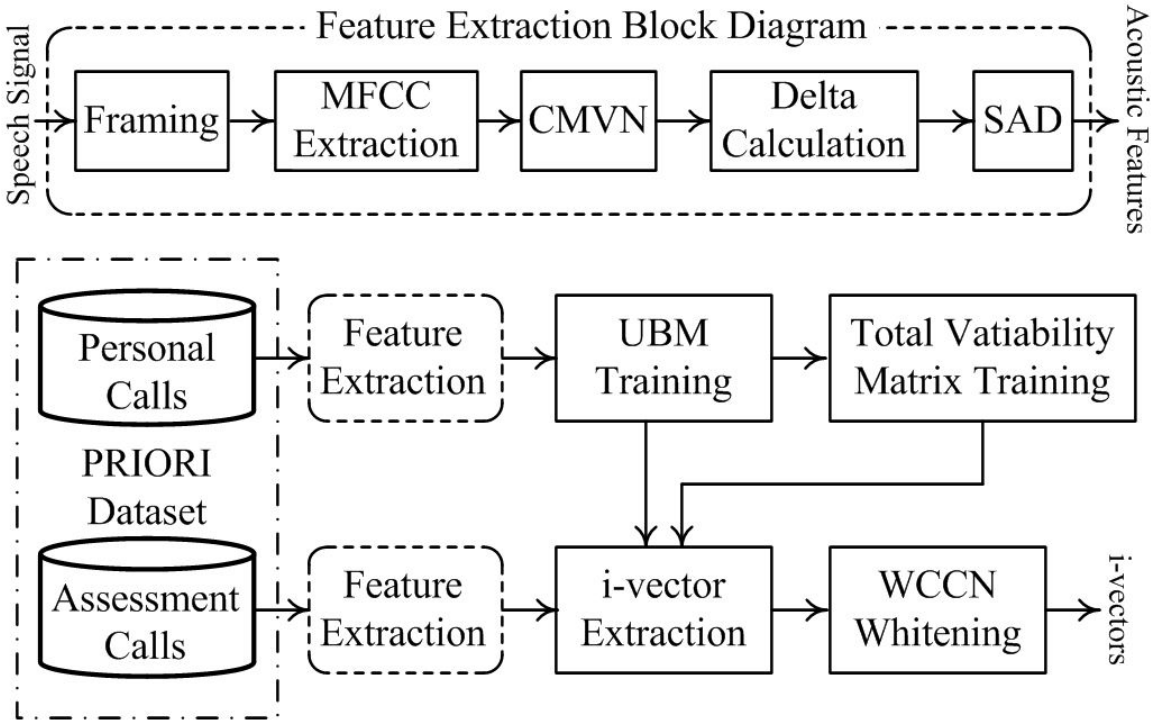


Figure 6.1: Schematic block diagram of i-vector extraction.

ming window with 10ms step from the calls with longer than five seconds of speech. They are normalized using utterance-level *Cepstral Mean and Variance Normalization (CMVN)* [208] to compensate for background and channel noise. This 20-dimensional feature vector is applied to a feature warping [163] with 3 second sliding window. The final feature set contains the MFCCs/log energy, their Δ, and $\Delta\Delta$.

i-vector Extraction: Figure 6.1 shows the system developed for extracting i-vectors. I train the UBM (2048 Gaussians) and total variability matrix (400 dimensions) using the acoustic features. Then I extract assessment i-vectors and apply WCCN [95] on them to compensate for residual channel effect.

6.3.3 Feature Normalization:

Both feature sets are normalized using the mean and standard deviation of each subject. Additionally, each fold is globally normalized so that a mean of zero and standard deviation of one is attained across all subjects.

6.4 Data Modeling

SVMs [45] are used to classify both types of features. SVMs find the boundary that maximally separates two classes. I use either a linear or Radial Basis Function (RBF) kernel. I weight the samples to accommodate for class imbalance. Finally, the output score is the signed distance to the hyperplane.

Various divisions between training, testing, and validation folds are used in this chapter and are defined below:

- **Population-General Validation:** One test subject is left out when training the model. This builds a system that is generalized to work on previously unseen individuals. I validate parameters by dividing the training subjects into four folds. I require that subjects have at least six calls, including at least two euthymic and two depressed, to ensure enough data for normalization and performance metric calculation. This *baseline system* was presented in Chapter V.

- **Subject-Specific Validation:** Only data from one subject is used. During training one test call is left out. This produces a system that can adapt to the features of an individual. Validation is performed over the training calls divided into ten folds. Only subjects with at least four euthymic and four depressed calls are used to ensure enough training data. This system is used for i-vectors because they are suited to subject-specific modeling, as explained in Section 6.6.

- **Hybrid Validation:** A hybrid approach of the above two systems. Calls across all subjects are used. I train the model by leaving one test call out. This system uses information from both the subject of the test call and the population. Validation is performed over training calls divided between 10 folds with calls from all subjects distributed across folds. Only subjects with at least six calls, including at least two euthymic and two depressed, are used.

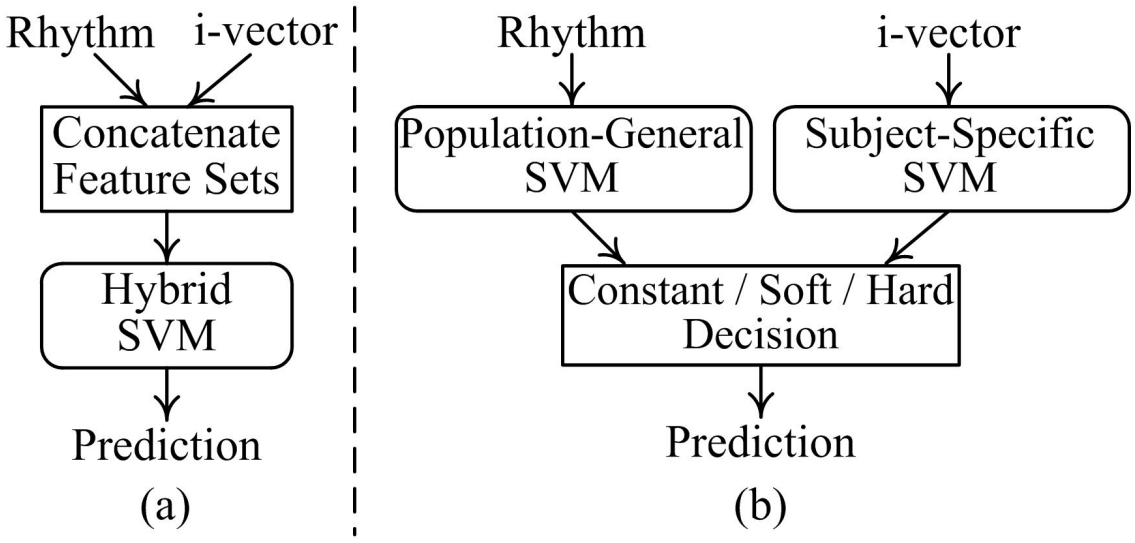


Figure 6.2: Diagrams of the system fusions. (a) Hybrid modeling with concatenated features. (b) Constant, soft, and hard decision fusions (all in one figure).

The kernel type and parameters (cost, gamma) as well as feature set size are selected during validation by maximizing UAR - the mean percentage of each class correctly identified. UAR gives minority classes equal weight. Features are ranked using a heuristic of Weighted Information Gain (WIG) to correct for subject label imbalances. This is implemented by weighted entropy as described in [192]. Each sample weight is set to the number of subject calls divided by the number of calls in the subject with the same label. This ensures that the sum of subject weights is proportional to its total number of samples, while also giving minority and majority labels equal weight. Only the test performances of subjects used in all systems are reported to make system results comparable.

To find the effect of combining cohort and person-specific knowledge, four fusion methods are considered (Figure 6.2):

Feature Fusion: Rhythm and i-vectors are concatenated into one feature vector. Hybrid validation is then performed.

Decision Fusion: I train a rhythm population-general model and an i-vector subject-specific model. SVM outputs from both models are normalized using a sig-

moid to ensure they are comparable. I determine the ideal weight (λ) to combine these systems for each test call using subject-specific validation. I find the population-general scores (PG) for the system trained using all but the test subject. I determine the subject-specific scores (SS) through validation by leaving one additional call out. This is necessary because test subject data is used in the model. Equation 6.2 shows the fusion scores (F):

$$F \;=\; PG \times \lambda \;\;+\;\; SS \times (1 - \lambda) \tag{6.2}$$

A higher weight indicates a higher contribution from the population-general system. Because score fusion is performed per call, each test call will have a different λ. During validation, I determine the weight based on the following three methods:

- **Constant:** λ is set to 0.5.
- **Soft:** λ is chosen between 0% and 100% by validating over increments of 1%. The best λ is selected by maximizing UAR measure. For subjects with fewer scores, there are often many weights that achieve the maximum UAR performance. A tie-breaking heuristic is used to determine which weight to choose in this case. The largest contiguous range of weights producing the maximum performance is found and the center weight is selected as λ. This mechanism was chosen to increase the stability of the fusion by selected a weight that is furthest from other weights that cause drops in performance.
- **Hard:** Same as soft, except λ is only allowed to be 0% or 100%. This results in only one classifier being selected.

System	UAR	AUC
Population-General (Rhythm)	.66±.11	.69±.15
Subject-Specific (i-vector)	.64±.17	.70±.18
Feature Fusion	.71±.14	**.76±.13***
Constant Decision Fusion	.72±.15	.74±.16
Soft Decision Fusion	**.73±.09***	**.78±.12***
Hard Decision Fusion	.71±.11	.76±.13

Table 6.2: Results for different systems (top) and fusions (bottom). Stared and bolded results mark significantly better performance than population-general baseline. (pairwise t-test, $p<0.05$).

6.5 Results

In addition to *UAR*, *AUC* is used in testing to compare the systems. *AUC* determines the system's ability to relatively rank test outputs. Unlike UAR, it does not have a set threshold. It is calculated as the area under the curve defined by the amount of true positives and false positives at all possible thresholds. Both measures have a chance performance of 0.5 and an ideal performance of 1. Table 6.2 shows a summary of results, including the two component systems and four fusions of systems.

Component Systems: The baseline system from Chapter V generalizes rhythmic symptoms across individuals. The subject-specific i-vector system learns individual patterns in the voice. Between the two, rhythm has the lower standard deviation of 0.11 UAR. Additionally, as seen in Table 6.3, no subjects perform worse than chance when using the rhythm model, showing greater stability. I hypothesize that this stability is due to the relatively larger number of samples across subjects used to train the population-general model. This means that the model remains mostly consistent between test subjects. Contrast this with the subject-specific i-vector system where the entire model changes between subjects. This results in a standard deviation of 0.17 UAR (Table 6.2) and four subjects performing below chance (Table 6.3). However, the i-vector model performs better than the rhythm model for six subjects. This difference in performance can be further quantified by the low correlation of

Rhythm	i-vector	Fusion	Mean λ	#Eut.	#Dep.	#Per.
.544	.730	.730	.10	9	14	474
.549	.761	.752	.18	19	14	513
.531	.813	.698	.22	6	16	2832
.707	.829	.829	.30	7	10	327
.570	.750	.740	.47	25	10	1660
.757	.714	.786	.56	5	7	780
.519	.385	.596	.57	13	4	348
.607	.400	.636	.62	7	20	769
.762	.774	.690	.63	7	6	131
.769	.625	.923	.64	26	4	1382
.714	.477	.618	.82	11	10	814
.739	.578	.683	.93	10	9	1483
.833	.417	.792	.93	6	12	558
.66±.11	.64±.17	.73±.09	.54±.27	12±7	10±5	929±741

Table 6.3: Subject AUCs of Soft decision fusion and component systems, ordered by mean weight (λ). The number of euthymic (Eut), depressed (Dep), and personal (Per) calls are shown. The last row is the column means and standard deviations. Highlighted rows show when soft decision performs best.

0.12 between SVM outputs of the two component systems. Prior work has shown that uncorrelated systems are more effectively fused [67, 131]. This indicates that fusing population-general rhythm and subject-specific i-vectors will likely produce better results than either of its components.

Fusion: Soft decision fusion attained the best performance of all experiments with a significant (paired t-test) improvement over baseline of 0.73±0.09 UAR (p=0.045) and 0.78±0.12 AUC (p=0.020). I hypothesize that soft decision fusion works best, because it allows for direct tuning of subject contributions from both features and validation methodologies. There are strong correlations of 0.70 and -0.67 between the respective rhythm and i-vector UARs and the mean selected weight. This demonstrates its effectiveness at selecting the best performing component system, unlike the constant decision version which is unable to moderate their contributions and has less consistent subject results (not significant, p=0.14). Additionally, there are four instances where the soft decision fusion performs better than both component

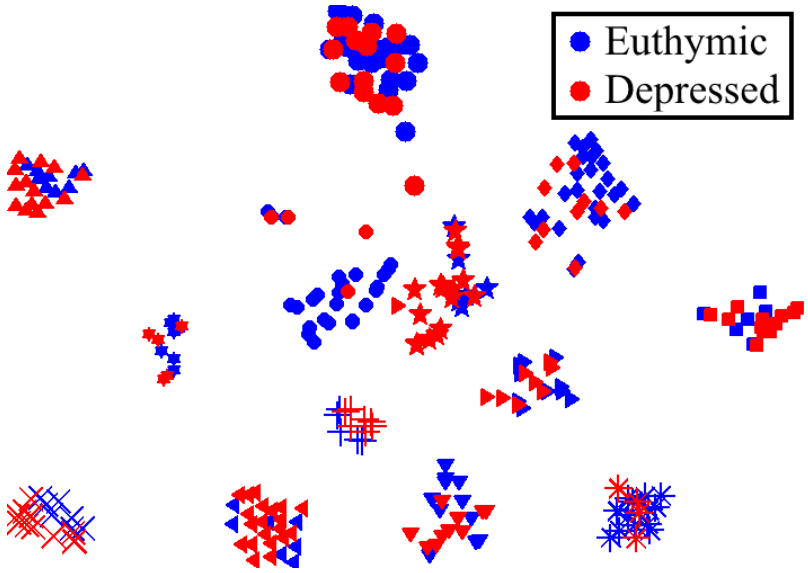


Figure 6.3: t-SNE plot of i-vectors showing subject separability. Shapes represent subjects, while the colors depict moods.

systems (highlighted in Table 6.3). This occurs when the weight is near 50% and there is contribution from both systems. This demonstrates its main advantage over the hard select technique - hard select can generally only perform as good as the best component system for each subject.

6.6 Discussion

The spatial distribution of the extracted i-vectors provides insight into the effectiveness of this feature at both the individual- and population-level. The i-vectors are mapped to a two-dimensional space using t-Distributed Stochastic Neighbour Embedding (t-SNE) [205], a dimensionality reduction algorithm that maps similar objects to nearby points and dissimilar objects to distant points. Figure 6.3 shows the distribution of the euthymic and depressed calls (blue and red dots in the figure, respectively). The individual groupings of the calls are at the subject-level. This separation at the subject-level suggests that the technique is effective for differentiating between mood within a speaker, but not between speakers due to the large speaker-effect. This

speaker-effect would preclude the use of a population-general i-vector classification. This effect can be mitigated using subject-specific normalization. However, this creates strong overlap between the two mood categories, suggesting that i-vectors may be most effective at the individual, rather than population, level.

6.7 Conclusion

This chapter demonstrates the importance of capturing both cohort and subject-specific variations in speech to effectively detect depression in bipolar disorder. This is important for a mental health monitoring system that both aims to be able to provide immediate help to new users and improved performance over time. This chapter introduces a soft decision fusion of population-general rhythm detection and subject-specific i-vector variation monitoring. The i-vectors are trained in a novel manner by using the unlabeled personal calls of individuals to learn patterns in the speech of subjects. This allows for a subject-specific system to be trained with relatively few assessment calls to effectively model individual changes in depression. The results show that the fusion significantly improves performance from the baseline of population-general rhythm.

The fusion experiments concentrated on learning the weights in a subject-specific manner because at least 4 samples of each label type were available. While this type of system worked well for depression, it may be difficult to adapt to mania. On average, subjects tend to be depressed three times more often than manic [141]. Due to this lack of data it may be necessary to consider similar subjects with more data as part of the fusion. However, this would require an investigation into subject similarities between speech phenotypes and mood. This is becoming increasingly important as I begin to model the personal calls, as population-general symptoms may be more difficult to find outside of assessments.

Part II

Addressing Variability in Emotion Recognition

CHAPTER VII

Emotion Recognition: Progressive Networks

7.1 Introduction

While Part I demonstrated methods to mitigate variability when detecting mood directly from speech, there is still much room for improvement. It is difficult to detect changes in mood solely relying on speech, due to the relatively large temporal disconnect between each. Because of this, the remainder of this dissertation is devoted to investigating and applying other high-level features for the automatic monitoring of mood. Emotion dysregulation is another common symptom of bipolar disorder [83]. As such, estimates of emotion could instead be used to augment predictions of mood made directly from speech. Part II of my dissertation explores improving the prediction of emotion from speech for eventual use in such a system.

Automatic emotion recognition has been actively explored by researchers for the past few decades. However, as seen in Chapter IV, the sizes of emotion datasets are relatively small when compared with other tasks such as automatic speech recognition (ASR) or speaker verification. This makes it difficult to create models that generalize beyond the recording conditions and subject demographics of a particular dataset. One possible approach to alleviate this problem is to incorporate related knowledge that can help in learning a better system. Many paralinguistic tasks are closely related and dependent on one another. As a result, I hypothesize that emo-

tion recognition models can be augmented with additional information from other paralinguistic information, such as speaker ID and gender, to improve classification performance. In this chapter, I explore transfer learning both as a method to leverage knowledge from other paralinguistic tasks and to augment an emotion model with another model trained on a different emotion dataset. In particular, I show that progressive neural networks are particularly well suited to achieve this transfer.

Previous work demonstrated that multi-task learning techniques (MTL) could be used to jointly model both gender and emotion [125, 207, 209], resulting in consistent performance increases compared to gender-agnostic models. Previous work also showed that emotion recognition systems could be improved by incorporating speaker identity as a feature, along with the other emotion-related features [191]. It has also been demonstrated that some features (e.g., pitch) are not only valuable for predicting emotions, but are also effective for detecting weight, height, gender, etc. [186]. Zhang et al. [222] explored MTL frameworks for leveraging data from different domains (speech and song) and gender in emotion recognition systems. Xia et al. [214] treated dimensional and categorical labels as two different tasks.

MTL is best suited to situations where it is possible to train with all data from scratch. However, in some cases an existing model needs to be adapted to a new situation. Transfer learning provides a framework from which to address this problem. The most common approach to transfer learning is to train a model in one domain and fine-tune it in a related domain [50, 55, 155]. This pre-training and fine-tuning (PT/FT) approach has been successful in cases where the available data for the source domain is abundant in comparison the data in the target domain. For example, in ASR, transfer learning is used to transfer knowledge from a richly-resourced language to an under-resourced language [50]. In emotion recognition, Deng et al. [55] presented a sparse auto-encoder method for transferring knowledge between six emotion recognition datasets. The authors used auto-encoders to transform a source domain

and its features to a domain that is more consistent with the target domain. The authors then used the transformed source domain to learn SVM classifiers of emotion in the target domain. Ng et al. [155] studied transfer learning for facial expression recognition. The authors first trained a model on a general large-scale dataset (ImageNet). They then fine-tuned the trained model on four datasets that contain relevant facial expressions similar to those in the target domain. Finally, the authors fine-tuned the network using the target data. The authors found that the two-step fine tuning approach provided improvements over one-step fine-tuning and no fine-tuning.

However, the PT/FT approach has a number of limitations. First, it is unclear how to initialize a model given learned weights from a sequence of related tasks [176]. Second, when a model is fine-tuned using initial weights learned from a source task, the end-model loses its ability to solve the source task, a phenomenon termed the "forgetting effect" [62]. Finally, transferring learned parameters between networks can be challenging if the networks have inconsistent architectures. In this work, I use the PT/FT approach as the baseline for transfer learning.

Another recent approach for transfer learning is progressive neural networks (ProgNets) [176, 177]. ProgNets train sequences of tasks by freezing the previously trained tasks and using their intermediate representations as inputs into the new network. This allows ProgNets to overcome the above-mentioned limitations associated with the traditional method of PT/FT, including the challenge of initializing a model from a sequence of models, at the expense of added parameters. Additionally, it prevents the forgetting effect present in the PT/FT methods by freezing and preserving the source task weights.

In this chapter, I investigate transfer learning between three paralinguistic tasks: emotion, speaker, and gender recognition, with a focus on emotion recognition as the target domain. In addition, I investigate the efficacy of transfer learning applied between two emotion datasets: IEMOCAP [31] and MSP-IMPROV [32]. Finally,

I study the effect of transfer learning between datasets when the target task has limited amount of data available. In all cases, I investigate three methods: (1) deep neural network (DNN); (2) DNN with PT/FT; and (3) progressive networks. The results demonstrate significant improvements over the conventional PT/FT methods when using ProgNets for transferring knowledge from speaker recognition to emotion recognition tasks. Furthermore, the results suggest that ProgNets show promise as a method for transferring knowledge from gender detection to emotion recognition tasks, as well as transferring knowledge across datasets, with results significantly better than a DNN without transfer learning.

7.2 Datasets

I use speech utterances from two datasets in this study: IEMOCAP [31] and MSP-IMPROV [32], further explained in Chapter IV. Both datasets were collected to simulate natural dyadic interactions between actors and have similar labeling schemes. I use utterances with majority agreement ground-truth labels. I only consider utterances with happy, sad, angry, and neutral labels.

IEMOCAP: I combine excitement and happiness utterances to form the happy category, as in [31]. The final dataset contains 5531 utterances (1103 angry, 1708 neutral, 1084 sad, 1636 happy).

MSP-IMPROV: The final dataset contains 7798 utterances (792 angry, 3477 neutral, 885 sad, 2644 happy).

7.3 Features

I use the eGeMAPS [61] feature set designed to standardize features used in affective computing. The eGeMAPS feature set contains a total of 88 features, including frequency, energy, spectral, cepstral, and dynamic information. The final feature

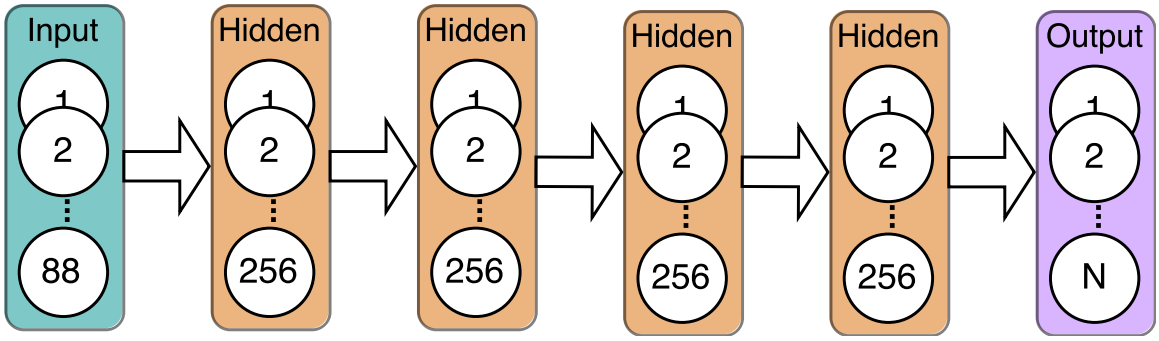


Figure 7.1: Deep Neural Network (DNN) used in the experiments. The arrows represent dense connections between each layer. The number of outputs (N) varies depending on the experiment.

vectors for each utterance are obtained by applying the following statistics: mean, coefficient of variation, 20-th, 50-th, and 80-th percentile, range of 20-th to 80-th percentile, mean and standard deviation of the slope of rising/falling signal parts, mean of the Alpha Ratio, the Hammarberg Index, and the spectral slopes from 0–500Hz and 500–1500Hz. I perform dataset-specific global z-normalization on all features.

7.4 Methods

I compare three methods in the context of transfer learning. As a baseline method, I consider the performance of a DNN trained on the target task without any extra knowledge. Additionally, I use the common transfer learning approach of pre-training a DNN on the source task and fine-tuning on the target task (PT/FT). The underlying assumption of PT/FT is that the target model can leverage prior knowledge present in the source task. This approach has been effective in many applications, including ASR [50] and natural language processing [148]. These two methods use the model depicted in Figure 7.1.

Both these methods are compared to the recently introduced progressive neural

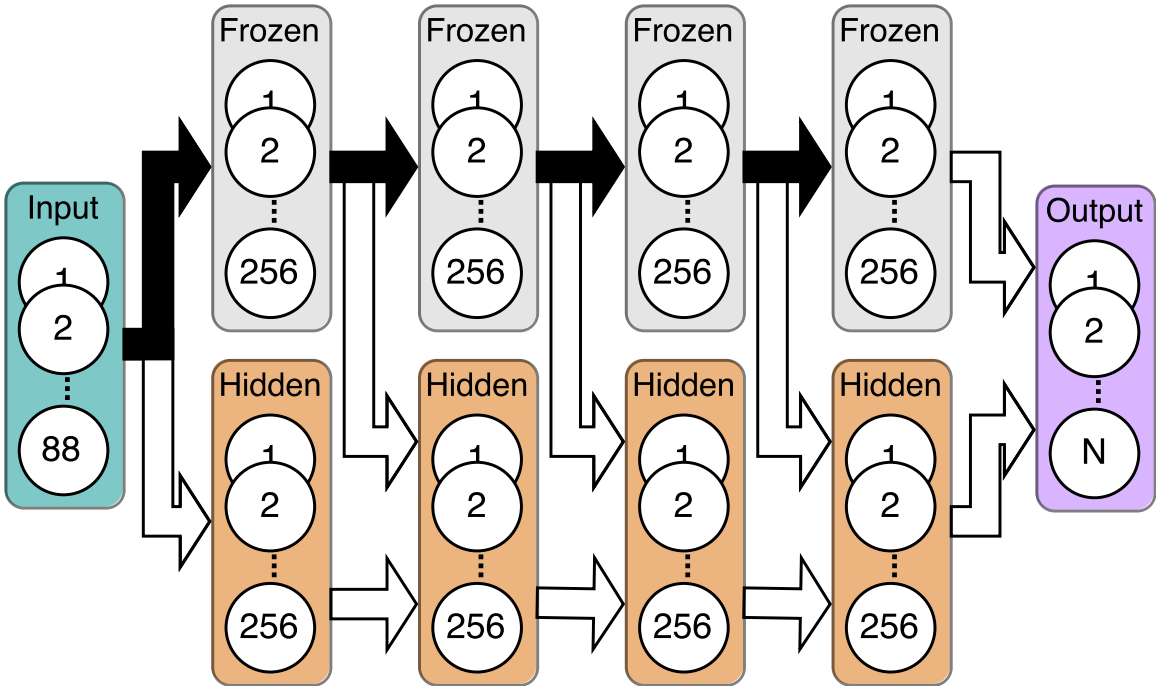


Figure 7.2: Progressive Neural Network (ProgNet) used in the experiments. The arrows represent dense connections between each layer. The black arrows show frozen weights from the transferred representations. The number of outputs (N) varies depending on the experiment.

networks (ProgNets) [176]. Instead of using learned parameters as a starting point for training a model on a target task, ProgNets do the following: (1) freeze all parameters of the old model; (2) add a new model that is initialized randomly; (3) add connections between the old (frozen) model and the new model; (4) learn parameters of the new model using backpropagation. ProgNets do not disrupt the learned information in existing source tasks, which avoids the forgetting effect present in PT/FT [176]. The ProgNet model is depicted in Figure 7.2.

In the construction of ProgNets, it is important to carefully select a method for combining representations across network and to identify where these representations will be combined. Adaptation layers can be included to transform from one task's representation to another. However, due to the small amount of data available for training, I use ProgNets without adaptation layers. For the same reason, I simplify the network by using an equal number of layers in each column and transfer the rep-

Table 7.1: The hyperparameters used in the experiments.

Hyper-parameter	Value
number of layers	4
layers width	256
hidden activation function	sigmoid
output activation function	softmax
dropout rate	0.5
learning rate	0.0005
maximum number of epochs	600

resentations between neighboring layers in a one-to-one fashion: the representations produced at layer k from the frozen column is fed as an input to layer $k+1$ of the new column.

Table 7.1 shows the neural network parameters used by all experiments in this chapter. These values were selected using a standard DNN without transfer learning to determine the best structure suited to the data (Figure 7.1).

I report unweighted average recall (UAR) as the comparison measure. UAR is an unweighted accuracy that gives the same weights to different classes and is a popular metric for emotion recognition, used to account for unbalanced datasets [183]. I evaluate the performance of the methods using a repeated ten-fold cross-validation scheme, as used in [27]. The folds are stratified based on speaker ID. In each step of cross-validation, one fold is used for testing, another is reserved for early stopping, and the remaining eight folds are used for training. I repeat this evaluation scheme ten times, resulting in ten UARs for each iteration. I calculate the mean and standard deviation UAR within folds and report the mean of these statistics over all iterations. I perform significance tests using a repeated cross-validation paired t-test with ten degrees of freedom, as shown in [27], and note significance when $p < 0.05$.

7.5 Paralinguistic Experiments

7.5.1 Experimental Setup

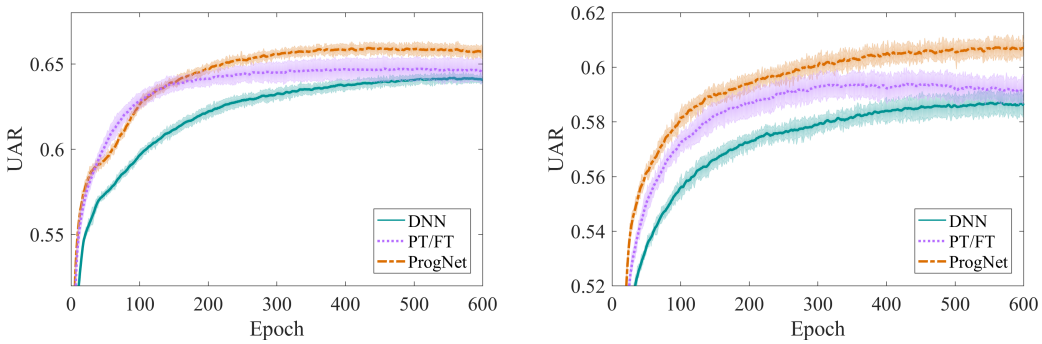
In the first set of experiments, I investigate the effectiveness of transferring knowledge from speaker or gender recognition to emotion recognition using the three methods mentioned above. In this section, I first report UAR of the systems on both IEMOCAP and MSP-IMPROV. I analyze the learning curves to compare the convergence behaviors of the systems. Prior work demonstrated that using the weights of a pre-trained model to initialize a new model to be trained on a related task can increase convergence speed [176].

7.5.2 Results

Table 7.2 summarizes the results obtained from speaker-emotion and gender-emotion transfer learning. When transferring from speaker recognition to emotion recognition, ProgNets significantly outperform both standard DNN (p = 2.6E-3) and PT/FT (p = 2.0E-2) for IEMOCAP and both standard DNN (p = 8.8E-4) and PT/FT (p = 1.2E-2) for MSP-IMPROV. The PT/FT system slightly outperforms the standard DNN, but the improvement is not significant. This suggests that ProgNets can efficiently incorporate representations learned by speaker recognition systems into

Table 7.2: Paralinguistic experimental results comparing different techniques for transferring knowledge from speaker/gender to emotion. Mean and standard deviation UARs are given for each method. A cross shows a result is significantly better than the other two methods for a given task, while an asterisk notes results significantly better than a standard DNN. The mean within-fold standard deviations are shown.

Source	Method	IEMOCAP	MSP-IMPROV
N/A	DNN	0.640±0.017	0.584±0.022
Speaker	PT/FT	0.645±0.017	0.592±0.018
Speaker	ProgNet	**0.657±0.018***†	**0.605±0.021***†
Gender	PT/FT	0.640±0.016	0.586±0.021
Gender	ProgNet	0.642±0.018	**0.593±0.022***



(a) IEMOCAP (b) MSP-IMPROV

Figure 7.3: The learning curves of different methods when transferring representations from speaker to emotion. The regions around each curve show the standard deviation of the UARs found by averaging across the folds of each iteration.

emotion recognition ones, but PT/FT cannot leverage this knowledge as effectively.

The performance of transferring gender information using ProgNets is not consistent between IEMOCAP and MSP-IMPROV. ProgNets significantly ($p = 1.2\text{E-}2$) outperform the standard DNN when transferring knowledge from gender recognition to emotion recognition in the case of MSP-IMPROV, but not IEMOCAP. I hypothesize that this is due to the stronger gender recognition performance on MSP-IMPROV. Gender recognition UAR on MSP-IMPROV and IEMOCAP are 98.1% and 93.1%, respectively. For both datasets, PT/FT is not effective at transferring gender information and performs no better than the standard DNN.

Figure 7.3 shows learning curves of the three reported systems for the case of transferring speaker knowledge to an emotion detection system. The figure shows that the PT/FT system reaches its best solution faster than the other two methods. The PT/FT system, however, achieves lower final performance than that of ProgNets. The learning curves of both transfer learning systems start with a larger slope compared to DNN. This slope vanishes quickly in PT/FT (after approximately 150 epochs), but the slope for the ProgNet preserves a positive value up to approximately 400 epochs. I hypothesize that this vanishing slope is due to PT/FT's inability to effectively incorporate representations learned for solving the source task.

7.6 Cross-Dataset Experiments

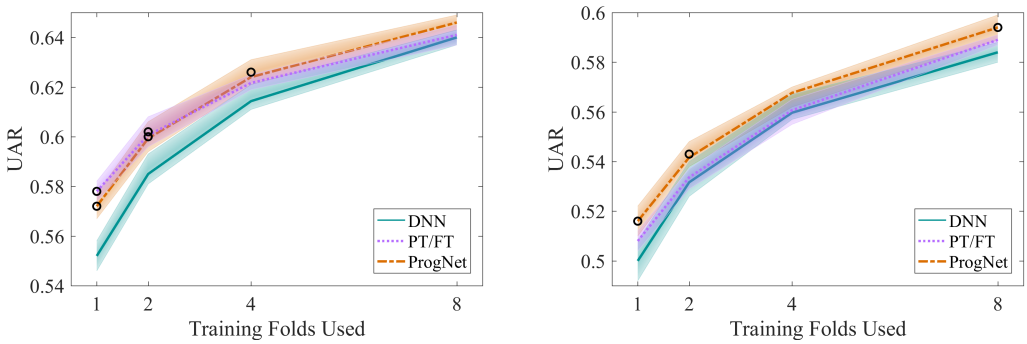
7.6.1 Experimental Setup

In this set of experiments, I explore transfer learning as a way to improve emotion recognition using an existing emotion model. In this experiment, the model is trained on the source dataset and is then adapted (PT/FT and ProgNet) to the target dataset. The standard DNN is trained only on the target dataset. I examine the impact of transfer learning when the target training data size is small by using different subsets of the training folds: 8, 4, 2, and 1. Previous work has shown that transferring knowledge from a large source data set to a smaller target datasets can be beneficial [50]. The source model is always trained using the full source dataset (all eight folds). I perform transfer learning by first treating IEMOCAP as the source and MSP-IMPROV as the target and I reverse the source/target designations.

7.6.2 Results

Figure 7.4 shows a summary of the results when transferring across different corpora. ProgNet significantly outperforms the standard DNN when transferring from MSP-IMPROV to IEMOCAP for training fold sizes of 1 ($p = 2.0\text{E-}3$), 2 ($p = 1.1\text{E-}2$), and 4 ($p = 1.6\text{E-}2$) and when transferring from IEMOCAP to MSP-IMPROV for training fold sizes of 1 ($p = 5.0\text{E-}3$), 2 ($p = 3.2\text{E-}2$), and 8 ($p = 3.6\text{E-}2$). PT/FT only achieves significant improvement versus the standard DNN baseline when transferring from MSP-IMPROV to IEMOCAP for training fold sizes of 1 ($p = 7.1\text{E-}4$) and 2 ($p = 1.0\text{E-}2$).

Because ProgNet has a larger number of weights to transfer knowledge, it is most beneficial when the target dataset is larger, compared with PT/FT. I hypothesize that this is what causes PT/FT to perform better in cases of smaller training fold amounts (1 and 2) on a smaller target dataset (IEMOCAP). This indicates that in



(a) Source: MSP-IMPROV; Target: IEMOCAP

(b) Source: IEMOCAP; Target: MSP-IMPROV

Figure 7.4: Cross-dataset experimental results under different amounts of training folds used (out of 8 total available training folds). Each test is run for ten iterations with different random folds to control for variations in selected data. All experiments use emotion as the source and target label. The regions around each curve show the standard deviation of the UARs found by averaging across the folds of each iteration. Circles mark results that are statistically significantly better than DNN.

some cases of small data, PT/FT may still be the better choice. However, in cases where the size of the target dataset is sufficient, ProgNet can effectively utilize the previous task representation better than PT/FT.

7.7 Conclusion

Transfer learning provides a method for using additional paralinguistic data, such as speaker ID, as well as a technique for combining models trained on different datasets. This chapter demonstrates the usefulness of progressive neural networks for this task. While pre-training a DNN on a source dataset has been previously used for transferring knowledge between tasks, progressive neural networks provide an alternative way of avoiding the forgetting effect by allowing the network to retain representations learned for solving the original task. ProgNets significantly outperformed the standard DNN and PT/FT networks when transferring knowledge between speaker identity and emotion. I also demonstrated that ProgNets can provide significant improvements for gender to emotion transfer tasks and dataset transfer tasks

when compared to systems that do not utilize source information. This suggests the importance of considering other factors including speaker, gender, and emotion for improved speech emotion recognition.

CHAPTER VIII

Emotion Recognition: Domain Generalization

8.1 Introduction

Speech emotion datasets are much smaller and less varied than their counterparts in many other machine learning tasks, including automatic speech recognition (ASR) [14]. As a result, even when an emotion model is successfully trained on one dataset, it often fails when applied to another [185]. This has motivated researchers to explore cross-corpus training methods to be able to utilize multiple datasets at once and to create systems more robust to unseen data. Chapter VII addressed this by using transfer learning to augment emotion with paralinguistic information and additional datasets. In this chapter, I take a different approach by introducing and exploring new methods for generalizing representations of speech for emotion by reducing the effect of non-emotion factors differing across datasets.

Emotion is only one of several factors that impacts the acoustics of speech. Some factors that change across datasets and can impact affect recognition include the environmental noise [54], the spoken language [185], the recording device quality [75], and the elicitation strategy (acted versus natural) [116]. Additionally, a mismatch in subject demographics between datasets can result in misclassification, due to the small numbers of participants common in speech emotion recognition datasets [14]. Early work in cross-corpus speech emotion recognition attempted to address these

differences with feature normalization [185, 224], sample selection [187], and decision fusion [188]. Most modern techniques of cross-corpus speech emotion recognition use deep learning to build representations over low-level acoustic features. Many of these techniques incorporate tasks in addition to emotion to be able to learn more robust representations [116, 161].

More recently, speech research has followed the popularization of adversarial methods, including Generative Adversarial Networks (GANs) [40, 80, 172, 179], Wassersteain GANS (WGANS) [16], and CycleGANs [110, 225, 226]. However, many of these generative speech transfer models introduce noise, as explored by Kaneko et al. [110]. To get around this issue, some cross-corpus research has instead explored discriminative adversarial methods, including Adversarial Discriminative Domain Adaptation (ADDA) [203] and Domain Adversarial Neural Networks (DANNs) [1, 5]. While ADDA has seen effective application in image recognition [203], it has not yet been successfully applied to speech emotion. This is likely because the target representation is learned independent of the output classifier and there is no guarantee that a lower varying characteristic, like emotion, would be preserved. DANNs work by using the GAN discriminator with the aim to "unlearn" domain from a target representation [5]. While this has been effectively applied to emotion, the authors note the difficulty in getting the method to converge in certain cases [1]. All of these methods are explained in greater detail in Section 8.2.

In this chapter, I investigate three models for speech emotion recognition across datasets. My baseline model is a Convolutional Neural Network (CNN), which is commonly employed in automatic speech emotion recognition [7, 106, 220]. CNNs are able to learn temporal filters across features and distill an entire utterance down into a static representation for more conventional fully connected layers to model. However, this model does not explicitly capture the effects of dataset or incorporate unlabelled data. I introduce Adversarial Discriminative Domain Generalization

(ADDoG) - a method for finding a more generalized intermediate representation for speech emotion across datasets. The network implementation is similar to CNN, except that an additional critic network is appended to the utterance-level representation layer. I adversarially train this network to iteratively move the different dataset representations closer to one another. I demonstrate that this "meet in the middle" approach always converges and improves upon previous, less stable, methods [1, 5]. Finally, I implement and explore Multiclass ADDoG, or MADDoG, which is able to incorporate many datasets at a time and build an even more robust and generalized representation.

I propose four sets of experiments to determine the effectiveness of the models for cross-corpus testing under different conditions. My first experiment examines the case of training on one dataset and testing on another, allowing ADDoG to also incorporate the unlabelled test features for training. This mirrors the transductive learning approach, seen in prior speech emotion work [194, 227]. I constrain the first experiment to only consider datasets recorded in a laboratory environment. Experiment 2 expands on this by introducing increasing amounts of labelled data available from the target dataset. Experiment 3 explores the impact of training on a laboratory dataset and testing on an in-the-wild dataset. I also look into incorporating three total datasets into training simultaneously, and present MADDoG as especially suited to this problem. Finally, Experiment 4 does the reverse of Experiment 3 and investigates training on in-the-wild data and testing on more traditionally recorded laboratory speech.

My results indicate that ADDoG consistently converges and is able to construct a more generalized representation for cross-corpus testing. This affirms the iterative "meet in the middle" approach to domain generalization. I find significant improvements in performance versus the baseline systems in all experiments with no added labelled target data. In addition to attaining higher performance, the ADDoG re-

sults have lower variance across repeated experiments, indicating better stability, when compared with CNN. Additional experiments show that ADDoG performs the best in the majority of cases when labelled target data is available, especially when the set is fairly small. However, the margin of improvement decreases with more added target data, implying that there is a trade-off between building a generalized model and specializing to the target domain. This trend holds true even with in-the-wild target data, demonstrating the robustness of the ADDoG technique. I find the improvement in performance to be at least as good as the benefit of doubling the amount of labelled target data. Finally, I show that MADDoG is able to improve further upon ADDoG when multiple source datasets are available by explicitly modelling all dataset differences.

The novelty of this chapter's work includes: (1) the ADDoG model, which allows for better generalized representation convergence by "meeting in the middle"; (2) the MADDoG method, which extends ADDoG to allow for many dataset differences to be explicitly modelled; (3) an analysis of cross-corpus experiments where both laboratory and in-the-wild data are trained and then tested on the other.

8.2 Related Works

Figure 8.1 gives a categorization of the main methods referenced in this chapter. See Section 2.5 for other prior methods used in emotion recognition.

8.2.1 Adversarial Methods

With the introduction of Generative Adversarial Networks (GANs) [80], there has been a large increase in the amount of adversarial methods for cross-corpus modelling. GANs work by iteratively training a generator and a discriminator. The generator aims to create data that matches a certain distribution of real examples from just a random seed. The discriminator is trained to be able to tell apart these generated and

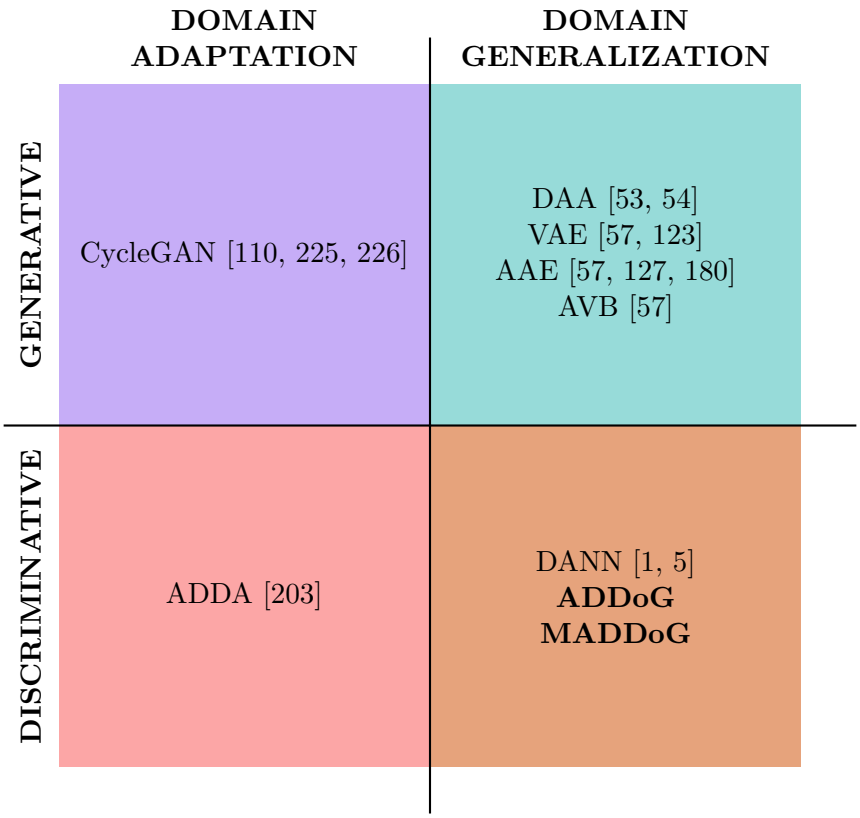


Figure 8.1: The main domain adaptation and domain generalization methods referenced in this chapter, divided by generative and discriminative methods. Prior work is listed with related citations and abbreviations defined in Section 8.2. Methods introduced in this chapter are bolded and are explained in Section 8.5.

real examples. A well-trained discriminator can be used to improve the authenticity of generated data by training a generator to fool the discriminator. To further expand on GANs, Radford et al. introduced Deep Convolutional GANs (DCGANs) [172]. They were able to improve convergence by using convolutions instead of fully connected layers. Automatic speech emotion recognition has begun to take advantage of these methods. Sahu et al. [179] explored how GANs could be used to augment utterance-level features, while Chang et al. [40] used DCGANs to improve performance on spectrograms.

The training of GANs can be unstable and relies on carefully tuned learning rates and numbers of generator versus discriminator iterations for convergence [172].

To address this issue, Arjovsky et al. introduced the Wasserstein GAN (WGAN) and was able to improve the convergence of GANs with a few minor changes [16]. The discriminator is replaced with a critic by removing the sigmoid activation on the output. Instead of trying to determine if examples are real or fake, like a discriminator, it instead learns to approximate the Wasserstein, or earth-mover's, distance between the real and fake distributions. This allows the system to have less sensitivity to over-training a discriminator, which often results in a saturated sigmoid function. It accomplishes this by clipping the critic weights to small values every iteration. This enforces the Lipschitz constraint and ensures that the output of the network does not grow infinitely. Instead, the network can only increase the output by finding the most succinct method of differentiating the real and fake examples. This allows a critic to be trained for many iterations before training the generator, resulting in more reliable gradients.

One of the first GAN-based methods to show promise for cross-domain applications was the CycleGAN by Zhu et al. [225]. This allowed for style transfer of images by converting from the style of one domain to another, while preserving the overall structure. A CycleGAN consists of two DCGANs working in tandem to convert from domain one to domain two and vice versa. Because CycleGANs can transfer in both directions, they are able to be trained with an additional reconstruction term that makes sure the overall structure of a transferred image is maintained. Zhu et al. demonstrated that is was possible to augment a facial emotion dataset by using CycleGANs to transfer between different emotions [226]. This allowed for training with balanced classes by transferring all utterances to each class, regardless of the original emotion. They found improved classification performance using this balacing method. Kaneko et al. explored subject conversion for speech using CycleGANs with some success, but found that there was still a large gap in quality for the real versus transferred samples [110]. I was unable to find published work on CycleGANs for

speech emotion recognition, possibly because of this lack of transfer quality.

To work around this transfer quality issue, some adversarial methods have instead forgone generative methods for discriminative ones. For example, Tzeng et al. explored the method Adversarial Discriminative Domain Adaptation (ADDA) for transferring to a target domain [203]. In the first stage of learning, the source domain data is encoded to an intermediate representation using a series of convolutions. The representation is then passed through a classifier and both networks are optimized based on the available labels. Next, a separate encoder is trained for the target dataset, using a discriminator to ensure that the source and target representations are similar. Finally, the target encoder is appended to the classifier that was trained on the source data and target predictions are output. This method produced significantly better and more balanced class performance on cross-corpus testing of numerical image datasets. However, this method makes the assumption that the representation trained by the second encoder will still preserve structure meaningful to the classifier.

8.2.2 Domain Generalization

Most of the prior cross-domain methods focus on transferring from one domain representation to another (domain adaptation). Domain generalization instead focuses on creating a middle-ground representation for all data [127]. Domain generalization methods can be divided into generative methods (usually autoencoder based) and discriminative methods.

One common method for finding a domain generalized representation is an autoencoder. Autoencoders work by converting the original features into a more compressed representation using smaller layer sizes, sparsity, or other regularization methods. Deng et al. examined the use of denoising autoencoders (DAAs) for cross-domain speech emotion recognition [53, 54]. DAAs introduce noise to the input features and

encourage the network to compress and reconstruct the features without the noise. This allows for the intermediate representation to discover a more noise robust representation that can work well across domains. Further work has examined different variations of autoencoders for speech emotion recognition, including Variational Autoencoders (VAE) [57, 123], Adversarial Autoencoders (AAE) [57, 127, 180], and Adversarial Variational Bayes (AVB) [57].

Another method that does not rely on autoencoders is Domain Adversarial Neural Networks (DANNs), introduced by Ajakan et al. [5]. DANNs have three main components: (1) feature extractor; (2) label classifier; (3) domain classifier. The input data is passed through the feature extraction layers. The representation is then fed to both the emotion classifier and the domain classifier. However, unlike multitask learning, where the domain is just another task, a reversal gradient layer is applied to the input of the domain classifier. This results in the network backpropagating a gradient to correctly classify the label but to incorrectly classify the domain, generalizing the intermediate representation. Abdelwahab et al. successfully applied this method to cross-corpus speech emotion recognition and showed significant improvement versus a model trained on the source dataset alone [1].

8.2.3 Transductive Learning

Another cross-domain framework that has been effectively applied within the speech emotion community is transductive learning. Traditional inductive learning first learns a model and then makes predictions on unseen data. Transductive learning instead aims to make predictions on a set of test data known in advance [157]. Because of this, it is possible to incorporate the unlabelled test data into the training procedure. Zong et al. explored an extension of linear discriminate analysis (LDA) for improving cross-corpus emotion recognition from speech [227]. Their method, called sparse transductive transfer LDA, or STTLDA, achieved significant improvement over

SVM. Further work by Song et al. demonstrated another extension of LDA, Transfer Supervised Linear Subspace Learning (TSLSL), which again provided improvement for speech emotion within the transductive framework [194].

8.2.4 Open Challenges

While many papers have explored cross-corpus speech emotion recognition, there are many challenges remaining for the field. Adversarial methods, led by CycleGANs [225], have shown promise for directly converting speech between datasets. Once all utterances are converted to one domain, the differences should no longer need to be considered during further modelling. However, there is much noise introduced in the output, making this currently impractical [110]. Other generative methods, such as autoencoders and their many variants, get around this by instead just using the intermediate representation for classification. Yet, it is unclear that compressing the representation preserves the emotion component of the signal, particularly given emotion's relatively slowly varying nature. In fact, prior work has shown that when the speech signal is compressed, the emotion content can be lost (e.g., Principal Component Analysis (PCA) [30, 117]).

Other discriminative methods have been introduced to avoid these issues, including ADDA [203] and DANN [1]. ADDA relies on training a feature transformation for the target dataset to match the mid-level representation for the source dataset. Yet, again, there is no guarantee that this transformation will preserve the emotion information present in the original example because the emotion classifier is trained separately. This is related to a well known issue with GANs, known as mode collapse, which results in the generator converging to just a few convincing examples, regardless of the input [16]. Additionally, emotion is less likely to be preserved when simply matching representation distributions, due to its relatively lower variability when compared with the entire speech signal (again, see the issues with emotion and

PCA [30, 117]). The DANN method improves on this by relying on a shared feature representation [5]. However, the researchers noted that DANN had issues converging on certain sets of parameters when training on speech emotion [1]. This could be due to the fact that DANNs attempt to "unlearn" domain, producing an unclear gradient. Finally, along with most other previously referenced papers in speech emotion recognition, the demonstration was on laboratory recorded datasets (IEMOCAP and MSP-Improv) rather than in-the-wild corpora. Further work is needed to incorporate more datasets simultaneously to improve generalization, as well as an exploration of the challenges of working with in-the-wild data.

8.3 Datasets

In this work, I use all emotion datasets explained in Chapter IV, including IEMOCAP [31], MSP-IMPROV [32], and PRIORI Emotion [115]. The PRIORI Emotion dataset is downsampled, selecting segments for which all assigned annotators were able to provide a rating, resulting in a dataset with 11,402 utterances over 21.7 hours. Table 8.1 gives a summary of the characteristics of each of the included datasets. Each of the datasets are segmented and rated on a dimensional scale for valence and activation by multiple annotators. I use these dimensional ratings for emotion, instead of discrete classes (as in Chapter VII), as I hypothesize they are more consistently interpretable across datasets [175]. Further, I focus only on valence in this study, as my preliminary experiments did not show a benefit to domain generalization for activation.

I follow the method similar to [40] and [8] to convert the dataset annotator ratings into a three bin vector for soft classification. The middle bin consists of valence ratings equal to 3 for IEMOCAP and MSP-Improv and 5 for PRIORI Emotion. The other two bins are valence ratings that fall below or above this midpoint. Each vector is formed by counting each of the ratings belonging to the bins. These counts are

Table 8.1: Summary of Emotion Datasets

	IEMOCAP	MSP-Improv	PRIORI Emotion
Subjects	10	12	12
Male	5	6	5
Female	5	6	7
Environment	Laboratory	Laboratory	Cell Phone Calls
Sample Rate	16 kHz	44.1 kHz	8 kHz
Valence Scale	1 - 5	1 - 5	1 - 9
Mean	2.79	3.02	4.86
Std.	0.99	1.06	1.12
Hours	12.4	9.6	21.7
Without Ties	8.6	8.9	16.4
Utterances	10039	8438	11402
Without Ties	6816	7852	8685
Low Valence	3181	2160	2809
Mid Valence	1641	2961	4779
High Valence	1994	2731	1097
Utterance Len.			
Mean (sec.)	4.5	4.1	6.8
Std. (sec.)	3.1	2.9	4.4

then divided by the total number of ratings so that the vector sums to one. For example, if three IEMOCAP annotators gave the ratings of 3, 4, and 5, the soft vector representation would be [0.0, 0.33, 0.66]. However, unlike [8], I do not include utterances with no clear majority bin to make the analysis more straightforward. This matches other speech emotion work that only used majority agreement [168]. Additionally, this corresponds with the finding by Schuller et al. that prototypical examples are more useful for cross-corpus speech emotion recognition [187].

The IEMOCAP dataset includes 3,181, 1,641, and 1,994 low, medium, and high valence utterances, respectively. The MSP-Improv dataset includes 2,160 low, 2,961 middle, and 2,731 high valence utterances. The PRIORI dataset includes 2,809 low, 4,779 middle, and 1,097 high valence utterances.

8.4 Features

I downsample the audio between datasets to match. For experiments involving just IEMOCAP and MSP-Improv the sample rate is 16 kHz. If PRIORI data are involved then all data are downsampled to 8 kHz. The audio is then normalized to 0dB FSD using the SoX command line tool [196]. I then extract 40 dimensional Mel Filter Banks (MFBs) using the Kaldi speech recognition toolkit [167]. The default options are used - a povey window with frame length of 25 ms, frame shift of 10 ms, preemphasis coefficient of 0.97, low cutoff of 20 Hz, and outputting log filterbanks. Because this produces features of different lengths for each utterance, batches have their MFBs padded by zeros to the length of the longest utterance.

8.5 Classification Models

In this section, I present the three different classification models used in this chapter: a simple Convolutional Neural Network (CNN), Adversarial Discriminative Domain Generalization (ADDoG), and Multiclass ADDoG (MADDoG), which is an extension of ADDoG that allows for more than one source dataset. All models consider MFBs as the input feature set and valence binned into a three dimensional vector as the output task. Each experiment will consist of labelled data from a source dataset (SRC) and data from a target dataset (TAR), some of which is labelled and some is not. TAR contains the test data and is available at train time without labels (transductive learning). The baseline CNN method is able to take advantage of the labelled data from all datasets, but does not use unlabelled data. Both ADDoG and MADDoG take advantage of the unlabelled test data to generalize the intermediate feature representation across datasets. In all methods, I use the Adam optimizer [118] with default parameters ($\alpha = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999$). All models described below are implemented in PyTorch version 0.4.0 [162].

8.5.1 CNN

Convolutional Neural Networks (CNN) have seen much success in speech emotion recognition [7, 106, 220]. Figure 8.2 shows my CNN implementation. It consists of two main components: (1) the feature encoder (convolutions + max pooling); (2) the emotion classifier (fully connected layers + softmax).

It is difficult to validate multiple sets of hyperparameters when conducting cross-dataset experiments, due to the lack of labelled data in the target domain. For this reason, I select hyperparameters based on those found to be commonly selected in prior work and keep them constant for all experiments. A channel size of 128 is used for all convolutional and fully connected layers, as commonly selected in prior work [113, 220]. ReLU is used as the activation function for all but the final layer, as it has been show successful in the field and is computationally efficient [66, 220]. I select a relatively large kernel size of 15 for the first convolutional layer, as previous work has shown large initial layers to be beneficial to emotion recognition using MFBs [7, 220]. I apply an additional convolutional layer of length 5 dilated by a factor of 2 to further extend the receptive field of the network. The global maximum is then taken over this convolution output, resulting in an encoded representation of 128 for the entire utterance. Previous work has shown that this is sufficient for recognizing emotion over short utterances [7, 220]. Dropout (p=0.2) is then applied to help prevent over-fitting. I next add three fully connected layers, with the final having three outputs for each of the valence bins, as in [8]. Finally, a softmax layer is applied, allowing for the output to be viewed as the probability distribution of valence. Biases are not used for any layers. Coupled with the ReLU activation and max pooling, this minimizes the effect of zero padding shorter utterances.

While older work in deep learning pretrained using autoencoders with unlabelled data, this has mostly subsided with the introduction of the ReLU activation, dropout, better initialization techniques, and larger datasets [78]. Because of this, the CNN

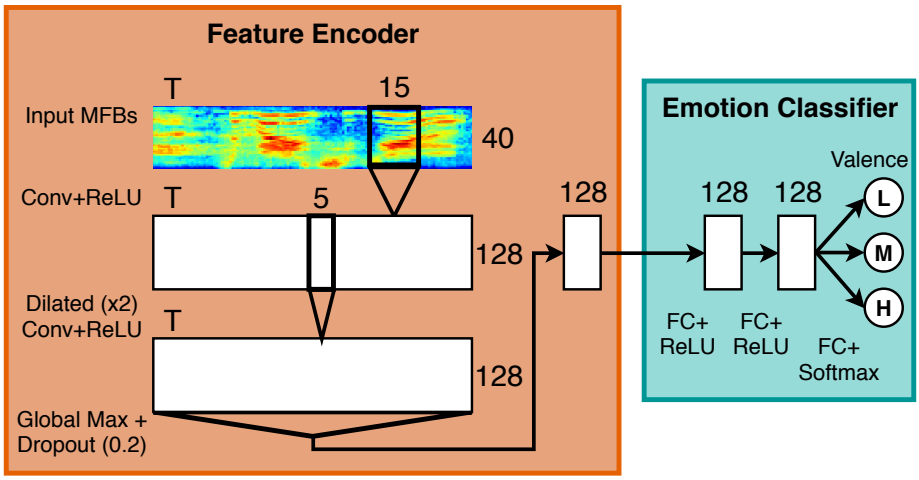


Figure 8.2: Convolutional Neural Network (CNN). Consists of two main parts: (1) feature encoder; (2) emotion classifier. The feature encoder uses a set of convolutions and global pooling to create a 128-dimensional utterance level representation. The emotion classifier then uses fully connected layers and a softmax layer to output the three bin valence probability distribution.

model does not use the unlabelled data, and only the labelled data from both SRC and TAR is used during training. Each epoch is divided into a total number of batches equal to the amount of labelled data divided by the batch size. After the MFBs are propagated through the network, I calculate loss using a weighted cross entropy measure. The classes are weighted so that all valence bins are given equal likelihood, regardless of class imbalance.

8.5.2 ADDoG

I introduce Adversarial Discriminative Domain Generalization (ADDoG), which addresses the open challenges of producing a generalized dataset representation using unlabelled target data, while still being able to consistently converge. Similar to CNN, it builds an intermediate 128-dimensional encoding of the utterances after global max pooling and dropout. However, in the case of ADDoG, there is a critic component, as in WGANs [16], that encourages the representations of the different datasets to be as close as possible. Unlike ADDA [203], the emotion classifier and database

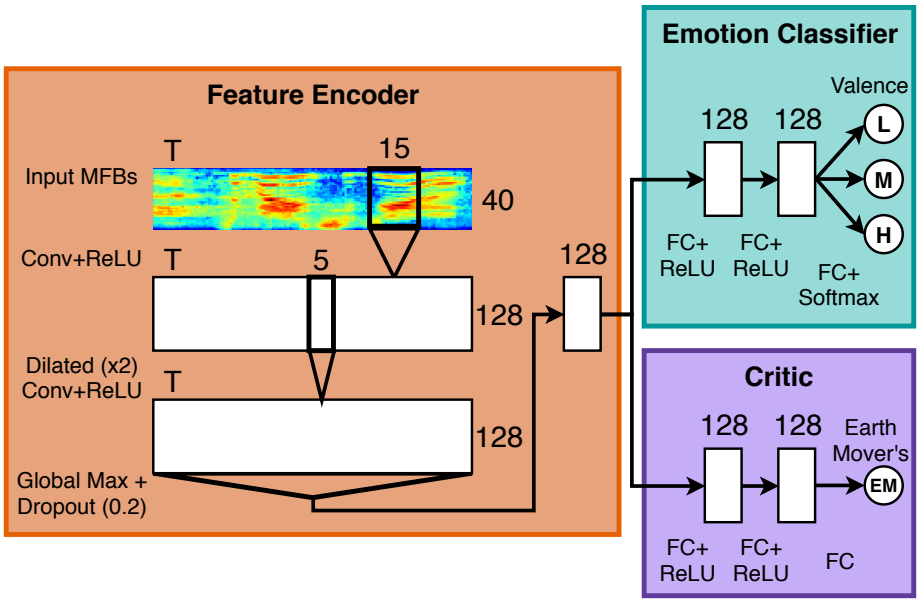


Figure 8.3: Adversarial Discriminative Domain Generalization (ADDoG) Network. Consists of three main parts: (1) the feature encoder; (2) emotion classifier; (3) critic. The critic learns to estimate the earth mover's or Wasserstein distance between the SRC and TAR dataset encoded feature representations. The emotion classifier ensures that valence is also preserved in the generalized representation.

critic are iteratively trained, ensuring the presence of emotion in the intermediate representation. I hypothesize that this creates a more generalized representation of emotion that will perform better across datasets, compared with CNN. This is because the representation will remove unrelated information that could mislead the emotion classifier (environment noise, microphone quality, subject demographics). Figure 8.3 shows the network structure of ADDoG. It consists of three main components: (1) the feature encoder (convolutions + max pooling); (2) the emotion classifier (fully connected layers + softmax); (3) the critic (fully connected layers + linear output).

The ADDoG hyperparameters are identical to those used in CNN. The critic network follows the same structure and hyperparameters as the emotion classifier. The only difference is that the critic is a linear activation instead of a softmax layer [16]. The training of the ADDoG network follows Algorithm 1 during each epoch. The number of training iterations per epoch is equal to the number of utterances in SRC divided by the batch size. Each iteration is divided into two main phases: (1)

training the critic; (2) training the feature encoder and emotion classifier.

Training the critic: I freeze the weights in the feature encoder and emotion classifier. First, unlabelled batches are sampled from SRC and TAR. Next, the MFBs are passed through the feature encoder to get the intermediate representations. These intermediate representations are then passed to the critic. I calculate the loss by subtracting the mean TAR output from the mean SRC output. I use the Adam optimizer on the critic weights with this loss to encourage TAR outputs to be as large as possible and SRC outputs to be as small as possible, estimating the Wasserstein, or earth mover's, distance [16]. The critic weights are then clipped to a range between -0.01 and 0.01, as in [16], to keep the outputs from growing infinitely. This critic training process is repeated five times to fully converge the critic before training the other systems, as in the original WGAN paper.

Training the feature encoder and emotion classifier: I freeze the critic weights. Next, I sample batches from SRC, TAR, and the subset of TAR that is labelled (if any). I then pass the MFBs through the entire network, getting outputs from the emotion classifier and the critic. As in the CNN training method, I calculate the emotion loss by weighted cross entropy using the SRC and labelled TAR sets. I add an additional term to the loss function for the critic that aims to move the dataset representations closer to one another. This is calculated by subtracting the mean SRC output from the mean TAR output, inverting the Wasserstein distance.

This training procedure iteratively moves the two dataset representations closer to one another, while following a clear gradient at each step. In contrast, DANNs [5] attempt to make a more generalized representation by "unlearning" domain. I attempted to implement DANN for my preliminary cross-corpus experiments, but had issues with getting the results to converge, as alluded to in [1]. My method of ADDoG gets around this by having a clear target at each step to "meet in the middle" and also utilizes the Wasserstein distance instead of a traditional discriminator.

Algorithm 1 Train ADDoG for one epoch. I use the default values of $n_{critic} = 5, c = 0.01, m = 32, \alpha = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999$

Require: The number of critic iterations per generator/classifier iteration n_{critic}, the critic clipping range c, the batch size m, Adam hyperparameters α, β_1, β_2.

Require: Generator parameters ϕ, critic parameters ψ, emotion classifier parameters θ.

Require: Emotion class weights for SRC S_w, emotion class weights for labelled TAR L_w

$n \leftarrow$ (Number of SRC samples) / m
for $batch = 1, ..., n$ **do**
  for $t = 1, ..., n_{critic}$ **do**
    Sample $\{S_X^{(i)}\}_{i=1}^m$ a batch from SRC data
    Sample $\{T_X^{(i)}\}_{i=1}^m$ a batch from TAR data
    $S_R \leftarrow G_\phi(S_X)$ ▷ Encoded SRC
    $T_R \leftarrow G_\phi(T_X)$ ▷ Encoded TAR
    $loss \leftarrow \frac{1}{m}\sum_{i=1}^m C_\psi(S_R^{(i)}) - \frac{1}{m}\sum_{i=1}^m C_\psi(T_R^{(i)})$
    $\psi \leftarrow$Adam$(\Delta_\psi[loss], \psi, \alpha, \beta_1, \beta_2)$
    $\psi \leftarrow \text{clip}(\psi, -c, c)$ ▷ Clip critic weights
  end for
  Sample $\{S_X^{(i)}, S_y^{(i)}\}_{i=1}^m$ a batch from SRC data
  Sample $\{T_X^{(i)}\}_{i=1}^m$ a batch from **all** TAR
  Sample $\{L_X^{(i)}, L_y^{(i)}\}_{i=1}^m$ a batch from **labelled** TAR
  $S_R \leftarrow G_\phi(S_X)$ ▷ Encoded SRC
  $T_R \leftarrow G_\phi(T_X)$ ▷ Encoded TAR
  $L_R \leftarrow G_\phi(L_X)$ ▷ Encoded labelled TAR
  $loss_C \leftarrow \frac{1}{m}\sum_{i=1}^m C_\psi(T_R^{(i)}) - \frac{1}{m}\sum_{i=1}^m C_\psi(S_R^{(i)})$
  $loss_E \leftarrow -\frac{1}{m}\sum_{i=1}^m S_y^{(i)} \times log(E_\theta(S_R^{(i)})) \times S_w$
    $-\frac{1}{m}\sum_{i=1}^m L_y^{(i)} \times log(E_\theta(L_R^{(i)})) \times L_w$
  $\phi \leftarrow$Adam$(\Delta_\phi[loss_C + loss_E], \phi, \alpha, \beta_1, \beta_2)$
  $\theta \leftarrow$Adam$(\Delta_\theta[loss_E], \theta, \alpha, \beta_1, \beta_2)$
end for

8.5.3 MADDoG

Multiclass Adversarial Discriminative Domain Generalization (MADDoG) expands the ADDoG algorithm to allow for more than two datasets. The MADDoG network structure is identical to ADDoG (Figure 8.3) except the critic has an output for each dataset instead of a single output. This allows for the method to account for the differences between all datasets while learning the representation. In contrast, ADDoG requires datasets to be grouped into target and source sets, not considering the differences within the sets.

The training of the MADDoG network follows Algorithm 2 during each epoch. The number of training iterations per epoch is equal to the number of utterances in SRC and TAR divided by the batch size. This is because data are drawn from all datasets simultaneously when training the critic instead of each separately. As in ADDoG, each iteration is divided into two main phases: (1) training the critic; (2) training the feature encoder and emotion classifier.

Training the critic: Training the critic is similar to ADDoG and begins with freezing the weights in the feature encoder and emotion classifier. One unlabelled batch is sampled from the combined SRC and TAR sets. The MFBs are then passed through the network to get the critic outputs, which are then modified as follows:

1. I calculate the proportion of each dataset versus the occurrence of all other datasets.

2. For each utterance in the batch, I flip the critic output corresponding to its dataset and multiply it by the previously calculated one-versus-all weight.

This makes each critic output a one-versus-all dataset critic and weights each of them so that samples from inside and outside the dataset are given equal total weight. The critic loss is calculated as the mean of the critic outputs, causing all of them to trend smaller. However, because the within dataset output is flipped, it is encouraged to

Algorithm 2 Train MADDoG for one epoch. I use the default values of $n_{critic} = 5, c = 0.01, m = 32, \alpha = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999, \lambda = 0.1$

Require: The number of critic iterations per generator/classifier iteration n_{critic}, the critic clipping range c, the batch size m, Adam hyperparameters α, β_1, β_2, the dataset generalization parameter λ.

Require: Generator parameters ϕ, critic parameters ψ, emotion classifier parameters θ.

Require: Emotion class weights for SRC S_w, emotion class weights for labelled TAR L_w, dataset one-versus-all weights DS_w

$n \leftarrow$ (Number of SRC and TAR samples) / m
for $batch = 1, ..., n$ **do**
  for $t = 1, ..., n_{critic}$ **do**
    Sample $\{X^{(i)}, ds^{(i)}\}_{i=1}^{m}$ a batch from all data
    $R \leftarrow G_\phi(X)$ ▷ Encoded data
    $D \leftarrow C_\psi(R)$ ▷ Get the 3 outputs of critic
    $D^{(:,ds)} \leftarrow D^{(:,ds)} \times -DS_w^{(ds)}$
    $loss \leftarrow \frac{1}{m}\frac{1}{3}\sum_{i=1}^{m}\sum_{j=1}^{3} D^{(i,j)}$
    $\psi \leftarrow$Adam$(\Delta_\psi[loss], \psi, \alpha, \beta_1, \beta_2)$
    $\psi \leftarrow \text{clip}(\psi, -c, c)$ ▷ Clip critic weights
  end for
  Sample $\{S_X^{(i)}, S_y^{(i)}, S_{ds}^{(i)}\}_{i=1}^{m}$ a batch from SRC data
  Sample $\{T_X^{(i)}, T_{ds}^{(i)}\}_{i=1}^{m}$ a batch from **all** TAR
  Sample $\{L_X^{(i)}, L_y^{(i)}\}_{i=1}^{m}$ a batch from **labelled** TAR
  $S_R \leftarrow G_\phi(S_X)$ ▷ Encoded SRC
  $T_R \leftarrow G_\phi(T_X)$ ▷ Encoded TAR
  $L_R \leftarrow G_\phi(L_X)$ ▷ Encoded labelled TAR
  $loss_C \leftarrow \frac{1}{m}\sum_{i=1}^{m} C_\psi(T_R^{(i)}) \times T_{ds}^{(i)}$
    $+ \frac{1}{m}\sum_{i=1}^{m} C_\psi(S_R^{(i)}) \times S_{ds}^{(i)}$
  $loss_E \leftarrow -\frac{1}{m}\sum_{i=1}^{m} S_y^{(i)} \times log(E_\theta(S_R^{(i)})) \times S_w$
    $-\frac{1}{m}\sum_{i=1}^{m} L_y^{(i)} \times log(E_\theta(L_R^{(i)})) \times L_w$
  $\phi \leftarrow$Adam$(\Delta_\phi[\lambda \times loss_C + loss_E], \phi, \alpha, \beta_1, \beta_2)$
  $\theta \leftarrow$Adam$(\Delta_\theta[loss_E], \theta, \alpha, \beta_1, \beta_2)$
end for

be larger. The critic loss estimates the one-versus-all Wasserstein distance for each dataset. As in ADDoG, the critic weights are clipped between -0.01 and 0.01 and the whole process is repeated five times.

Training the feature encoder and emotion classifier: The critic weights are first frozen. I then sample batches from SRC, TAR, and the subset of TAR that is labelled (if any). The MFBs are passed through the network, providing the emotion classifier and critic outputs. I calculate the emotion loss as before, using weighted cross entropy over the SRC and labelled TAR sets. For each utterance, the contribution to the critic loss is the critic output from the same dataset as the utterance (ignoring the other outputs). This encourages the one-versus-all Wasserstein distance to be reduced and the dataset to start looking like the others. Because this results in a more complex learning procedure than before, in practice I need to provide a weighting parameter λ=0.1 (found in preliminary experiments) to allow for the representation to converge. The total loss is the emotion loss added to the critic loss times λ.

The novelty of MADDoG is its ability to incorporate multiple datasets into its generalization procedure while still maintaining a clearly defined target (the other datasets) at each step. This creates cross-dataset representations that become more similar as the system is trained and that continue to encode the emotion information in the signal. As long as a sufficiently small learning rate is used, the intermediate dataset representations should converge somewhere in the middle. If instead, SRC datasets were considered as a group, the training procedure would not do anything to generalize between the SRC datasets. MADDoG considers these differences to enforce a more generalized representation that allows for better cross-corpus performance.

8.6 Experimental Design

I design four sets of experiments to examine different types of cross-dataset emotion classification. Each experiment examines the effect of the inclusion or absence of labelled data in the target dataset. The final two experiments focus on incorporating both laboratory and in-the-wild datasets.

All experiments begin by dividing the data into three folds: train, validation, and test. I run each experiment for 30 epochs, recording validation performance and test set predictions at each step. In this chapter, I use Unweighted Average Recall (UAR) as the performance metric. This ensures that each valence class is given equal weight, despite possible imbalance, and results in a chance performance of 0.33 UAR. Once an experiment is complete, I record test predictions from the highest validation epoch to prevent overfitting and calculate the UAR for each test subject. Each experiment is repeated a total ten times (fifty times for Experiment 1, Section 8.6.1), resulting in a final performance matrix of size (Number of Repeats $\times$ Number of Subjects). Folds are kept consistent between different methods so that these performance matrices can be compared. These experiments were split between two machines with GPUs - one with four GeForce GTX 1080s and another with one GeForce GTX 1080 and two Titan X's.

8.6.1 Experiment 1: Cross-Dataset

I determine the effect of training and testing on different datasets when labelled data is unavailable in the target dataset. Additionally, I constrain this initial experiment to only consider data from similar environments - using IEMOCAP and MSP-Improv, which were both recorded in a laboratory. I form the train and validation sets by splitting the SRC data randomly on a 80:20 split, respectively. In this experiment, I compare the effectiveness of a CNN versus ADDoG, which uses the unlabelled test data in the training process to learn a more generalized represen-

tation for emotion. I run each experiment for 50 total repeats, so that I get enough per-epoch data to perform a convergence analysis.

8.6.2 Experiment 2: Increasing Target Labels

The next experiment augments the training data with varying amounts of labelled examples from TAR. This experiment continues focusing on the laboratory recorded datasets IEMOCAP and MSP-improv. I train the network with 0, 200, 400, 800, 1,600, and 3,200 labelled TAR utterances, in addition to the labelled SRC utterances and unlabelled test utterances.

I follow the fold scheme shown in Figure 8.4 to get test predictions for all utterances in TAR. The number of SRC utterances and test utterances is kept constant through all experiments. SRC utterances are only used in the train set, unless no labelled TAR data is available. In that case, the SRC data follows a random 80:20 split between train and validation, as in Experiment 1. I split the TAR data randomly in half to allow for some labelled data for training, while reserving the other half for testing. If labelled data is used for an experiment, these samples are drawn from one of the halves and split 80:20 between the train and validation sets. Figure 8.4 depicts the case of having 200 labelled TAR utterances, resulting in a validation set of 40 labelled TAR and a train set including 160 labelled TAR and all of SRC. The remaining TAR data in the fold is discarded so the amount of unlabelled data is kept constant. This procedure results in test predictions for half of TAR. The TAR folds are then swapped, a new model is trained, and test predictions on the other half are output. Finally, I calculate the UARs for each subject in TAR using the concatenated predictions.

ADDoG is able to use the unlabelled test data along with the labelled SRC and TAR data during training for learning a more generalized dataset representation. The baseline CNN method is provided labelled data from both SRC and TAR when

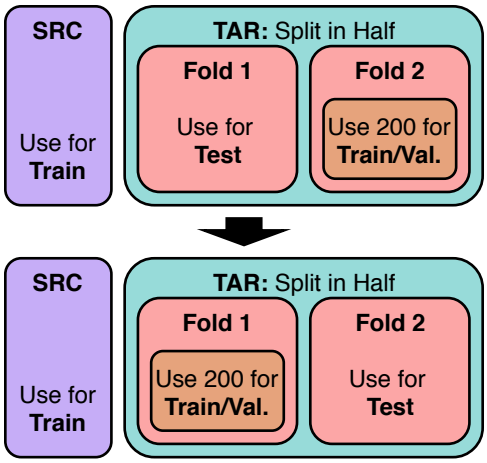


Figure 8.4: Folds used for Experiments 2, 3, and 4 when 200 labelled TAR are available. The SRC set is always used as part of the train set. The TAR set is split in half - part for testing and part for randomly sampling the 200 labelled TAR. The TAR data not selected in Fold 2 is discarded. After getting test predictions, the TAR folds are swapped and the process is repeated.

available. I also introduce another baseline method that specializes (SP) on the available labelled target data. SP uses the same network and training procedure as CNN, but only uses labelled TAR. Because of this, it is unable to be trained when 0 labelled TAR utterances are provided. I run this an all other experiments using 10 total repeats.

8.6.3 Experiment 3: To In-the-Wild Data

I next examine the effect of training on a laboratory recorded dataset (IEMOCAP and/or MSP-Improv) and testing on emotion in-the-wild (PRIORI Emotion). I expect this experiment to be more difficult than the previous two, due to the difference in recording environment (combining lab and cellphone call), recording quality (previously 16 kHz, now 8 kHz), and elicitation strategy (combining acted and natural conversation). I examine the effect of training on IEMOCAP or MSP-Improv alone, as well as training on them together. Each test follows the same procedure as Experiment 2, using the folds seen in Figure 8.4. I again compare the CNN, SP, and ADDoG methods. The experiment combining IEMOCAP and MSP-Improv training

data also employs the MADDoG method to take advantage of all three datasets.

8.6.4 Experiment 4: From In-the-Wild Data

My final experiment examines the reverse of Experiment 3 - training on in-the-wild data (PRIORI Emotion) and testing on laboratory recorded emotion (IEMOCAP or MSP-Improv). The experiment follows the same strategy as Experiment 2, using the fold scheme seen in Figure 8.4. I use the CNN, SP, and ADDoG models. Unlike Experiment 3, MADDoG is not used, as PRIORI Emotion is the only dataset used for training.

8.7 Results

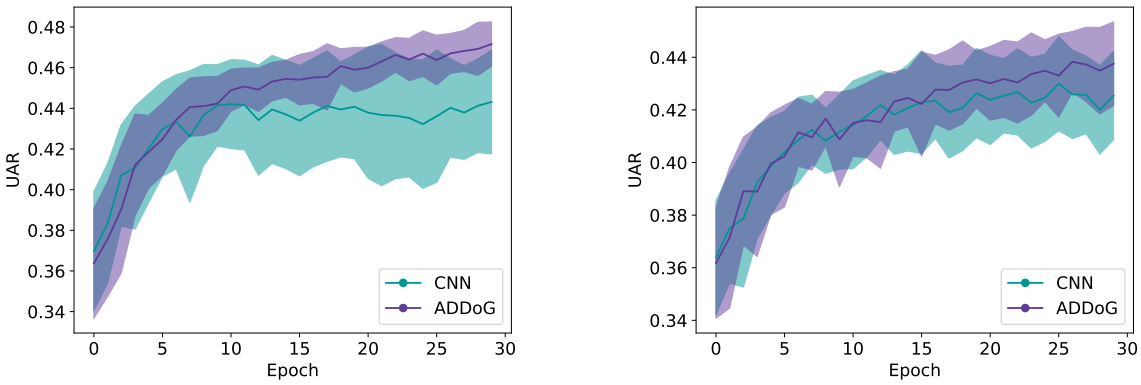
In all of the presented UARs, errors are calculated by first taking the mean subject UAR within each repeat of an experiment. The reported errors are the standard deviation of these means across all repeats, showing the stability of the findings. Matplotlib [107] was used to generate all result plots with the error shown as shaded error bands. Significance is determined using an analysis of variance in R [170] over the matrix of subject UARs output by the compared methods, as explained in Section 8.6. Significant results in each experiment are indicated by dots on the plots and/or bolded and starred values in the tables.

8.7.1 Experiment 1: Cross-Dataset

Table 8.2 shows the results of using IEMOCAP as SRC and MSP-Improv as TAR, as well as the reverse experiment. Two different methods are compared, including the

Table 8.2: Experiment 1 Results

	MSP-Improv to IEMOCAP	IEMOCAP to MSP-Improv
CNN	0.439±0.022	0.432±0.012
ADDoG	**0.474±0.009***	**0.444±0.007***



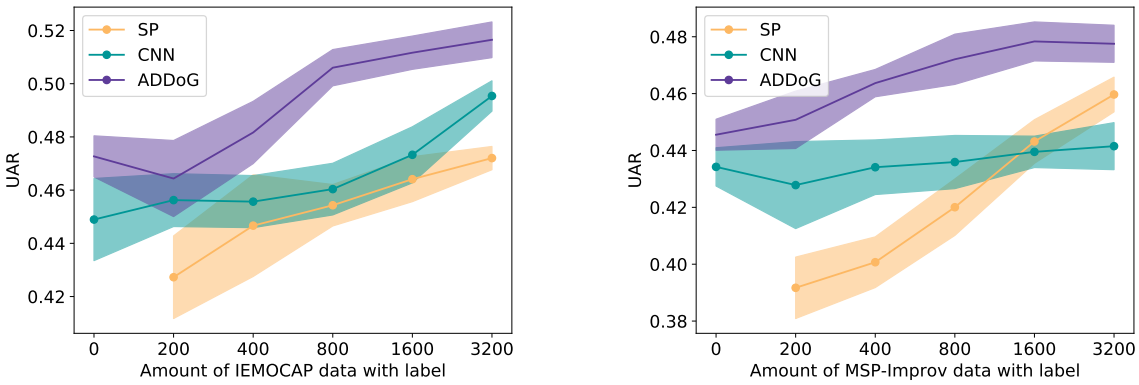
(a) MSP-Improv to IEMOCAP

(b) IEMOCAP to MSP-Improv

Figure 8.5: The test set mean subject UAR at different epochs when training on one dataset and testing on another. In particular, Figure 8.5a demonstrates how ADDoG reduces the variance of the output, improving cross-corpus testing, regardless of the mismatched validation set.

CNN, which is trained only using the SRC data, and ADDoG, which is additionally trained using the unlabelled TAR data, creating a more generalized intermediate representation. I find that ADDoG significantly outperforms CNN in both cases, implying that a more generalized representation can be used to improve cross-corpus testing without added labelled data.

In addition, I note that the standard deviation across experiment repeats is much lower for ADDoG versus CNN. This can also be seen in the convergence of results, as seen in Figure 8.5. Figure 8.5a in particular shows ADDA with a much smaller error at each epoch, compared with CNN. This is especially important for cross-corpus testing where labelled data is not available in the target dataset for validation. Using SRC data for validation is necessary in these experiments, but can still be unreliable, due to the mismatch. This may be less of a problem for ADDoG because of the results stability, contributing to the overall better performance.



(a) MSP-Improv to IEMOCAP

(b) IEMOCAP to MSP-Improv

Figure 8.6: Results of training on either IEMOCAP or MSP-Improv and testing on the other with increasing amounts of labels from the target dataset. Dots indicate methods significantly different from ADDoG using an analysis of variance in R (p=0.05).

8.7.2 Experiment 2: Increasing Target Labels

Figure 8.6 shows the results for Experiment 2, when I begin to incorporate labelled TAR data into the training and validation methodology. The left most point on both plots is the case when only unlabelled TAR data is available. This is slightly different than Experiment 1, as only half the amount of unlabelled data is available due to the fold structure shown in Figure 8.4. I find that ADDoG significantly improves on the baseline method in all cases, although the margin of improvement decreases with larger amounts of labelled target data. This may indicate that generalizing the representation may have diminishing returns once there is sufficient labelled data in the target domain. However, coupling even a small amount of labelled data and ADDoG results in significant improvement over baseline methods.

Adding labelled data to the ADDoG method increases its performance in all but one case - training on MSP-Improv and testing on IEMOCAP with only 200 labelled IEMOCAP utterances. While still significantly better than CNN and SP with the same amount of labelled data, better performance is actually attained using ADDoG without labelled IEMOCAP data. This could be due to the relatively small validation

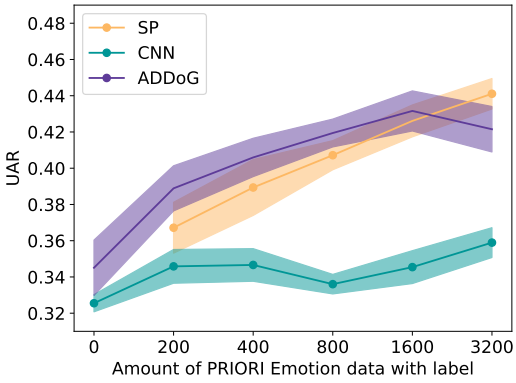
set, only consisting of 20% of the labelled data, or 40 utterances. This may not provide a reliable enough estimate of test performance, resulting in the larger error band around the result. It may be better to instead incorporate some additional SRC data in validation when very small amounts of TAR data are only available.

I also find that the SP method begins to outperform the CNN method when training on IEMOCAP and testing on MSP-Improv once a large amount of MSP-Improv labelled data is included. This may imply that appending SRC data to TAR data only complicates the training when not considering the effect of dataset. This is even more apparent when considering very different datasets, as seen in the next section.

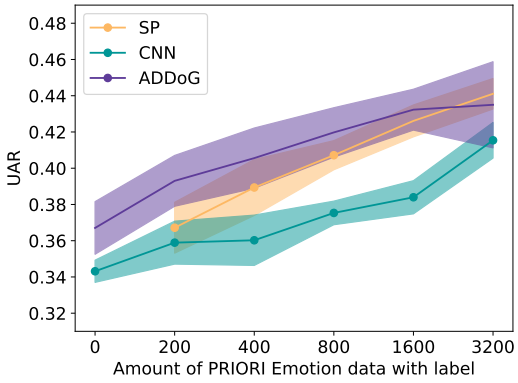
8.7.3 Experiment 3: To In-the-Wild Data

The results of Experiment 3 are shown Figure 8.7. Experiment 3 considers the effect of training on a laboratory recorded data (IEMOCAP and/or MSP-Improv) and testing on an in-the-wild set (PRIORI Emotion). Because all experiments use the same test set, the y-axis (UAR) range is kept constant. The SP results are the same between all figures, as it does not rely on the SRC data.

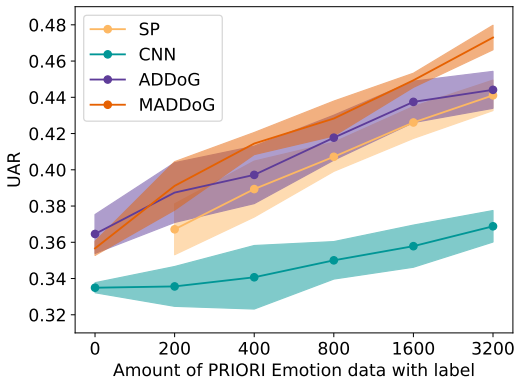
The first two figures 8.7a and 8.7b examine the case where just one laboratory dataset is used to train the model. I find much lower performance than prior experiments, due to the mismatch in recording conditions and elicitation strategy (acted versus a natural phone conversation). Combining SRC and TAR data together with the CNN method entirely fails, with the results consistently being the worst, due to the extreme mismatch. However, ADDoG is still able to provide a significant improvement in performance when none or a small amount (800 or fewer) of labelled TAR utterances are used. For these experiments with smaller labelled TAR data, the advantage of using ADDoG is similar to that attained by doubling the amount of labelled TAR data. However, this trend is broken with larger amount of labelled TAR



(a) IEMOCAP to PRIORI



(b) MSP-Improv to PRIORI



(c) IEMOCAP and MSP-Improv to PRIORI Emotion

Figure 8.7: Results of training on IEMOCAP and/or MSP-Improv and testing on PRIORI Emotion with increasing amounts of labels from PRIORI Emotion. Dots indicate methods significantly different from ADDoG in (a) and (b) and MADDoG in (c) using an analysis of variance in R (p=0.05).

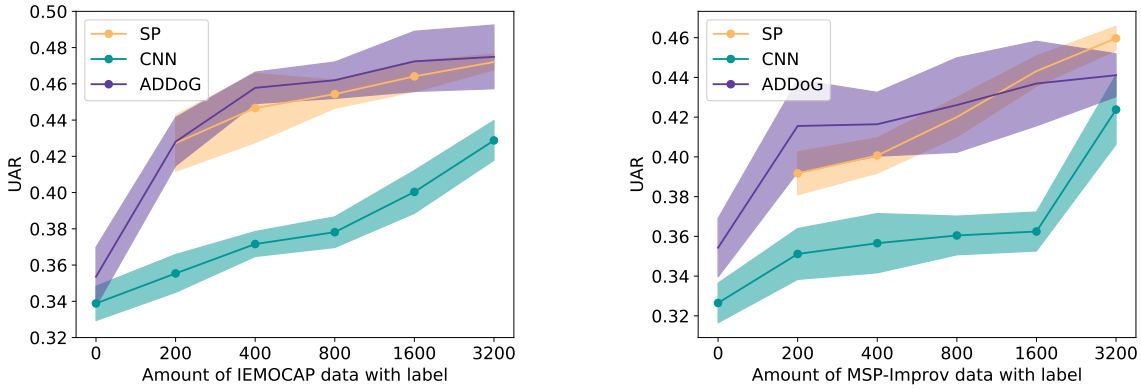
data where ADDoG no longer performs better and is in one case significantly worse (IEMOCAP to PRIORI, 3200 labelled samples). Due to the mismatch in dataset, it is better to specialize a model to the dataset characteristics, instead of generalizing, once a certain critical mass is attained. Both CNN and ADDoG perform slightly better when trained with MSP-Improv data, implying that it may be the more similar of the two datasets to PRIORI Emotion. This could potentially be due to the included more natural speech in the MSP-Improv dataset recorded in between scenarios.

Figure 8.7c shows the results for the last case where IEMOCAP and MSP-Improv are both simultaneously considered as SRC datasets. Despite the added data, the

CNN method is unable to perform better than with just MSP-Improv data, implying that the additional dataset is just confusing the classifier. The ADDoG classifier is able to take advantage of the additional data to at least perform the same as, if not better than, the MSP-Improv ADDoG method. While the method is not hurt by the addition of IEMOCAP, in most cases it does not help. However, MADDoG performs better than all methods using labelled data (significantly in all cases but ADDoG with 200 samples). This is likely due to the fact that it is able to effectively integrate together information from all datasets and come up with an even more generalized representation. ADDoG still seems to perform significantly better in the case where there is no labelled TAR data. Perhaps the labels from the other two datasets dominate the representation when none are available for MADDoG.

8.7.4 Experiment 4: From In-the-Wild Data

Because of my success in generalizing a representation across laboratory and in-the-wild datasets, I was interested in cross-corpus testing in the reverse direction. Figure 8.8 shows the results when training on PRIORI Emotion and testing on either IEMOCAP or MSP-Improv. In these experiments I just use the CNN, SP, and ADDoG methods, as there is only one SRC dataset included, making MADDoG unnecessary. My results again show that the CNN method performs the worst, demonstrating that appending together datasets does not work effectively when the datasets are too different. ADDoG also behaves similarly to Experiment 3 with significant improvements in most cases without labelled data or small amounts of labelled data. SP has similar or better performance than ADDoG once a substantial amount of labelled TAR data is available, implying that a method that trades off between generalization and specialization may instead be needed in these cases.



(a) MSP-Improv to IEMOCAP

(b) IEMOCAP to MSP-Improv

Figure 8.8: Results of training on PRIORI Emotion and testing on another dataset with increasing amounts of labels from the target dataset. Dots indicate methods significantly different from ADDoG using an analysis of variance in R (p=0.05).

8.8 Discussion and Conclusion

In this chapter, I investigate methods of controlling for the unwanted factors of variation when conducting cross-corpus experiments. These factors can include environmental noise, recording device differences, elicitation strategies (acted versus natural), and subject demographics. In cross-corpus speech emotion recognition, these factors can distract from the underlying emotion and decrease performance, especially because of the often smaller dataset sizes.

I introduce two new methods, ADDoG and MADDoG, which aim to generalize the representation of speech emotion across datasets. Both methods iteratively move their dataset representations closer to one another and have a clearly defined target at each step, following a "meet in the middle" approach. Experiments 1 and 2 focus on more traditional laboratory datasets to introduce the models and explore convergence. Experiments 3 and 4 take advantage of the PRIORI Emotion Dataset to examine the effect of training with in-the-wild data. Experiment 3 also explores training with three simultaneous datasets using the MADDoG method.

My results indicate that ADDoG is able to consistently converge and produce a more generalized representation across datasets, even when no labelled target data is

available. Significant improvement is found with no added labelled data in all four experiments, regardless of the number of datasets or whether they are laboratory or in-the-wild recordings. These results reinforce the idea that the "meeting in the middle" approach of ADDoG can reach the same generalized representation as "unlearning", seen with DANNs [1, 5]. However, convergence of the algorithm is easier to attain because of the more straightforward training paradigm. This generalized representation not only improves performance, but also decreases variance over different repeats of the experiment with different data. Because of this stability, less emphasis needs to be placed on validation. This is particularly important, since the validation and test sets are mismatched when conducting cross-corpus experiments.

Further experiments demonstrate how to effectively use small amounts of target labelled data when available. Simply combining the labelled data together from both datasets performs reasonably well when the recording conditions closely match, such as those in the two laboratory datasets - IEMOCAP and MSP-Improv. However, this method fails when substantially different data is introduced, such as PRIORI Emotion, demonstrated by the low CNN results in Experiments 3 and 4. ADDoG takes a more elegant approach to combining these datasets by building a generalized model and ensuring the representation is valid for the provided TAR data. Additionally, for the case of more than two datasets, MADDoG is able to recognize the differentiating factors in all SRC and take advantage of them. Next chapter will expand on MADDoG's ability to train a generalized emotion model for yet unseen domains and will focus on detecting emotion in individuals with suicidal ideation.

Part III

Applications

CHAPTER IX

Emotion Recognition in Individuals with Suicidal Ideation

9.1 Introduction

This chapter presents a real world application of emotion detection using MAD-DoG – the prediction of emotion dysregulation in individuals with suicidal ideation. Suicide is an increasingly serious public health issue, with the suicide rate increasing from 10.46 to 14.48 deaths per 100,000 between 1999 and 2017 [38]. A recent meta-analysis suggests that our ability to predict suicide is only slightly above chance levels, and has not improved over the past 50 years [69]. Early detection of suicidal ideation is crucial for prevention and intervention. However, relying on self-report of suicide risk is problematic, as the majority of patients deny suicidal ideation and intent in their last communication before their death by suicide [29, 108]. This points to the need for additional, objective monitoring strategies to better know when to intervene. Prior research has shown that individuals experiencing suicidal thoughts manifest changes in their speech [46]. This presents an opportunity for effective monitoring, as speech can be easily collected and relates to an individual's underlying condition.

Most prior work into automatically detecting suicidal or depressed speech has

focused on laboratory collected datasets [10, 46, 122, 182]. However, these datasets are not necessarily representative of the variations in environment and mood present under real world conditions. Furthermore, characteristics of speech change at the sub-second scale, while thoughts of suicide can be longer in duration [213] and vary considerably [212].

Suicidal ideation is also related to the manner in which emotion is expressed and prior work has examined this link [120, 124]. Self-reports of momentary suicidal ideation have been strongly associated with negative affect among psychiatric inpatients [17]. By first detecting emotional variations from speech, it may be possible to use emotion as a predictor of suicide. This still requires emotion detection of real world speech, which is a difficult task due to the confounding factors of environment, noise, and subject differences. Furthermore, due to the sensitive nature of suicidal data, there are no publicly available datasets linking naturally recorded speech, emotion, and suicide.

In this chapter, I examine the Ecological Measurement of Affect, Speech, and Suicide (EMASS) Dataset, introduced in Section 4.4. It contains recordings of natural phone conversations, as well as regular self-reports of emotion, mood, and suicidal thoughts using ecological momentary assessment (EMA) methods [26]. The participants include individuals with recent suicidal ideation or behavior, as well as psychiatric and clinical controls. The collection is ongoing, and the dataset is still relatively small. I demonstrate how outside data can be used to generate emotionally salient features. I then train a model to accurately predict a set of emotion measures, despite restrictive real-world conditions and small amounts of data. These measures were found to be indicative of suicidal ideation in prior work [17]. Finally, I show how emotion predicted from speech can be used to separate healthy controls from those with recent suicidal ideation. This system could eventually allow for the detection of the onset of suicidal ideation, making early intervention possible.

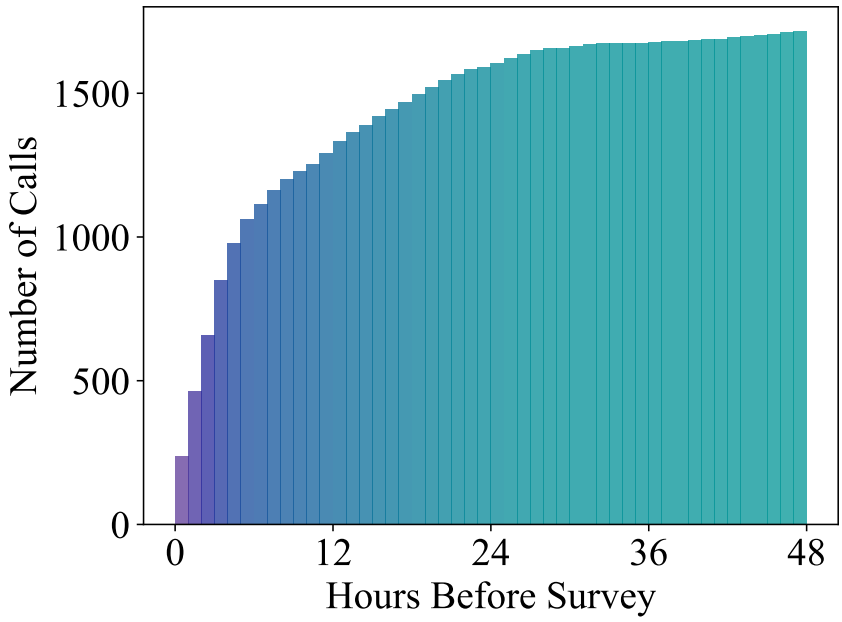


Figure 9.1: Cumulative histogram of the hours from calls to the following survey. There are a total of 239 calls with a survey within one hour afterwards (216 from subjects with at least five).

9.2 Data

I associate EMASS phone call recordings and surveys to enable automatic prediction (see Section 4.4). My initial experiments concluded that calls occurring after surveys were less related to the rated emotions than calls occurring before surveys. As such, call recordings are labelled with the emotion present in the closest following survey. Furthermore, I hypothesize the more time that separates a call and a survey, the weaker the certainty in the rated emotion. Because of this, I examine the impact of cutting off the training and testing data at different hours of separation. Survey separation is measured from the start of a call to the survey response. Figure 9.1 displays the number of calls present at different cutoffs ranging from one hour to two days.

I require at least five calls to be within the cutoff for a subject to be included in experiments (not necessarily from five unique surveys). Table 9.1 gives the amount

	All	HC	PC	SI	SA
Subjects	16	13	0	3	0
Calls	216	201	0	15	0
Hours	25	23	0	2	0

Table 9.1: The amounts of data for different groups of subjects, when considering only those calls occurring within one hour before a survey.

of data available when considering a one hour cutoff. While this severely reduces the data, it increases the certainty in the results. As such, I focus analysis on these 16 subjects and 216 calls from the HC and SI groups.

9.3 Features

Due to the relatively small dataset, I focus on three knowledge based, feature sets - eGeMAPS, Rhythm Statistics, and Emotion Statistics. eGeMAPS is a state-of-the-art emotion recognition feature set [61] and Rhythm Statistics have been used effectively in prior work in mood recognition [201]. I compare the efficacy of these features with Emotion Statistics - features generated by a deep learning system trained on existing emotion corpora. My hypothesis is that Emotion Statistics will outperform the other two, due to the small amount of data available for training emotion recognition in the EMASS corpus.

All call recordings are segmented using the ComboSAD algorithm, introduced in [178] and adapted for contiguous segments in [75]. The algorithm estimates the presence of speech using six signals - harmonicity, clarity, prediction gain, periodicity, perceptual spectral flux, and energy. These are then combined using principle component analysis (PCA) and then grouped into segments ranging from 2-30 seconds. See Section 5.3 for more details. All calls must at least contain at least three speech segments to ensure enough data for accurate feature extraction.

9.3.1 eGeMAPS

The eGeMAPS feature set was introduced in [61] and is extracted using OpenSMILE [59]. Low level descriptors (LLDs) are extracted for frequency, energy, amplitude, and spectral parameters in each segment and result in 23 values per frame.

9.3.2 Rhythm Statistics

Segments are subdivided into two second subsegments using a sliding window with a one second step size. Seven-dimensional representations of rhythm are then extracted for each subsegment, following the work by Tilsen and Arvaniti [201]. See Section 5.4 for further details.

9.3.3 Emotion Statistics

I extract segment-level emotion using the previously trained MADDoG model, introduced in [74] and described in Chapter VIII. This allows us to use outside data to generate a set of features indicative of emotion fluctuations. Furthermore, the training methodology results in a more generalized representation of emotion so that the model can be used across yet unseen datasets.

In this work, I incorporate emotional speech from three other datasets - IEMOCAP [31], MSP-Improv [33], and PRIORI Emotion [115]. I then train two separate MADDoG models for activation and valence using the three combined datasets. Segments from the EMASS dataset are then input to the models, resulting in three bins of emotion for both activation and valence, or six values per segment.

9.3.4 Call-Level Statistics

I apply 31 statistics across the concatenated segments to produce call-level features, as in [75]. These include the mean, standard deviation, skewness, kurtosis, minimum, maximum, and range of the signal. I perform linear regression on the

signal and use the fit parameters and error as statistics. I then extract the various percentiles and percentile differences and calculate the percentage of the signal above different thresholds.

9.4 Emotion Modeling

I compare all three feature sets using a DNN, trained to classify one of the 11 emotion measures in the EMASS dataset. The selected target emotion for each experiment is converted from a five point Likert Scale to a fuzzy binary scale for classification. The purpose of this scheme is to eventually allow for the system to distinguish between baseline and atypical emotion. The emotion baseline is estimated with the median subject rating, which produces a baseline of 1, 2, or 3 for each of 11 emotions. Each emotion scale is binarized with a fuzzy value of 0.5 between baseline and atypical values, as follows:

- Baseline of 1 or 3: $1 \rightarrow 0.0 \quad 2 \rightarrow 0.5 \quad (3,4,5) \rightarrow 1.0$
- Baseline of 2: $(1,2) \rightarrow 0.0 \quad 3 \rightarrow 0.5 \quad (4,5) \rightarrow 1.0$

Each experiment begins by randomly dividing the subjects into five sets for cross validation. One of the sets is reserved for testing, ensuring a subject-independent analysis. Each of the remaining subjects has their data randomly divided between training and validation, with 1/5 of their data used for validation. This process is repeated 100 times for each subject, resulting in 100 splits. I calculate the standard deviation of the target emotion within the two folds and use the split that maximizes their product. This ensures enough emotion variability in each fold. I found that applying Z-normalization was beneficial only for the eGeMAPS feature set and normalize it based on train data. All experiments are repeated a total of 100 times with different fold assignments to achieve more stable results.

I use a DNN for classification with four hidden layers (widths of 1024, 512, 256, and 256) using a RReLU activation function, found to work best in [215]. The output layer employs a sigmoid activation function and is trained with binary cross entropy loss. This loss is weighted by the inverse of the count of each emotion value (0.0, 0.5, 1.0) in the training set. The Adam optimizer [118] is used with a learning rate of 0.0001 and default parameters. This DNN was found to outperform random forest and support vector machines (SVMs) in early experiments and is the focus of this work. Training is performed over ten epochs with batches selected to contain all of one subject's data. This ensures the model focuses on learning within-subject variations versus cross-subject biases. I determine the stopping epoch by maximizing Pearson's correlation of the actual emotion and predictions across all data in the validation set.

The test predictions are then estimated using the held-out subjects and selected model. For each test subject, I calculate an Area Under the Receiver Operating Characteristic Curve (AUC) as the performance measure. AUC represents the ability of a system to correctly rank pairs of instances and has a chance rating of 0.5 and ideal rating of 1. Subjects must have at least one negative instance (0.0) and one positive instance (1.0) to be able to calculate a valid AUC. Instances with the fuzzy value of 0.5 are not used to calculate test AUC. Because of this, each emotion experiment will have a different set of test subjects that have enough data for AUC calculation (see Table 9.3).

9.5 Results

In this section, I explore speech emotion classification using different feature sets, survey cutoffs, and emotion measures.

I first examine the effect of feature set choice and focus on only testing with calls within one hour of a survey. I employ varying amounts of data in the training set and allow for a cutoff of 1, 2, 4, 8, 16, 24, 36, or 48 hours. Table 9.2 gives the

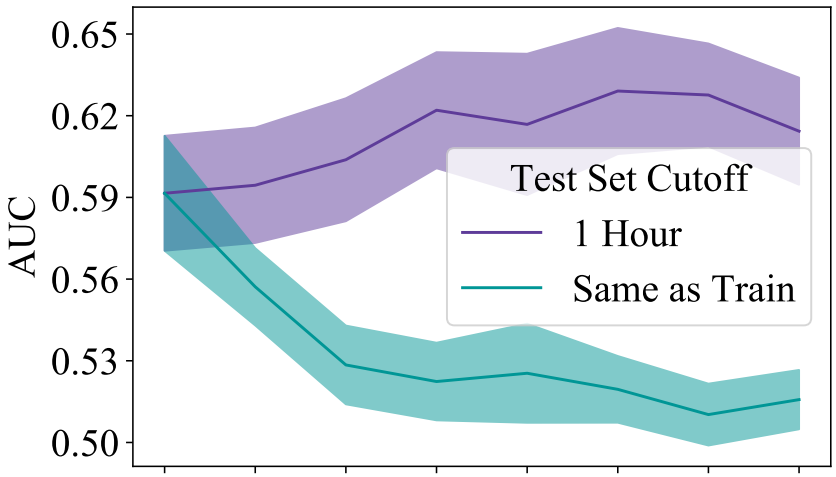
Features	Best Cutoff Hr.	AUC
eGeMAPS	8	0.53 ± 0.12
Rhythm Statistics	16	0.54 ± 0.09
Emotion Statistics	**24**	**0.63 ± 0.10**

Table 9.2: The results for each feature set on calls within one hour before surveys. AUCs are averaged across iterations, subjects, emotions. The best hour cutoff is found for each feature set. The AUC error is the standard deviation across subjects.

performance of each feature set averaged across all subjects, emotions, and iterations using its best performing training cutoff. I find that the Emotion Statistics perform substantially better than the others, which are close to chance. This is likely due to the Emotion Statistics already containing estimates of emotion at the segment level. This allows the model to overcome the lack of data, which makes classification difficult for other feature sets. Because of this, the following analyses only focus on the Emotion Statistics feature set.

Figure 9.2 shows the analysis of performance at different training set cutoffs, averaged over subjects, iterations, and emotion measures. While a larger cutoff allows for more training calls, it lowers the certainty in training labels. However, because subjects usually participate in three surveys per day, the median separation between calls and surveys is still only 4.03 hours even with a 48 hour cutoff. There are diminishing returns for the amount of added data with each increase in cutoff (Figure 9.1). When testing on the newly added data at each increase in cutoff, I find decreased performance. However, if I only test on the 216 calls within one hour of a survey, I see performance increase with a maximum at 24 hours. This provides the most data for classification without overly diluting the labels.

I lastly examine the model's capability to detect different types of emotion using the Emotion Statistics feature set, a 24 hour training cutoff, and the 216 test calls within one hour of a survey. Table 9.3 presents the AUC for each emotion, averaged over all iterations and subjects with enough data for testing. Due to the lack of data,



Cutoff Hr.	1	2	4	8	16	24	36	48
Median Hr.	0.43	0.96	1.78	2.63	3.27	3.71	3.93	4.03
Subjects	16	20	24	25	25	25	27	28
Calls	216	448	838	1151	1405	1574	1661	1703

Figure 9.2: Mean AUC over all emotions, subjects, and iterations using emotion statistic features at different cutoffs. The error bands show the standard deviation between iterations. The table displays the amount of data at each cutoff.

Emotion	Subjects	Calls	AUC
Confident	7	66	0.64 ± 0.19
Excited	7	92	0.51 ± 0.19
Happy	7	87	0.63 ± 0.21
Sad	6	65	0.54 ± 0.08
Guilty	4	59	0.66 ± 0.25
Worried	4	50	0.62 ± 0.23
Shame	3	45	0.57 ± 0.12
Hopeless	2	21	0.72 ± 0.04
Anger at Others	**8**	**89**	**0.78 ± 0.16**
Anger at Self	5	65	0.60 ± 0.23
Irritable	3	24	0.69 ± 0.34

Table 9.3: Results on emotion measures using emotion statistic features with a 24 hour cutoff. Only calls with a survey within one hour afterwards are used in testing. The amount of subjects and non-fuzzy calls available to calculate AUCs are shown.

it is difficult to draw conclusions about individual measures. One exception is "Anger at Others", which has the most subjects (8), highest AUC (0.78), and a relatively small standard deviation between subjects (0.16). In total, 8/11 emotions have an AUC of at least 0.6, demonstrating an overall trend in the model's ability to capture emotion in natural speech.

9.6 Suicidal Ideation Analysis

In this section, I explore the relationship between suicidal ideation and emotion estimated from speech. I focus on HC and SI subjects, as they are the groups with the most data (19 and 12 subjects, respectively). Emotion measures are extracted from the 2,988 HC and SI calls using the previously trained DNNs. Each estimate is only taken from models where the subject was unused during training. I exclude measures of Excited, Sad, and Shame, as they were previously predicted with less than 0.6 AUC. I calculate the within-subject standard deviation of each emotion to gauge each emotion's variability.

Figure 9.3 shows the overall variability of the two groups across the different emotions. I consistently find that subjects with SI have lower levels of emotional variability. I then use those emotions with significant differences (Guilty, Hopeless, Anger at Others, Anger at Self, Irritable) to classify HC versus SI. I average each of the five emotion standard deviations for subjects and use this as an estimate. Using this method, I attain a performance of 0.79 AUC.

These findings, though preliminary, are inconsistent with existing research on affective instability and suicide. One study found that heightened affective instability was associated with suicidal behaviors [217], while another found no link [130]. However, individuals in these samples were recruited based on a borderline personality disorder diagnosis and were not compared to healthy controls, unlike the present analysis. It is possible that self-reported affect differs from more objective measures (i.e.

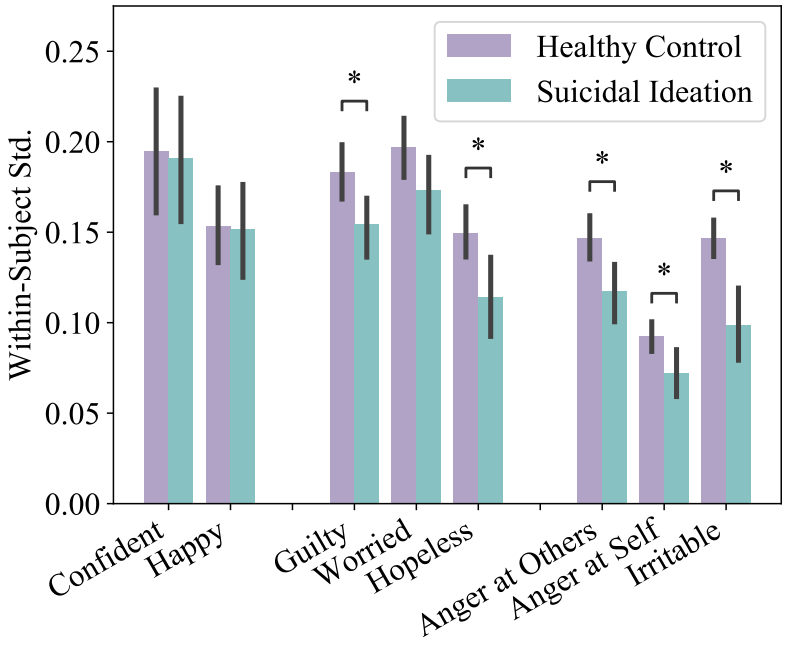


Figure 9.3: The within-subject standard deviation of emotions. * Designates a significant difference (t-test, $p<0.05$).

speech), or that the experiments consider too few subjects.

9.7 Conclusions

In this work, I introduced the EMASS dataset, which allows for an investigation into the relationship between speech from natural conversations, emotion fluctuations, and suicidal ideation. I successfully detected emotion using the still relatively small EMASS dataset by first generating features representative of emotion dynamics on outside data. This shows that the MADDoG algorithm is capable of training sufficiently general representations for use in real-world applications. Finally, I examined how emotion fluctuations detected from speech are able to distinguish subjects with recent suicidal ideation from healthy controls, linking speech, affect, and suicidality.

Collection of the EMASS dataset is currently ongoing and extracted features will be made available through the NIH Data Archive. While this work focused on the momentary ratings of affect, the participant surveys also include questions related to

mood and suicidal ideation. Furthermore, the EMASS dataset includes weekly clinical assessments, which could give a more reliable indication of subject progression. Future work will aim to detect the onset of suicidal ideation and explore this interplay between self-assessed and clinician-assessed mood. The emotion dysregulation features introduced in this work will be key to making this future analysis possible. Furthermore, I will explore the usefulness of these features in other domains, including bipolar disorder.

CHAPTER X

When to Intervene: Detecting Abnormal Mood using Everyday Smartphone Conversations

10.1 Introduction

Previous efforts in bipolar mood monitoring have focused on predicting the intensity of mania and depression symptoms. However, symptom severity thresholds alone are insufficient. Different individuals may have different levels of symptom severity considered healthy, which we refer to as their baseline, and this baseline may change over time. More nuanced information is needed in order to make clinical adjustments or interventions for disease management [6]. This may include the specific characteristics of an individual [139]. This chapter investigates techniques to predict the need for clinical intervention based on deviations from a subject's estimated mood baseline.

Intervention prediction requires intervention labels, or knowledge of when a clinician would choose to act. These labels were obtained by annotating a subset of the PRIORI corpus, referred to as the PRIORI Annotated Mood dataset (PRAM). The collection of the PRAM dataset is described in Section 3.3. I found that clinicians typically identified interventions based on symptom severity ratings that were abnormally high compared to an *individual's* baseline mood.

The PRAM labels were used to create an intervention detection system that could personalize over time, using techniques from the anomaly detection literature, which we refer to as *Temporal Normalization*, or TempNorm. TempNorm initializes with a baseline (a description that captures the range of typical behavior) for the general user population. As the system receives information from an individual, it first transforms these ratings into a continuous value indicative of the abnormality of the symptom severity. It then updates the baseline to personalize to the patterns *of each individual*. I validate the TempNorm framework on the intervention dataset. In particular, we investigate the trade-off between a conservative model that slowly adapts to each subject's baseline and one that instead reacts more strongly to recent mood. I show that TempNorm can be used to transform the symptom severity ratings to effectively predict if an intervention should occur. TempNorm significantly improves on a system using only a single population threshold, achieving an unweighted average recall (UAR) of 0.93±0.04, versus a UAR of 0.80±0.15.

I next investigated the ability of the system to automatically predict interventions from speech, rather than the clinician-assessed symptom severity measures. I combined a neural network with a middle layer consisting of TempNorm. I achieved a UAR of 0.70±0.14 and 0.68±0.12 for clinical and personal conversations, respectively. I find that transcript features perform best for the clinical calls, most likely due to their structured format, while both transcript and emotion features work well for natural speech. These results establish the first results for detecting interventions using clinically-collected and, critically, unstructured natural speech.

The novelty of this work includes: (1) An outcome-based annotation of bipolar mood that identifies the need for clinical interventions; (2) The first work framing bipolar mood intervention detection in the context of anomaly detection using TempNorm; (3) The detection of anomalous mood solely from unstructured, natural speech, enabling real-world applications.

10.2 Related Works

10.2.1 Shortcomings in Speech Mood Recognition

My prior work in Chapter VI has shown that mood recognition can be improved by adapting a system to subject data, allowing it to pick up on subject-specific symptomatology. However, this requires labeled data for all subjects, making the scaling of such systems difficult. Another approach from affect prediction is to use speaker normalization to reduce differences between subjects [9, 173, 185]. While this doesn't require labeled data, it often breaks causality by using all data to calculate the normalization parameters, thus using future subject data in forming the predictions of earlier samples. In a practical system, this would require an enrollment period to determine normalization parameters before being able to make predictions. Enrollment may also need to be carefully constrained to certain types of speech (e.g., non-symptomatic) or else the system may learn incorrect parameters.

The bipolar speech monitoring research explored in Section 1.2 are all steps in the right direction, but most works still require some sort of active participation from the subjects. One of the challenges in mobile health engagement is *app fatigue* – individuals tire of interactions with programs over time or ignore requests to complete evaluations [190, 198]. Passive techniques are likely to be more successful in longer-term monitoring. These techniques should also be able to work in a variety of environments to facilitate continuous monitoring. Faurholt-Jepson et al. demonstrated the feasibility of detecting bipolar mood in everyday phone conversations [65]. Matton et al. reported transcript-based features extracted with Automatic Speech Recognition (ASR) are indicative of depression in bipolar disorder using natural speech [138]. However, these and other methods have not shown that interventions could be automatically triggered. This limits their use in intervention-driven applications, since the predictions are not directly related to clinical action. The output and evaluation

of such systems should instead match their proposed clinical use [56].

Researchers have used reinforcement learning (RL) to explore this concept of interpretable and actionable systems for interventions. Typical supervised machine learning requires pairing predictions with ground truth labels. However, this is not always feasible when there is an unclear relationship between single actions and meaningful outcomes. RL instead defines and optimizes for long-term goals using domain-specific *reward functions*. For example, work in epilepsy has been evaluated using a reward function that penalizes the occurrence of seizures and learns the optimal pattern of deep brain stimulation [86]. Research in sepsis has employed a reward function tied to patient mortality and proposes the ideal personalized clinical intervention strategy [119]. Work in HIV has used a reward function based on changes in blood test measures and selected the best combination of drugs for therapy [160]. This work differs from that in RL in that our ground truth directly pairs each sample of speech with an indication of the need for intervention.

10.2.2 Anomaly Detection

Anomaly detection identifies unusual measures in data, with common applications including outlier removal and fraud detection [39]. Basic methods of anomaly detection can take advantage of the distribution of data and designate points above a certain standard deviation or other measure as anomalies. For example, different forms of the moving average (MA) and variance can be used to de-trend and scale sequential data, as in [22, 99, 153]. Autogressive (AR) models are commonly used in sequence anomaly detection to forecast the likely value of the next sample using a certain number of prior samples [88]. Anomalies can then be designated as a deviation of an actual data-point from the predicted value. Autoregressive-moving-average models, or ARMA, combine both the MA and AR models, and are effective at detecting anomalies in a variety of fields [109, 143, 165]. All of these models may take

into account prior knowledge about the domain, such as seasonal trends and and the base rate of occurrence of anomalies.

Recent work has used neural networks to detect anomalies in data with unknown distributions. Autoencoders learn a compressed representation of the data, with anomalies identified by higher reconstruction error [13]. Malhotra et al. trained an LSTM to forecast time series predictions and then classified deviations from actual values as anomalies [136]. Generative adversarial networks, or GANs, learn a latent space where anomalous data is more clearly distinguished [219]. These methods are unsupervised and focus on capturing aspects of the input with higher variance. Similar approaches have been ineffective at representing affect, as emotion and mood have a much lower varying nature when compared to other aspects of speech [30, 117].

Supervised anomaly detection is typically trained using standard classification methods with one category for normal data and one for anomalous data [39]. The main difference from typical classification problems is that the label distribution is strongly unbalanced and biased against anomalies. This bias can result in being unable to learn a robust representation for anomalies, due to a lack of examples. Additionally, it may be difficult to find clear representative samples of anomalies [79]. This is because anomalies are not necessarily defined by the presence of certain attributes, but are instead defined by the amount of deviation from a typical distribution. Modelling the problem as regression instead of classification can help avoid this issue by defining a continuous label for the abnormality of each sample [39]. However, previous work has not yet examined mood monitoring in the context of anomaly detection, leaving the definition of mood abnormality undefined. This work establishes this definition using the exponential moving average (EMA) and exponential moving variance (EMVar), similar to [22, 99, 153], to track typical mood for subjects. I then define mood abnormality as a mood rating's deviation from this continuously updated baseline, as discussed in the following section.

10.3 Temporal Normalization

During clinical annotation of the PRAM dataset (Section 3.3), the most common rationale for flagging an intervention was mood ratings substantially above a subject's baseline. Because of this, we focus on how to model a subject's baseline using the YMRS and HDRS ratings to best predict anomalous mood and the need for interventions. In order to be successful, our system should be able to do the following: (1) estimate a baseline for subject mood, (2) predict anomalies based on a deviation from this baseline, (3) produce actionable predictions, even when little subject data has been seen. This section presents an example of how to model bipolar mood, motivated by these goals and conversations with our clinical collaborators.

I formalize the problem as anomaly detection, using the EMA and EMVar, similar to [22, 99, 153]. This converts the problem from contextual anomaly detection to an easier point anomaly detection. For simplicity, we denote this procedure as *Temporal Normalization* (TempNorm) [1]. I ground TempNorm in its application to the YMRS and HDRS ratings. The full process is described in Algorithm 3 and shown in Figure 10.2, with each step explained in the following text.

I first subtract the YMRS and HDRS ratings by six and divide by four, based on an initial estimate of mean and standard deviation. These values were selected to closely match the within-subject rating means and standard deviations. This normalization also maps a rating of ten to one standard deviation from the mean. This is desirable, as ten was the threshold used in previous PRIORI experiments in the definition of a symptomatic state [75, 114]. Because the mood ratings should now have an approximate mean of zero and standard deviation of one, we initialize the EMA to zero and the EMVar to one. This establishes a subject's starting baseline as the population's baseline. I call this initial state the *population prior*.

[1] An interactive demonstration of Temporal Normalization is available at `http://www.johngideon.me/projects/TempNorm/`

Algorithm 3 Temporal Normalization (TempNorm)

Input: X, the 1d array to be normalized
Input: $t_{1/2}$, the half-life parameter
Output: Y, the 1d normalized output array

$\lambda \leftarrow 1.0 \quad - \sqrt[t_{1/2}]{0.5}$ ▷ Get decay from half-life
$\mu \leftarrow 0$ ▷ Initialize EMA to 0
$\sigma^2 \leftarrow 1$ ▷ Initialize EMVar to 1
for $i = 1, ..., length(X)$ **do** ▷ Loop through all samples
    $\Delta \leftarrow X[i] - \mu$ ▷ Get sample and EMA delta
    $Y[i] \leftarrow \Delta / \sigma$ ▷ Normalize current sample
    $\beta \leftarrow \lambda \times \Delta$ ▷ Scale delta based on decay
    $\mu \leftarrow \mu + \beta$ ▷ Update EMA
    $\sigma^2 \leftarrow (1.0 - \lambda) \times (\sigma^2 + (\beta \times \Delta))$ ▷ Update EMVar
end for

I then approximate each subject's baseline over time using the EMA and EMVar. I normalize new data points by subtracting the EMA and dividing by the EMVar before updating these running statistics using the new values. This results in the first sample being unchanged because the original population prior is EMA=0 and EMVar=1. As the system sees new samples, the mood baseline and subsequent normalization adapts to that subject's patterns. TempNorm does not require the data to be sampled at a fixed rate, which is beneficial as subjects periodically have missing clinical assessments.

TempNorm requires one parameter, half-life ($t_{1/2}$), in order to control the contribution of new data to the running mean and variance. Half-life is described in units of the number of new samples needed to diminish the weight of old data to 50% (Figure 10.1). As half-life increases, the baseline takes longer to adapt to subject mood and remains conservatively closer to the original population prior. A half-life of infinity results in a system that only relies on the population prior and does not adapt to subject mood ratings. This makes it comparable to a system without TempNorm – one that has a single threshold across all subjects. Conversely, decreasing the half-life results in a system that concentrates more on recent data. A half-life of zero would

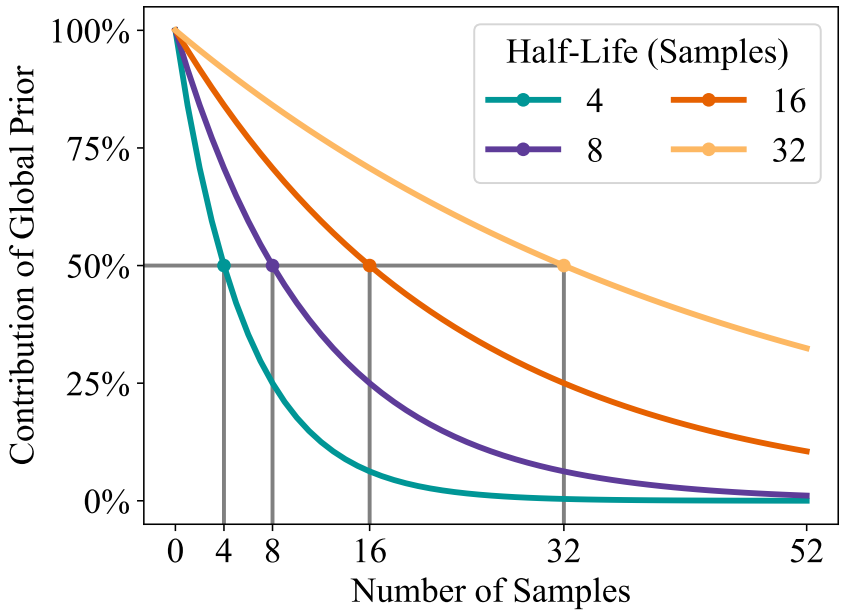


Figure 10.1: The contribution of the original population prior distribution for different half-lives after certain numbers of samples have been observed.

only use the newest sample to calculate mean and variance. This would always result in a baseline variance of zero and produce divide-by-zero numerical issues in the normalization. Because of this, we restrict half-life to be any value greater than zero. Much of this work is focused on investigating the impact of half-life, as the model is heavily affected by its choice. I determine the weight of new samples, λ, using the following equation:

$$\lambda = 1.0 \quad - \sqrt[t_{1/2}]{0.5}$$

The EMA and EMVar are updated with each new sample, weighted by λ. This causes a hybrid global/speaker normalization of the data, depending on the half-life parameter and how many samples have been observed. This has the desired quality of gradually adapting to the most recent subject data, while also providing a measure of how anomalous a sample is versus the baseline.

The mood ratings are normalized separately, as the intensity of each subject's

baseline mania and depression may differ. Furthermore, different subjects may have different correlations between these ratings. So for simplicity, we assume that both mood ratings are independent and model them with two separate TempNorms. This assumption will be later validated using the PRAM dataset.

I select a lower threshold of one standard deviation to represent a subject's typical mood. This matches the threshold for symptomatic mood used in our previous work [75, 114]. TempNormed YMRS and HDRS below this threshold are considered typical. I define an upper limit of two standard deviations as atypical mood, based on conversations with our clinical team. I define a *mood anomaly* as a sample with a TempNormed YMRS or HDRS above the upper limit. I leave the range between one and two standard deviations undefined, focusing only on regions with clear behaviors as in [71]. This work presents just one example of how to model bipolar mood anomalies. Future work will investigate alternative models.

I validate our model using the PRAM annotations. I hypothesize that weeks flagged for intervention are weeks that our model should designate as anomalous (over two standard deviations from the EMA), while weeks marked without an intervention should be typical (within one standard deviation of the EMA). I do not assess model performance in the undefined region. It is important to note that the anomalous and typical category changes with half-life: TempNorm transforms the mood ratings based on this value (see Table 10.1).

I evaluate performance using unweighted average recall (UAR), an average of the recall over each category. This is desirable, since the distribution of the annotations is biased toward non-interventions. I sweep through different half-lives to gain insight into how annotators balance historical and new mood symptom information. I only consider subjects that have at least one typical and one anomalous week so that a valid UAR can be calcuated. As a result, the number of available subjects varies with half-life (see Table 10.1).

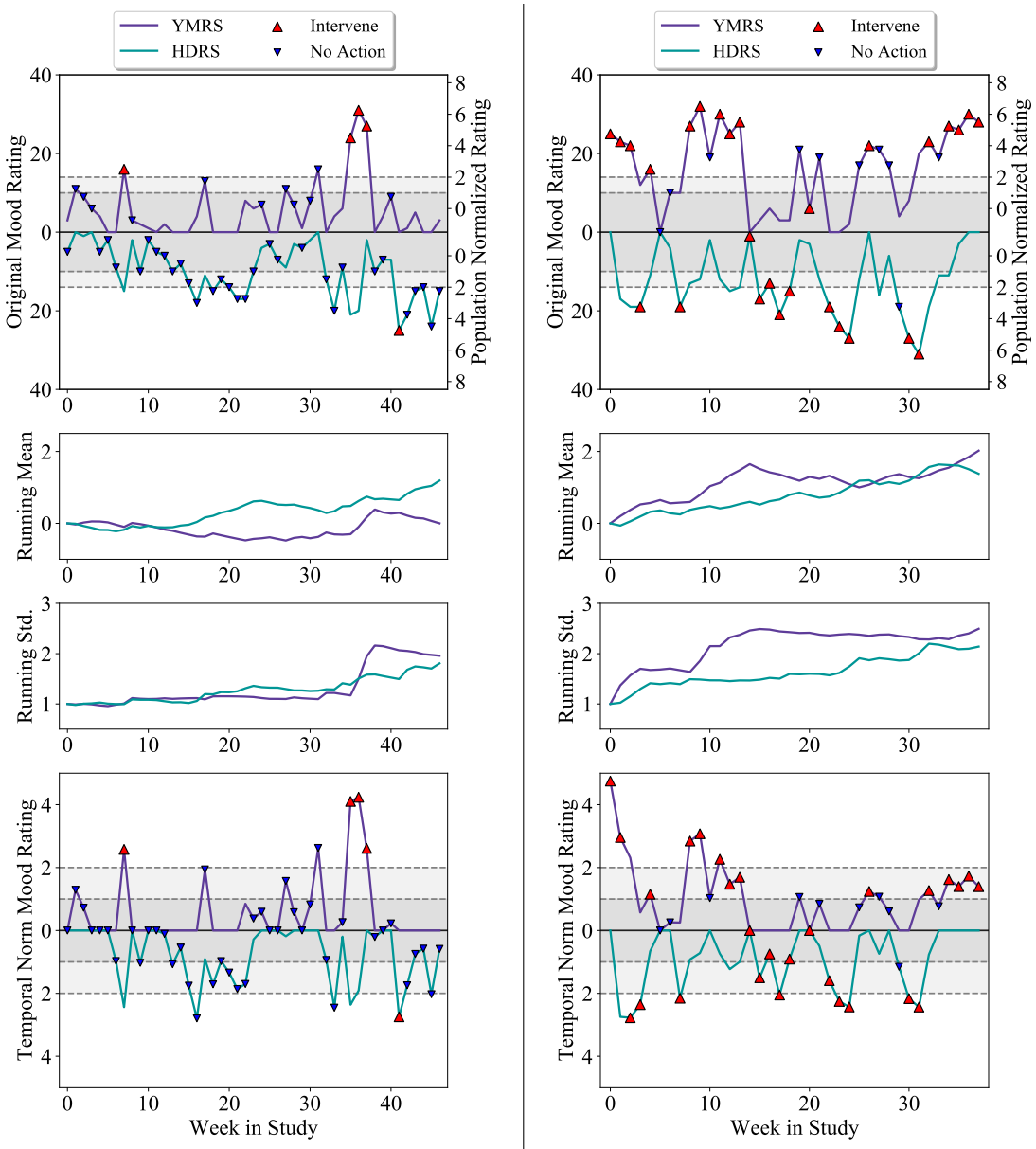
Table 10.1: TempNorm Mood Ratings Compared with Annotation. Ratings below one are *typical*; those above two are *anomalies*; ratings between one and two are *unused*. The number of samples flagged for intervention in each region is shown in parentheses. Relying only on the global prior ($t_{1/2} = \infty$) results in a system with many false positives. Highlighted results are not significantly different from one another.

	Number	Number Samples			UAR
$t_{1/2}$	Subjects	Typical	Unused	Anomaly	Mean ± Std.
1	13	322 (22)	94 (24)	136 (25)	0.72 ± 0.19
2	11	352 (20)	113 (26)	87 (25)	0.83 ± 0.13
4	11	366 (16)	116 (28)	70 (27)	0.87 ± 0.12
8	12	339 (11)	117 (25)	71 (35)	0.90 ± 0.07
16	13	317 (6)	126 (18)	84 (47)	0.93 ± 0.04
32	13	282 (2)	146 (16)	99 (53)	0.91 ± 0.07
64	13	268 (2)	131 (5)	128 (64)	0.89 ± 0.09
∞	13	263 (2)	71 (1)	193 (68)	0.80 ± 0.15

I fit a linear mixed-effect (LME) model in R [19, 171] to determine if the UAR of different half-lives are significantly different. I treat half-life as the fixed effect and subject as the random effect. All tests use a 0.05 significance threshold. I perform an analysis of variance (ANOVA) over the LME model to determine if there is any significant effect of half-life. I then perform a post-hoc pairwise comparison test with Tukey weighting using the emmeans package [126].

My findings in Table 10.1 indicate that a half-life of 16 provides the best match to clinical annotation, with 0.93 ± 0.04 UAR, although multiple half-lives achieve comparable performance. I highlight half-lives that are not significantly different from one another. The rows that are not highlighted (1 and infinity) are significantly worse than at least one of the highlighted rows.

I show the full TempNorm procedure for two subjects using the half-life of 16 in Figure 10.2. Figure 10.2a shows a subject on which TempNorm performs well, while Figure 10.2b presents a difficult case (the subject in Figure 10.2b has particularly unstable mood). Figure 10.2a demonstrates how TempNorm can detect mood anomalies even given an increasing depression baseline (the turquoise lines). Blue markers outside the shaded regions indicate false positives, while red markers in the middle-most



(a) Subject with good TempNorm performance

(b) Subject with difficult ratings for TempNorm

Figure 10.2: TempNorm using a half-life of 16 samples for two subjects. Each gives four plots. (1) Depicts the original mood ratings and flags for intervention on the maximum of the mania or depression rating. The right y-axis gives the initial population normalized mood. The dashed lines and shaded regions depict the upper and lower mood thresholds of one and two standard deviations, respectively. These are used to differentiate typical and anomalous mood. (2) Gives the running scaled mean mood rating. (3) Gives the running scaled standard deviation of the mood rating. (4) Shows the TempNorm output with similar thresholds to the first plot. Normalized mood ratings below zero are truncated to zero.

region are false negatives. The top plot shows the population prior system, which only relies on a fixed threshold and does not adapt to the increasing baseline. This results in 12 false positives, while TempNorm (shown at the bottom) decreases this to just four. Figure 10.2b gives an example of a subject with highly fluctuating mood and many intervention flags. The system normalizes many weeks to between one and two standard deviations because of the frequency with which this subject experiences heightened mood. This highlights the importance of clinical judgment in determining an intervention threshold – with this individual potentially benefiting from a lower cutoff. Despite this, TempNorm makes distinguishing extreme examples easier. The population prior with a fixed threshold results in eight false positive and two false negatives. TempNorm removes all false positives at the expense of two additional false negatives, but with a large increase in uncertain mood predictions (between one and two standard deviations).

This section has demonstrated the first advantage of TempNorm – it transforms the ground truth to resemble anomaly detection and creates more actionable predictions. While this section introduced TempNorm in the context of the YMRS and HDRS mood ratings, it would be easy to extend the procedure to other sequential data. The remainder of the work explores the other two main benefits of TempNorm: (1) It initially behaves as a hybrid global/speaker normalization. After sufficient data, depending on the selected half-life, it acts like speaker normalization, providing the performance benefits of speaker normalization, without a requiring an enrollment period. (2) Each subject's mood ratings are self-normalizing, removing individual biases and resulting in reduced biases between subjects. This balances the dataset, making the learning of both typical and anomalous mood more straightforward.

10.4 Features and Preprocessing

I now focus on predicting the need for intervention from speech. In particular, we are interested in two different types of experiments, predicting mood anomalies from: (1) recorded clinical assessment calls, or (2) personal calls (non-clinical) from the same day as each assessment. I denote these as the *assessment* and *day-of* experiments, respectively. Note that the assessment calls themselves are never included in the day-of experiments. In this work, we focus only on calls from the day of the assessment because we hypothesize that they are most associated with the assessment label. Previous research has demonstrated that recency bias affects retrospective recall, causing clinical ratings to be strongly impacted by the most recent events [4]. For each experiment, we use different combinations of two speech feature sets – emotion and transcript.

10.4.1 Emotion Features

I have previously shown that there is a connection between fluctuations in emotion and mood in Chapter IX. In this Chapter, we validate this hypothesis by extracting measures of emotion and relating statistics derived from these measures to the clinical mood measures. I estimate emotion from the recorded speech using the MADDoG model introduced in Chapter VIII. MADDoG allows the model to learn emotion, while also finding a representation that is similar across different datasets. I train MADDoG using the same three emotion datasets as in Chapter VIII – PRIORI Emotion, IEMOCAP, and MSP-Improv.

Using this MADDoG model, we extract features for the final mood analysis. These features consist of binned segment-level estimates of both activation and valence that represent emotion dynamics over each assessment or day. However, we are interested in ensuring that no emotion labels are used in the eventual test set. To accomplish this, we train six different MADDoG models. The first five models are trained and

tested in a round-robin manner using five folds and produces test predictions for all labelled data. One additional model is trained with all the 13,611 emotion-annotated segments and used to predict the remaining unlabeled segments. This results in three activation bins and three valence bins per PRIORI segment, or six total dimensions.

I hypothesize that the distribution of emotion over the course of an assessment or day is indicative of mood. To quantify this, we first concatenate all segments over the course of an assessment or day, depending on the experiment. I then take 31 statistics across the segments, which we previously demonstrated were related to mood [73]. This results in a final 186-dimensional feature set. This includes the mean, standard deviation, skewness, kurtosis, minimum, maximum, and range of the emotion bin predictions. I extract different percentiles (1, 10, 25, 50, 75, 90, 99) and percentile differences (25-50, 50-75, 25-75, 10-90, 1-99). I perform linear regression on the segment emotion estimates and incorporate the fit parameters and error (R^2, mean error, MSE) as features. Finally, we determine the percentage of the binned predictions above various thresholds (10%, 25%, 50%, 75%, 90% of the range).

10.4.2 Transcript Features

I transcribe the calls using an ASR model, which was implemented in Kaldi, an open-source, freely available speech recognition toolkit [167]. The model was built following the 'nnet2' recipe and was trained on the Fisher English Corpus [42]. When tested on the transcribed subset of the PRIORI dataset, it obtained a word error rate of 39.7%. I recognize that this is high. However, our data consists of unconstrained, natural speech in the presence of noise, so we expect imperfect transcriptions. Previous work showed that transcript-based features extracted from the ASR transcripts were useful [138].

I extract call-level features from the assessment transcripts to use in our assessment experiments. For our day-of experiments, we concatenate the transcripts of

all *personal* calls made on the day-of an assessment for each subject and assessment date (this excludes assessment calls). I extract day-level features from these merged transcripts. This results in a 208-dimensional feature set, which I divide into five categories. See [138] for more details.

I employ the Linguistic Inquiry and Word Count (LIWC) tool [164], a psycholinguistic analysis resource used in previous work to detect mental health states [44, 51], to compute the percentage of words belonging to 63 different language categories. Some of these categories are measures of **semantic content** and are related to psychological constructs (e.g. affect, biological processes) and personal concerns (e.g. work, death). The other categories measure aspects of **linguistic style**; these include 18 part of speech (POS) categories, three verb tense categories, and swear word, non-fluency, and filler categories. I extract 22 supplemental measures of linguistic style, including five additional POS categories, five POS ratios (e.g. adjectives:verbs), and 12 measures of speech complexity and verbosity (e.g. mean words per speech segment).

I apply speech graph analysis, introduced by Mota et al. to quantify thought disturbances in individuals with mania and schizophrenia [36, 147], as our final means of measuring linguistic style. I form speech graphs by representing each unique word as a node and inserting an edge for every pair of words uttered consecutively within the same speech segment. I create three graphs from each transcript that: (1) use the words directly, (2) use the lemmatized form of each word, and (3) represent each word as its associated POS. I extract 12 measures per graph, including average degree, density, diameter, the size of connected components, and loop, node, and edge counts (see [147] for a full list). I also include a version of each feature that is normalized by total word count, providing us with 72 total graph measures.

I use Kaldi to generate aligned word and phone timing annotations for each transcript. From this output, we extract 43 features that quantify **speaker timing**

patterns. I extract the same features for words, phones, and pauses: (1) statistics (mean, median, standard deviation, min, max) applied to the durations of all instances (e.g. mean word duration), (2) statistics (same set) applied to the per second timing within all segments (e.g. mean words per second across segments), (3) total count, and (4) per second timing over the whole transcript. I also extract total call duration, total subject speaking duration, ratio of subject speaking duration to total duration, total pause duration, ratio of pause duration to total duration, segment count, segments per minute, count of short utterances (lasting less than 1-second), and short utterances per minute, some of which were motivated by [149].

I use ASR confidence scores as measures of **speaker intelligibility** based on the idea that ASR has higher confidence for well enunciated speech. I apply statistics (mean, median, standard deviation, min) to the segment-level confidence scores to obtain four features (max was not used because it was almost always 1). Lastly, we quantify the presence of **non-verbal expressions** by extracting counts of instances of laughter and noise detected by the ASR model, normalized by word count.

10.4.3 Data Selection

I reduce our dataset to the highest quality data subset in order to focus on the impact of TempNorm on bipolar mood ratings (Table 10.2). I first remove all healthy controls to ensure the model specializes in individuals with bipolar disorder. I then remove subjects with phones other than the Samsung Galaxy S5s, as prior work has highlighted challenges with the other phone models [75]. I require data to have at least five segments, 100 words, and valid ASR transcripts (as in [138]). Finally, we require subjects to have at least eight samples (assessments/days), as this work focuses on adapting to a subject's baseline over time. This results in a total of 23 subjects and 533 samples (calls) for the assessment experiments and 17 subjects and 369 samples (days) for the day-of experiments. Note that removing data changes the ground truth

Table 10.2: Restrictions causing the reduction of data for both the assessment and day-of experiments.

Restriction	Number of Samples Assessment	Day-of
None	1515	1515
No healthy controls	1319	1319
Only S5 devices	680	680
5 segments, 100 words, Valid ASR	556	417
Subjects need 8 samples	**533**	**369**

because those samples no longer contribute to the baseline used in TempNorm (see Section 10.3).

10.5 Modelling

The goal of the model is to predict the abnormality of the mood (TempNorm symptom severity) and to use this prediction to determine whether or not an intervention is needed. I use the same model and training methodology on the assessment and day-of experiments. I train two unimodal systems (i.e., only transcript and only emotion) and a multimodal early fusion system, resulting in a total of six systems, three for assessment and three for day-of.

I use a dense neural network (DNN), which has been effectively employed for mood recognition from static features [216]. It consists of six fully connected layers with Randomized Leaky Rectified Linear Activation (RReLU) activations after the hidden layers, as seen in Figure 10.3. The output layer has a linear activation and predicts Temporally Normalized YMRS and HDRS ratings.

I perform TempNorm in the feature space to match the label space. This makes both the labels and features relative to a subject baseline and allows for the detection of anomalies in both. Preliminary experiments without feature TempNorm had poor performance because the feature and label baselines drifted apart. I ap-

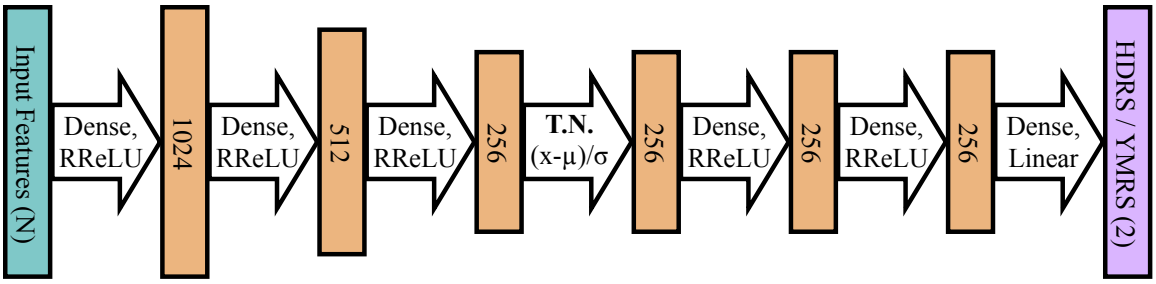


Figure 10.3: The DNN used to predict mood abnormality, modified with a TempNorm Layer after the third hidden layer to learn a feature baseline.

ply a *TempNorm Layer* to the 256-dimensional representation after the third fully connected layer. This applies TempNorm independently for each of the third layer's inputs using the same half-life parameter as the one used for the mood ratings. The feature model adapts to each subject's samples and, over time, performs subject-normalization over the mid-level representation. The model requires valid features to ensure that the label and feature EMA and EMVar remain synchronized. Future work will explore how to handle missing data (features or labels) and the placement of the TempNorm Layer.

The model is trained and tested in a round robin manner with five folds. Folds are kept subject independent by randomly assigning each subject to one fold. Data from each of the remaining subjects are split randomly with one fifth of samples used for development and the rest used for training.

Samples are batched by subject and are randomly reordered for data augmentation when training. Reordering the samples changes the baseline produced by the model and the ground truth at each time step, effectively increasing the amount of training data. Prior work has shown that mood changes in bipolar disorder are Markovian and that current mood is the most predictive of mood at the next assessment [145]. Due to the lack of long-term connection, we hypothesize that this reordering is an acceptable trade-off to augment the training data.

I fit the model using a weighted mean squared error (WMSE) loss to compensate

for the relative rarity of anomalous mood. I calculate this WMSE loss for each mood output and sum them to form the total loss. The model makes the assumption that our measure of mood abnormality approximately follows a standard normal distribution. I examine this assumption further in Section 10.7.2. Given this assumption, we set the weight of a sample to be inversely proportional to the probability density function (PDF) of the ground truth value. This gives higher weight to less common moods. I cap this weight at a maximum of 25 since the PDF increases rapidly for higher values, making rare samples dominate if unchecked. This method results in each sample having a different weight each epoch, depending on its order of appearance after randomization. This is because the abnormality (and weight) of samples is determined from the context of what comes before.

I train for 50 epochs. The first 10 epochs are used to pre-train the model without TempNorm. The next 40 epochs use TempNorm and validate to find the best stopping epoch for testing. The validation performance is also evaluated using WMSE. The WMSE takes the maximum of the two output mood predictions and the maximum of the TempNorm ground truth. This estimates the ability of the system to measure the abnormality of samples – focusing on the most extreme of the two moods. I found that this method of validation more closely matched the test setup and provided better performance.

Each subject's test performance is measured in UAR. I denote typical mood as TempNorm rating under one standard deviation, an anomaly as above two standard deviations, and unused as between one and two standard deviations. This strategy is similar to one by Georgiou et al., where they predict the amount of blame expressed in speech recordings of couple's therapy and just focus on the upper and lower 20% of samples [71]. The amount of prototypical data and available test subjects varies for each half-life, as discussed in Section 10.3.

It is important to note that although the test data we evaluate are only in the

range of typical or anomalous, it is possible that our system will predict values in the unused range. I binarize the unused range to calculate UAR by introducing a threshold at 1.5 standard deviations. I assign estimates below 1.5 to typical mood and those above to anomalous mood.

For each of the two main experiments (assessment and day-of), we examine three combinations of features: (1) emotion, (2) transcript, (3) an early fusion of both. My initial experiments found that applying global normalization (z-normalization using all training data) to transcript features worked best. No normalization was necessary for emotion. As such, we apply global normalization to the transcript features, in both the unimodal and early fusion sets. I run each experiment for a total of 100 iterations, each with a different random seed. This helps to compensate for the randomness of selecting subjects for different folds and the neural network initialization. This produces a final UAR matrix of size $100 \times \#Subjects$ for each if the six tested methods.

10.6 Results

UAR measures our ability to judge the need for interventions. The performance is given as the UAR averaged over all subjects with enough data for the experiment (at least one typical and one anomalous sample). The number of subjects and samples changes depending on the half-life (see Section 10.3).

I test for the significance of each feature set and half-life on UAR by fitting a linear mixed-effect model, similar to Section 10.3. I consider feature set and half-life as fixed effects and subject and random seed as random effects. All tests use a 0.05 significance threshold. I first perform an analysis of variance (ANOVA) to determine if there is any significant effect of either feature set or half-life and find significance for both in each experiment. Next, we perform a post-hoc pairwise comparison, as in Section 10.3. I denote significantly better results than the emotion feature set with

Table 10.3: Assessment and day-of experiment results. The amount of subjects, typical samples, and anomalous samples is shown. Highlighted results show half-lives that do not produce significantly different results, given a certain feature set. An asterisk indicates results significantly better than the emotion features for the same half-life.

(a) Assessment Results (533 total assessment calls)

	Total Number			Feature Set UAR (Mean ± Std.)		
$t_{1/2}$	Subjects	Typical	Anomalous	Emotion	Transcript	Fusion
1	23	312	117	0.49 ± 0.09	0.61 ± 0.10*	0.59 ± 0.09*
2	21	320	72	0.58 ± 0.14	0.67 ± 0.10*	0.68 ± 0.12*
4	21	326	59	0.61 ± 0.12	0.68 ± 0.13*	0.70 ± 0.14*
8	19	293	68	0.59 ± 0.11	0.66 ± 0.13*	0.68 ± 0.12*
16	19	270	78	0.60 ± 0.09	0.68 ± 0.13*	0.69 ± 0.14*
32	19	245	93	0.59 ± 0.12	0.68 ± 0.13*	0.70 ± 0.14*
64	19	237	124	0.57 ± 0.10	0.67 ± 0.12*	0.70 ± 0.15*
∞	19	223	174	0.54 ± 0.07	0.67 ± 0.12*	0.68 ± 0.13*

(b) Day-of Results (369 total days)

	Total Number			Feature Set UAR (Mean ± Std.)		
$t_{1/2}$	Subjects	Typical	Anomalous	Emotion	Transcript	Fusion
1	17	216	86	0.53 ± 0.08	0.55 ± 0.08	0.56 ± 0.09
2	17	231	55	0.59 ± 0.10	0.61 ± 0.11	0.63 ± 0.13*
4	17	223	49	0.59 ± 0.12	0.64 ± 0.13*	0.65 ± 0.11*
8	15	192	43	0.63 ± 0.14	0.66 ± 0.12	0.68 ± 0.12*
16	15	176	61	0.62 ± 0.11	0.63 ± 0.12	0.65 ± 0.11
32	15	166	76	0.57 ± 0.10	0.60 ± 0.12*	0.61 ± 0.13*
64	15	159	96	0.54 ± 0.07	0.59 ± 0.13*	0.59 ± 0.13*
∞	16	163	127	0.51 ± 0.05	0.56 ± 0.11*	0.56 ± 0.13*

an asterisk in Table 10.3.

There were no significant differences between the transcript and fusion sets in either the clinical or the day-of calls. I describe the patterns in more detail in the following sections. I highlight the best performing half-lives for each feature set that are not significantly different from one another in Table 10.3.

10.6.1 Assessments

The assessment results are given in Table 10.3a. Regardless of half-life, the transcript and fusion feature sets significantly outperform emotion. There are no significant differences when appending the emotion features to transcript features (fusion). Assessment calls consist of the YMRS and HDRS interviews, and as such have a structure not present in natural speech. Because of this, the transcript features likely capture aspects of the questionnaires, giving them an advantage.

The performance of mood anomaly detection is mostly insensitive to half-life. The only exception is for a half-life of one. I find that a half-life of one changes the baseline too rapidly and results in significantly worse performance, using both emotion and fusion features. However, all half-lives between two and infinity provide stable results, and are not significantly different from one another. This stability is particularly evident in the transcript features. Again, this is likely due to the close relationship between the transcript features and the answers to the interview questions.

10.6.2 Day-of

The day-of results are given in Table 10.3b. The performance of the transcript features decreased, relative to the assessment experiment. This demonstrates how the efficacy of transcript features change when the assumptions of interview structure are no longer present – a result also shown recently in [138].

The emotion features attained similar performance in the two experiments. The similarity of the performance of emotion features between experiments indicates that they are capable of capturing mood-related aspects of speech present in both structured and natural conversations. In fact, the transcript features now only significantly outperform the emotion features in about half of the results.

I find that the day-of experiment is more sensitive to different half-lives, when

compared to the assessments. When working with non-clinical conversations it is especially important to establish a clear baseline. While a small half-life devalues old data too quickly, an overly large half-life takes too long to converge to subject normalization. In effect, the large half-lives result in systems that depend on the population prior distribution. This causes each subject's mood ratings to be biased, depending on how closely the population prior distribution matches the subject's actual distribution. This mismatch of subject mood distributions can complicate classification and result in worse performance.

10.7 Discussion

The choice of half-life has a strong effect on the outcome of most results in this work. Figure 10.4 shows the UAR from the annotation experiment, as well as the assessment and day-of experiments. Assessment and day-of experiments use the fusion features and the shaded error represents the standard deviation between iterations. Because there are no iterations for the annotation experiment, no error is shown. While the assessments are mostly insensitive to half-life (see Section 10.6.1), the half-lives of 8 and 16 provide the highest UAR for the day-of experiment and annotations, respectively. This section explores two factors that contribute to their performance: (1) enrollment length and (2) the distribution of the normalized ratings.

10.7.1 Enrollment Length

TempNorm begins with a population prior distribution and eventually learns a subject baseline. After sufficient samples it behaves like subject normalization, favoring recent data. Because of this, half-life controls both how quickly to disregard the population prior and the effective window length constructing a subject baseline. A half-life of 16 versus 8 will effectively incorporate double the samples into the subject baseline, but should also take about twice as long to converge.

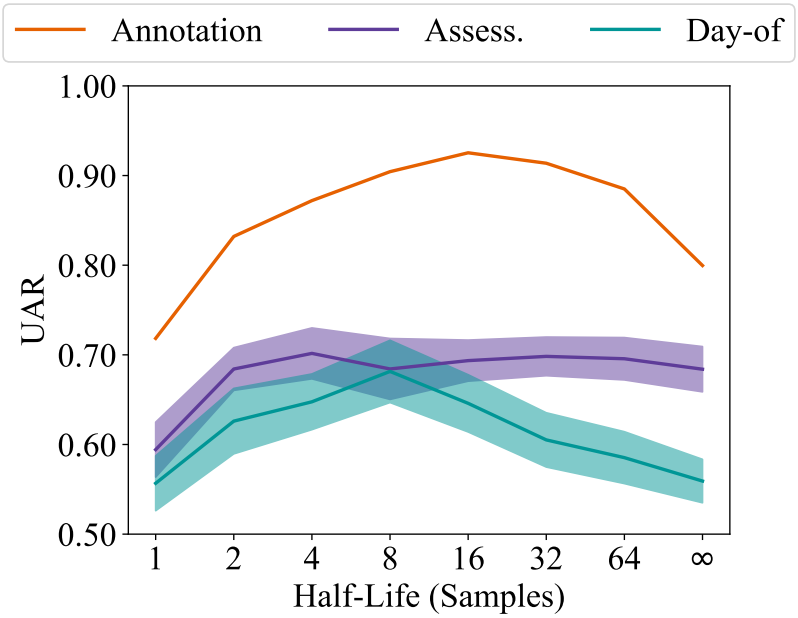


Figure 10.4: Summary of experiment results. Assessment and day-of experiments use fusion features with the shaded region showing the standard deviation between random iterations.

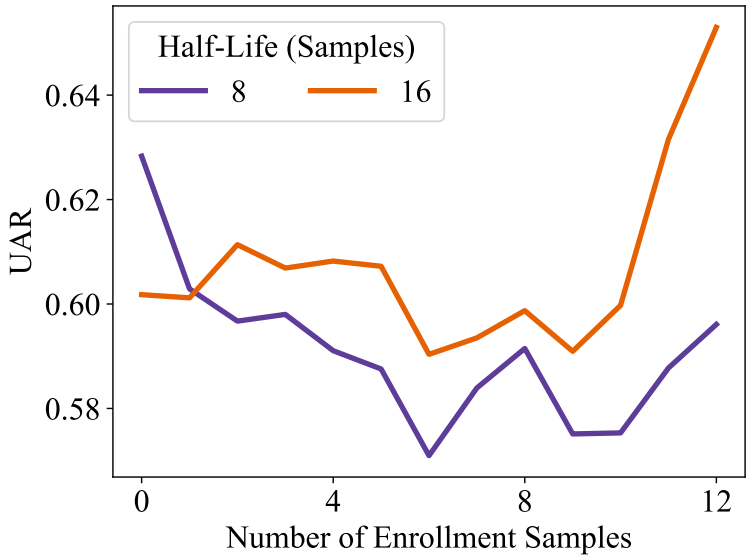


Figure 10.5: The performance of the day-of fusion experiment, considering different enrollment periods and half-lives.

I examine our day-of results after differing enrollment periods to see the change in performance during the adaptation process. For example, an enrollment period of four samples uses those samples to calculate the initial subject baseline, but not to

make predictions. The remaining samples both continue to adapt the baseline and are the sole focus of the UAR calculation. I focus only on those seven subjects with enough data to consider an enrollment period of 12 weeks, so that the subjects stay the same with increasing enrollment amounts.

Figure 10.5 shows how the mean subject UAR changes as the system accumulates subject data. The figure shows two half-lives, eight (purple) and 16 (orange). As before, we see that a half-life of eight provides the best performance when there is no enrollment data (enrollment of 0). However, it tends to worse performance when the subject baseline is initialized with enrollment data. Conversely, a half-life of 16 produces results that are fairly stable with enrollment periods of up to ten samples. This stability is likely associated with the broader effective window length for a half-life of 16. It then sharply increases in performance. While the method takes longer to achieve higher performance, the wider window provides a better subject baseline for comparison, versus a half-life of eight with no enrollment restriction. This demonstrates that higher half-lives can be beneficial – but only after sufficient data. Because there is not sufficient time for a half-life of 16 to reach its full potential, a half-life of eight provides better overall performance when considering no enrollment restrictions in the personal call experiment (see Table 10.3b).

10.7.2 Distribution of the Normalized Mood Ratings

While subject adaptation is one potential improvement caused by TempNorm, another is the self-normalization of the ground truth. Once a subject's baseline is determined, the mood is de-biased and scaled so that it has a mean of zero and a standard deviation of one. As explained in Section 10.5, our model makes the assumption that the ground truth during training should have a unit normal distribution. Given this assumption, we weight each sample's loss inversely with respect to the unit normal PDF evaluated at the ground truth mood rating.

Table 10.4: Mean and Standard Deviation of Normalized Mood Ratings.

	Mania			Depression		
$t_{1/2}$	Mean	Std.	R^2	Mean	Std.	R^2
1	0.92	23.79	0.03	0.07	1.87	0.98
2	0.08	2.48	0.38	0.09	1.38	0.99
4	-0.06	1.25	0.82	0.13	1.20	0.98
8	-0.14	1.12	0.86	0.21	1.15	0.98
16	-0.20	1.14	0.85	0.34	1.17	0.97
32	-0.24	1.22	0.83	0.49	1.23	0.97
64	-0.27	1.31	0.82	0.62	1.32	0.97
∞	-0.28	1.57	0.78	0.96	1.69	0.94

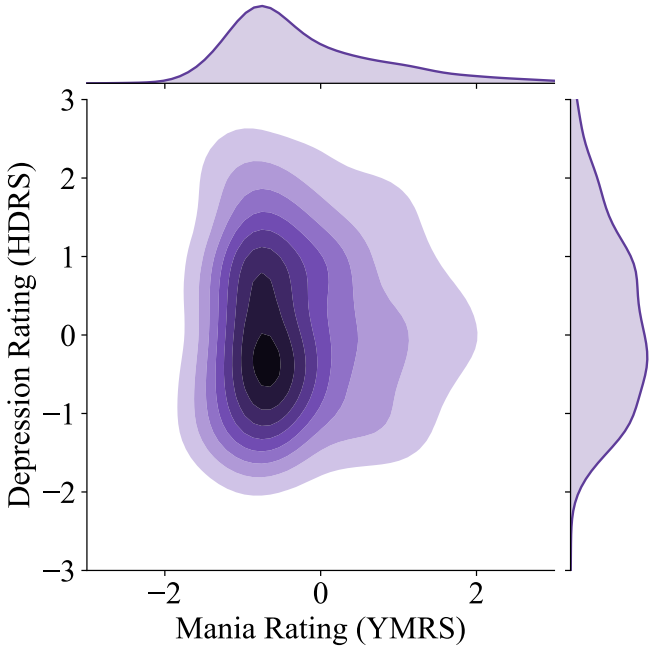


Figure 10.6: Mood Rating Distribution After TempNorm ($t_{1/2} = 8$).

In order to verify if this assumption is valid, we evaluated the distribution of the TempNorm mood at different half-lives. Table 10.4 shows the calculated means and standard deviations, as well as the R^2 value from a normal probability plot. I find that decreasing the half-live down to eight consistently causes the mean to trend toward zero, the standard deviation to trend to one, and the R^2 to get closer to one. However, this trend does not hold with very small half-lives. In particular, a half-life of one causes the standard deviation of normalized mania to increase sharply and the

R^2 value to approach zero. Because the effective window of the EMVar becomes so small, only a few identical values in a row can cause it to approach zero. This is particularly a problem for mania because the distribution is skewed toward ratings of zero (Figure 3.2). After a few weeks with no mania symptoms, the next week with relatively higher mania symptoms will be substantially scaled upwards.

The compromise half-life between these two trends is at eight, whose normalized mood distribution is shown in Figure 10.6. This produces ratings that are self-normalizing, while still having a large enough window to prevent near-zero standard deviations when the original mood distribution is biased.

10.8 Conclusion

In this work, we investigate not only how to estimate mood from natural speech, but also how to make meaningful predictions for each subject. In order to accomplish this, we collected the PRIORI Annotated Mood dataset – a set of annotations indicating when an intervention was needed, based on a variety of clinical factors. I then framed the problem of intervention detection as anomaly detection using TempNorm to transform mood ratings into a more actionable measure. This framework allows us to measure mood abnormality in clinical and natural speech using emotion and transcript features. Across all experiments, we determined that a half-life of eight or 16 provided the best compromise between subject baseline stability and adaptation. These half-lives allowed the model to learn a baseline quickly enough so that the mood ratings were transformed to a unit normal distribution – balancing the dataset and increasing classification performance.

The results of this study could form the basis for an intervention-driven clinical trial. TempNorm detects a continuous rating of mood abnormality and allows for variable thresholds at a personal level. For example, individuals experiencing app fatigue could have their anomaly threshold raised, while individuals with elevated

mood instability may require a lower limit. Additionally, data from medical records may be used to more effectively initialize the system in a subject-specific manner. This would reflect actual clinical monitoring which takes into account any clinical history to better target interventions.

CHAPTER XI

Conclusions and Future Directions

In this dissertation, I have investigated methods of addressing variability in speech when automatically detecting emotion and mood. I have also introduced two applications of emotion and mood monitoring relying solely on natural, unstructured speech. This section highlights the key findings of these works and also proposes future work that further extends these applications.

11.1 Main Results and Contributions

Part I of the dissertation explored methods of reducing speech variability when classifying mood. In Chapter V, I explored how mood recognition from speech can be complicated when the speech is recorded on different types of devices. This is due to the quality of the microphone in different devices sometimes dramatically varying. I demonstrated how preprocessing techniques including declipping, normalization, and noise-robust segmentation could be used to lessen the impact of device variability. Additionally, I showed how multi-task learning could be employed to attain significantly better performance, compared with simply concatenating device data. In Chapter VI, I investigated how the previously introduced system trained across an entire cohort of individuals could be specialized using subject-specific modeling. In particular, I found that a soft weighted fusion between a population-general model

and a subject-specific model provided the best recognition of depressed speech. This allowed for subjects with more data to rely more heavily on a specialized model and attain better performance.

Part II of the dissertation investigated methods of reducing speech variability when classifying emotion. In Chapter VII, I investigated how to augment speech emotion classification with additional paralinguistic information (speaker an gender) and datasets. I used conducted the first speech emotion experiments using progressive neural networks and demonstrated how they improved on the standard pre-train/fine-tune method when testing in a subject-dependent manner. In Chapter VIII, I introduced MADDoG – a new approach for more generalized representation of emotion for cross-corpus testing. While previous related methods, such as DANNs [1, 5], have had issues converging, MADDoG follows a "meet in the middle" paradigm for consistent convergence. I then demonstrated that MADDoG was significantly better than traditional methods when combining differing amounts of labeling and unlabeled data from across different emotion datasets. In particular, I demonstrated how MADDoG was able to combine laboratory recordings with the newly introduced PRIORI Emotion dataset for improved in-the-wild speech emotion recognition.

Part III of the dissertation introduced two different real-world speech monitoring applications for individuals with mental disorders. Both applications used emotion features extracted using the MADDoG model from Chapter VIII, validating its ability to generate a robust representation of speech emotion. In Chapter IX, I demonstrated that MADDoG could be used to classify self-rated emotion in individuals with suicidal ideation. While most prior speech emotion work has required time-consuming annotation of data, this work solely relied upon these participant self-ratings. I then predicted the speech emotion of all phone calls in the EMASS dataset and found that individuals with suicidal ideation had a lower amount of emotion variability, compared with healthy controls. In Chapter X, I investigated how to create an ac-

tionable bipolar mood monitoring system using only natural conversations (PRIORI personal calls). I coordinated the collection of the PRAM dataset – an annotated subset of the PRIORI corpus that identifies the need for clinical interventions. I then examined how TempNorm could be used to convert mood symptom severity ratings to a continuous value representative of the abnormality of mood, taking into account each subject's baseline. I then combined TempNorm with a neural network, to detect anomalous mood using emotion and transcript-based features over unstructured, natural speech. This prediction of mood abnormality will enable future applications to better monitor bipolar mood and detect when intervention is needed.

11.2 Future Work

The work given in this dissertation aims to address the variability of speech to allow for real-world detection of emotion and mood. In particular, I accomplish this by either removing these factors of variability (Chapters V and VIII) or adapting using subject data (Chapters VI and X). Future work will expand on both of these approaches and combine techniques explored across this dissertation.

11.2.1 MADDoG for Multiple Modalities and Factors of Variability

The MADDoG algorithm introduced in Chapter VIII and further validated in Chapter IX demonstrated an emotion recognition system capable of working effectively across datasets. It accomplished this by building a representation of speech emotion based on MFB features that removed the variability due to differences in dataset. I propose future work that extends the MADDoG algorithm to include additional features sets and controls for multiple factors of variability simultaneously.

Additional features will be added to the MADDoG model using different types of fusion techniques. For example, other low-level descriptors shown effective in emotion classification, such as those in the eGeMAPS set [61], could be combined with MFBs.

Work will explore whether to use early or mid-fusion and whether to apply a separate critic for each feature set. The transcript-based features used in Chapter X and introduced by Matton et al. in [138] could also be employed. However, these features operate on call-level recordings instead of at the segment-level and will need to be adapted to the shorter length emotion ground truth time scale.

I also propose experiments investigating the use of MADDoG for controlling other types of speech variability – device, gender, noise, and speaker. Augmenting the system to control for device, gender, or noise variability would be relatively straightforward by adding an additional critic for each. Including environmental noise variability in the analysis would either require either using a dataset with different types of noise annotated or artificially introducing noise to the recordings. The critic for each type of variability would require one output for category considered. For example, the device variability critic would require three outputs if considering three different device types. However, a different approach would likely be needed for speaker variability because there are no clearly defined categories and the number of speakers is only limited by the size of the dataset. Instead, the critic training process would need to be changed to a pairwise method, as commonly used in speaker verification [97]. The training procedure would instead estimate and minimize the distance between paired subject representations. By explicitly reducing these other factors of variability, MADDoG should be able to produce even more robust representations of speech emotion capable of working on unseen data.

11.2.2 Extensions to Natural Speech Mood Monitoring

Chapter X introduced a speech mood monitoring system capable of detecting the need for interventions in everyday speech. However, my analysis required a reduction of the data in order to first focus on the higher quality samples (see Section 10.4.3). Future work will address additional factors of variability to increase the usable data

and make the system more versatile.

For example, I excluded healthy controls from this work to focus on mood in individuals with BPD. However, future work will include healthy controls by allowing for a separate set of parameters. These will presumably be within the normal range of human behavior and should not require interventions. Additionally, techniques will be employed to control for speech variability to allow for more devices and no longer limit the data to S5 recordings [74, 75]. This will include the declipping and normalization techniques explored in Chapter V, as well by using the improved MADDoG model for emotion features, as proposed in Section 11.2.1.

Future work will also investigate other feature sets and incorporate other modalities to ensure that lack of speech does not result in an inability to detect mood anomalies. To facilitate this, the newest version of the PRIORI app captures movement and location data, in addition to information representative of phone usage (battery and data). Finally, future work will explore causal models to take advantage of any data occurring between mood ratings, in addition to day-of features. While previous work has shown a lack of causality for bipolar mood, this was on a week-to-week basis and may not apply to daily mood dynamics [145]. These improvements will allow us to increase the scope of PRIORI and get closer to an application capable of real-world use.

11.3 Work Published

The work presented in this dissertation was published in the following papers:

- Part of Chapters III and V, in **John Gideon**, Emily Mower Provost, and Melvin McInnis. "Mood State Prediction from Speech of Varying Acoustic Quality for Individuals with Bipolar Disorder." *ICASSP*. 2016. (*oral presentation*)

- Part of Chapters III and VI, in Soheil Khorram, **John Gideon**, Melvin McIn-

nis, and Emily Mower Provost. "Recognition of Depression in Bipolar Disorder: Leveraging Cohort and Person-Specific Knowledge." *INTERSPEECH*. 2016. (*oral presentation*)

- Part of Chapters IV and VII, in **John Gideon**, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. "Progressive Neural Networks for Transfer Learning in Emotion Recognition." *INTERSPEECH*. 2017. (*oral presentation*)

- Part of Chapters I - IV, in Soheil Khorram, Mimansa Jaiswal, **John Gideon**, Melvin McInnis, Emily Mower Provost. "The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild." *INTERSPEECH*. 2018. (*oral presentation*)

- Part of Chapters I - IV and VIII, in **John Gideon**, Melvin McInnis, and Emily Mower Provost. "Barking up the Right Tree: Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDoG)." *IEEE Transactions on Affective Computing*. 2018. (*accepted*)

- Part of Chapters IV and IX, in **John Gideon**, Heather T Schatten, Melvin G McInnis, Emily Mower Provost. "Emotion Recognition from Natural Phone Conversations in Individuals With and Without Recent Suicidal Ideation." *INTERSPEECH*. 2019. (*poster presentation*)

- Part of Chapters II, III, X, and XI in **John Gideon**, Katie Matton, Melvin G McInnis, Emily Mower Provost. "When to Intervene: Detecting Abnormal Mood in Everyday Smartphone Conversations." *IEEE Transactions on Affective Computing*. 2019. (*in submission*)

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Mohammed Abdelwahab and Carlos Busso. "Domain Adversarial for Acoustic Emotion Recognition". In: *arXiv:1804.07690* (2018).

[2] Mohammed Abdelwahab and Carlos Busso. "Supervised domain adaptation for emotion recognition from speech". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 5058–5062.

[3] A. Aguilera and F. Muench. "There's an App for that: Information technology applications for cognitive behavioral practitioners". In: *The Behavior therapist/AABT* 35.4 (2012), p. 65.

[4] Adrian Aguilera, Stephen M Schueller, and Yan Leykin. "Daily mood ratings via text message as a proxy for clinic based depression assessment". In: *Journal of affective disorders* 175 (2015), pp. 471–474.

[5] Hana Ajakan et al. "Domain-adversarial neural networks". In: *arXiv:1412.4446* (2014).

[6] Martin Alda and Mirko Manchia. "Personalized management of bipolar disorder". In: *Neuroscience letters* 669 (2018), pp. 3–9.

[7] Zakaria Aldeneh and Emily Mower Provost. "Using regional saliency for speech emotion recognition". In: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE. 2017, pp. 2741–2745.

[8] Zakaria Aldeneh et al. "Pooling acoustic and lexical features for the prediction of valence". In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM. 2017, pp. 68–72.

[9] Sharifa Alghowinem et al. "Cross-Cultural Depression Recognition from Vocal Biomarkers." In: *Interspeech*. 2016, pp. 1943–1947.

[10] Sharifa Alghowinem et al. "Detecting depression: a comparison between spontaneous and read speech". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 7547–7551.

[11] S. Alghowinem et al. "A comparative study of different classifiers for detecting depression from spontaneous speech". In: *Proceeding from the IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2013, pp. 8022–8026.

[12] S. Alghowinem et al. "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech." In: *FLAIRS Conference*. 2012.

[13] Jinwon An and Sungzoon Cho. "Variational autoencoder based anomaly detection using reconstruction probability". In: *Special Lecture on IE* 2 (2015).

[14] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011". In: *Artificial Intelligence Review* 43.2 (2015), pp. 155–177.

[15] J. Angst, R. Sellaro, and F. Angst. "Long-term outcome and mortality of treated versus untreated bipolar and depressed patients: a preliminary report". In: *International Journal of Psychiatry in Clinical Practice* 2.2 (1998), pp. 115–119.

[16] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein gan". In: *arXiv:1701.07875* (2017).

[17] Michael F Armey et al. "Ecologically assessed affect and suicidal ideation following psychiatric inpatient hospitalization." In: *General hospital psychiatry* (2018).

[18] American Psychiatric Association et al. *DSM 5*. American Psychiatric Association, 2013.

[19] Douglas Bates et al. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01.

[20] RH Belmaker. "Bipolar disorder". In: *New England Journal of Medicine* 351.5 (2004), pp. 476–486.

[21] Mohamed Benzeghiba et al. "Automatic speech recognition and speech variability: A review". In: *Speech communication* 49.10-11 (2007), pp. 763–786.

[22] Jarosław Bernacki and Grzegorz Kołaczek. "Anomaly detection in network traffic using selected methods of time series analysis". In: *IJCNIS* 7.9 (2015), p. 10.

[23] Amelia J Birney et al. "MoodHacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial". In: *JMIR mHealth and uHealth* 4.1 (2016).

[24] William V Bobo et al. "A randomized, open comparison of long-acting injectable risperidone and treatment as usual for prevention of relapse, rehospi-

talization and urgent care referral in community-treated patients with rapid cycling bipolar disorder". In: *Clinical neuropharmacology* 34.6 (2011), p. 224.

[25] Tobias Bocklet et al. "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines". In: *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE. 2008, pp. 1605–1608.

[26] Niall Bolger, Angelina Davis, and Eshkol Rafaeli. "Diary methods: Capturing life as it is lived". In: *Annual review of psychology* 54.1 (2003), pp. 579–616.

[27] Remco R Bouckaert and Eibe Frank. "Evaluating the replicability of significance tests for comparing learning algorithms". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2004, pp. 3–12.

[28] Margaret M Bradley and Peter J Lang. "Measuring emotion: the self-assessment manikin and the semantic differential". In: *Journal of behavior therapy and experimental psychiatry* 25.1 (1994), pp. 49–59.

[29] Katie A Busch and Jan Fawcett. "A fine-grained study of inpatients who commit suicide". In: *Psychiatric Annals* 34.5 (2004), pp. 357–364.

[30] Carlos Busso and Shrikanth S Narayanan. "Interrelation between speech and facial gestures in emotional utterances: a single subject study". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.8 (2007), pp. 2331–2347.

[31] Carlos Busso et al. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language resources and evaluation* 42.4 (2008), p. 335.

[32] Carlos Busso et al. "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception". In: *IEEE Transactions on Affective Computing* (2016).

[33] Carlos Busso et al. "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception". In: *IEEE Transactions on Affective Computing* 1 (2017), pp. 67–80.

[34] Carlos Busso et al. "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception". In: *IEEE Transactions on Affective Computing* 8.1 (2017), pp. 67–80.

[35] M. Cannizzaro et al. "Voice acoustical measurement of the severity of major depression". In: *Brain and cognition* 56.1 (2004), pp. 30–35.

[36] Facundo Carrillo et al. "Automated speech analysis for psychosis evaluation". In: *MLINI*. Springer, 2013, pp. 31–39.

[37] Facundo Carrillo et al. “Emotional intensity analysis in Bipolar subjects”. In: *arXiv:1606.02231* (2016).

[38] Centers for Disease Control and Prevention NCfIPaC. *Web-based Injury Statistics Query and Reporting System (WISQARS)*. [online] Available from URL: `www.cdc.gov/injury/wisqars`.

[39] Varun Chandola, Arindam Banerjee, and Vipin Kumar. “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3 (2009), p. 15.

[40] Jonathan Chang and Stefan Scherer. “Learning representations of emotional speech with deep convolutional generative adversarial networks”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE. 2017, pp. 2746–2750.

[41] Chi-Hua Chen et al. “Explicit and implicit facial affect recognition in manic and depressed states of bipolar disorder: a functional magnetic resonance imaging study”. In: *Biological Psychiatry* 59.1 (2006), pp. 31–39.

[42] Christopher Cieri, David Miller, and Kevin Walker. “The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text.” In: *LREC*. Vol. 4. 2004, pp. 69–71.

[43] J. Cohn et al. “Detecting depression from facial actions and vocal prosody”. In: *Proceedings from the International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*. IEEE. 2009, pp. 1–7.

[44] Glen Coppersmith, Mark Dredze, and Craig Harman. “Quantifying mental health signals in Twitter”. In: *CLPsych*. 2014, pp. 51–60.

[45] C. Cortes and V. Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297.

[46] Nicholas Cummins et al. “A review of depression and suicide risk assessment using speech analysis”. In: *Speech Communication* 71 (2015), pp. 10–49.

[47] Nicholas Cummins et al. “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, pp. 970–974.

[48] N. Cummins et al. “An Investigation of Depressed Speech Detection: Features and Normalization.” In: *Proceedings from the Conference of the International Speech Communication Association (Interspeech)*. 2011, pp. 2997–3000.

[49] N. Cummins et al. “Modeling spectral variability for the classification of depressed speech.” In: *Proceedings from the Conference of the International Speech Communication Association (Interspeech)*. 2013, pp. 857–861.

[50] Amit Das and Mark Hasegawa-Johnson. "Cross-lingual transfer learning during supervised training in low resource scenarios." In: *Interspeech*. 2015, pp. 3531–3535.

[51] Munmun De Choudhury et al. "Predicting depression via social media". In: *AAAI*. 2013.

[52] Najim Dehak et al. "Front-end factor analysis for speaker verification". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 19.4 (2011), pp. 788–798.

[53] Jun Deng, Zixing Zhang, and Björn Schuller. "Linked Source and Target Domain Subspace Feature Transfer Learning–Exemplified by Speech Emotion Recognition". In: *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE. 2014, pp. 761–766.

[54] Jun Deng et al. "Autoencoder-based unsupervised domain adaptation for speech emotion recognition". In: *IEEE Signal Processing Letters* 21.9 (2014), pp. 1068–1072.

[55] Jun Deng et al. "Sparse autoencoder-based feature transfer learning for speech emotion recognition". In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE. 2013, pp. 511–516.

[56] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv:1702.08608* (2017).

[57] Sefik Emre Eskimez, Zhiyao Duan, and Wendi Heinzelman. "Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5099–5103.

[58] T. Evgeniou and M. Pontil. "Regularized multi–task learning". In: *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM. 2004, pp. 109–117.

[59] 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*. ACM. 2010, pp. 1459–1462.

[60] Florian Eyben, Martin Wöllmer, and Björn Schuller. "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit". In: *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on*. IEEE. 2009, pp. 1–6.

[61] Florian Eyben et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE Transactions on Affective Computing* 7.2 (2016), pp. 190–202.

[62] Scott E Fahlman and Christian Lebiere. "The cascade-correlation learning architecture". In: (1990).

[63] Maria Faurholt-Jepsen et al. "Daily electronic monitoring of subjective and objective measures of illness activity in bipolar disorder using smartphones–the MONARCA II trial protocol: a randomized controlled single-blind parallel-group trial". In: *BMC psychiatry* 14.1 (2014), p. 309.

[64] Maria Faurholt-Jepsen et al. "The effect of smartphone-based monitoring on illness activity in bipolar disorder: the MONARCA II randomized controlled single-blinded trial". In: *Psychological medicine* (2019), pp. 1–11.

[65] M Faurholt-Jepsen et al. "Voice analysis as an objective state marker in bipolar disorder". In: *Translational psychiatry* 6.7 (2016), e856.

[66] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. "On the Correlation and Transferability of Features Between Automatic Speech Recognition and Speech Emotion Recognition." In: *Interspeech*. 2016, pp. 3618–3622.

[67] Luciana Ferrer, M Kemal Sönmez, and Elizabeth Shriberg. "An anticorrelation kernel for improved system combination in speaker verification." In: *Odyssey*. Citeseer. 2008, p. 22.

[68] D. France et al. "Acoustical properties of speech as indicators of depression and suicidal risk". In: *IEEE Transactions on Biomedical Engineering* 47.7 (2000), pp. 829–837.

[69] Joseph C Franklin et al. "Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research." In: *Psychological Bulletin* 143.2 (2017), p. 187.

[70] E. Friedman and G. Sanders. "Speech timing of mood disorders". In: *Computers in Human Services* 8.3-4 (1991), pp. 121–142.

[71] Panayiotis G Georgiou, Matthew P Black, and Shrikanth S Narayanan. "Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments". In: *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. ACM. 2011, pp. 7–12.

[72] S. N. Ghaemi, A. L. Stoll, and H. G. Pope Jr. "Lack of insight in bipolar disorder the acute manic episode". In: *The Journal of Nervous and Mental Disease* 183.7 (1995), pp. 464–467.

[73] John Gideon, Melvin G McInnis, and Emily Mower Provost. "Emotion Recognition from Natural Phone Conversations in Individuals With and Without Recent Suicidal Ideation." In: *Interspeech*. 2019.

[74] John Gideon, Melvin McInnis, and Emily Mower Provost. "Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDoG)". In: *IEEE Transactions on Affective Computing* (2019).

[75] John Gideon, Emily Mower Provost, and Melvin McInnis. "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder". In: *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE. 2016, pp. 2359–2363.

[76] John Gideon et al. "Progressive neural networks for transfer learning in emotion recognition". In: *Interspeech 2017* (2017).

[77] John Gideon et al. "When to Intervene: Detecting Abnormal Mood in Everyday Smartphone Conversations". In: *IEEE Transactions on Affective Computing* (2019). (*in submission*).

[78] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011, pp. 315–323.

[79] Prasanta Gogoi, Bhogeswar Borah, and Dhruba K Bhattacharyya. "Anomaly detection analysis of intrusion data using supervised & unsupervised approach". In: *Journal of Convergence Information Technology* 5.1 (2010).

[80] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[81] F. K. Goodwin and K. R. Jamison. *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression*. Oxford University Press, 2007.

[82] E. Granholm et al. "Mobile Assessment and Treatment for Schizophrenia (MATS): a pilot trial of an interactive text-messaging intervention for medication adherence, socialization, and auditory hallucinations". In: *Schizophrenia bulletin* (2011), sbr155.

[83] Melissa J Green, Catherine M Cahill, and Gin S Malhi. "The cognitive and neurophysiological basis of emotion dysregulation in bipolar disorder". In: *Journal of affective disorders* 103.1-3 (2007), pp. 29–42.

[84] A Grunerbl et al. "Smart-phone based recognition of states and state changes in bipolar disorder patients". In: (2014).

[85] A. Grunerbl et al. "Smartphone-based recognition of States and state changes in bipolar disorder patients". In: *IEEE Journal of Biomedical and Health Informatics* 19.1 (2015), pp. 140–148.

[86] Arthur Guez et al. “Adaptive Treatment of Epilepsy via Batch-mode Reinforcement Learning.” In: *AAAI*. 2008, pp. 1671–1678.

[87] A. Guidi et al. “Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients”. In: *Biomedical Signal Processing and Control* (2014).

[88] Nikou Günnemann, Stephan Günnemann, and Christos Faloutsos. “Robust multivariate autoregression for anomaly detection in dynamic product ratings”. In: *WWW*. ACM. 2014, pp. 361–372.

[89] Ruben C Gur et al. “Facial emotion discrimination: II. Behavioral findings in depression”. In: *Psychiatry research* 42.3 (1992), pp. 241–251.

[90] M. Hamilton. “Hamilton depression scale”. In: *ECDEU Assessment Manual For Psychopharmacology, Revised Edition. Rockville, MD: National Institute of Mental Health* (1976), pp. 179–92.

[91] Kun Han, Dong Yu, and Ivan Tashev. “Speech emotion recognition using deep neural network and extreme learning machine”. In: *Fifteenth annual conference of the international speech communication association*. 2014.

[92] M. J. Harvilla and R. M. Stern. “Least squares signal declipping for robust speech recognition”. In: *Proceedings from the Conference of the International Speech Communication Association (Interspeech)*. 2014.

[93] Mark J Harvilla and Richard M Stern. “Efficient audio declipping using regularized least squares”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 221–225.

[94] Ali Hassan, Robert Damper, and Mahesan Niranjan. “On acoustic emotion recognition: compensating for covariate shift”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.7 (2013), pp. 1458–1468.

[95] Andrew O Hatch, Sachin S Kajarekar, and Andreas Stolcke. “Within-class covariance normalization for SVM-based speaker recognition.” In: *Interspeech*. 2006.

[96] Xiaofei He and Partha Niyogi. “Locality preserving projections”. In: *Advances in neural information processing systems*. 2004, pp. 153–160.

[97] Georg Heigold et al. “End-to-end text-dependent speaker verification”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 5115–5119.

[98] Brian S Helfer et al. “Classification of depression state based on articulatory precision”. In: *Interspeech*. 2013, pp. 2172–2176.

[99] Jordan Hochenbaum, Owen S Vallis, and Arun Kejariwal. "Automatic anomaly detection in the cloud via statistical learning". In: *arXiv:1704.07706* (2017).

[100] R. Hoffman, S. Stopek, and N. Andreasen. "A comparative study of manic vs schizophrenic speech disorganization". In: *Archives of General Psychiatry* 43.9 (1986), pp. 831–838.

[101] Hao Hu, Ming-Xing Xu, and Wei Wu. "GMM supervector based SVM with spectral features for speech emotion recognition". In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. IEEE. 2007, pp. IV–413.

[102] Kun-Yi Huang, Chung-Hsien Wu, and Ming-Hsiang Su. "Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses". In: *Pattern Recognition* 88 (2019), pp. 668–678.

[103] N. E. Huang et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis". In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 454.1971 (1998), pp. 903–995.

[104] Zhaocheng Huang, Julien Epps, and Dale Joachim. "Speech Landmark Bigrams for Depression Detection from Naturalistic Smartphone Speech". In: *ICASSP*. IEEE. 2019, pp. 5856–5860.

[105] Zhaocheng Huang et al. "Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions." In: *Interspeech*. 2018, pp. 3393–3397.

[106] Zhengwei Huang et al. "Speech emotion recognition using CNN". In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 801–804.

[107] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[108] Erkki T Isometsa et al. "The last appointment before suicide: is suicide intent communicated?" In: *The American journal of psychiatry* 152.6 (1995), p. 919.

[109] Farid Kadri et al. "Seasonal ARMA-based SPC charts for anomaly detection: Application to emergency department systems". In: *Neurocomputing* 173 (2016), pp. 2102–2114.

[110] Takuhiro Kaneko and Hirokazu Kameoka. "Parallel-data-free voice conversion using cycle-consistent adversarial networks". In: *arXiv:1711.11293* (2017).

[111] Zahi N Karam et al. "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech". In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 4858–4862.

[112] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel. "Eigenvoice modeling with sparse training data". In: *Speech and Audio Processing, IEEE Transactions on* 13.3 (2005), pp. 345–354.

[113] Soheil Khorram et al. "Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition". In: *arXiv:1708.07050* (2017).

[114] Soheil Khorram et al. "Recognition of Depression in Bipolar Disorder: Leveraging Cohort and Person-Specific Knowledge." In: *Interspeech*. 2016, pp. 1215–1219.

[115] Soheil Khorram et al. "The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild". In: *Interspeech 2018* (2018), pp. 1903–1907.

[116] Jaebok Kim et al. "Towards Speech Emotion Recognition" in the wild" using Aggregated Corpora and Deep Multi-Task Learning". In: *arXiv:1708.03920* (2017).

[117] Yelin Kim, Honglak Lee, and Emily Mower Provost. "Deep learning for robust feature generation in audiovisual emotion recognition". In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, pp. 3687–3691.

[118] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv:1412.6980* (2014).

[119] Matthieu Komorowski et al. "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care". In: *Nat. Med.* 24.11 (2018), p. 1716.

[120] Amy Kranzler et al. "Emotional dysregulation, internalizing symptoms, and self-injurious and suicidal behavior: Structural equation modeling analysis". In: *Death studies* 40.6 (2016), pp. 358–366.

[121] S. A. Langenecker et al. "Intermediate: cognitive phenotypes in bipolar disorder". In: *Journal of Affective Disorders* 122.3 (2010), pp. 285–293.

[122] Mark E Larsen et al. "The use of technology in suicide prevention". In: *2015 37th annual international conference of the IEEE engineering in Medicine and biology society (EMBC)*. IEEE. 2015, pp. 7316–7319.

[123] Siddique Latif et al. "Variational Autoencoders for Learning Latent Representations of Speech Emotion". In: *arXiv:1712.08708* (2017).

[124] Keyne C Law, Lauren R Khazem, and Michael D Anestis. "The role of emotion dysregulation in suicide as considered through the ideation to action framework". In: *Current Opinion in Psychology* 3 (2015), pp. 30–35.

[125] Chul Min Lee and Shrikanth S Narayanan. "Toward detecting emotions in spoken dialogs". In: *IEEE transactions on speech and audio processing* 13.2 (2005), pp. 293–303.

[126] Russell Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means.* R package version 1.3.5.1. 2019. URL: `https://CRAN.R-project.org/package=emmeans`.

[127] Haoliang Li et al. "Domain generalization with adversarial feature learning". In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*. 2018.

[128] R. LiKamWa et al. "Moodscope: building a mood sensor from smartphone usage patterns". In: *ACM International Conference on Mobile Systems, Applications, and Services*. ACM. 2013, pp. 389–402.

[129] Yi-Lin Lin and Gang Wei. "Speech emotion recognition based on HMM and SVM". In: *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*. Vol. 8. IEEE. 2005, pp. 4898–4901.

[130] Paul S Links et al. "Affective instability and suicidal ideation and behavior in patients with borderline personality disorder". In: *Journal of personality disorders* 21.1 (2007), pp. 72–86.

[131] Gang Liu and John HL Hansen. "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios". In: *Audio, Speech, and Language Processing, IEEE/ACM Transactions on* 22.12 (2014), pp. 1978–1992.

[132] Alan D Lopez et al. "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data". In: *The Lancet* 367.9524 (2006), pp. 1747–1757.

[133] Paula Lopez-Otero, Laura Dacia-Fernandez, and Carmen Garcia-Mateo. "A study of acoustic features for depression detection". In: *Biometrics and Forensics (IWBF), 2014 International Workshop on*. IEEE. 2014, pp. 1–6.

[134] H. Lu et al. "StressSense: Detecting stress in unconstrained acoustic environments using smartphones". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM. 2012, pp. 351–360.

[135] D. D. Luxton et al. "mHealth for mental health: Integrating smartphone technology in behavioral healthcare." In: *Professional Psychology: Research and Practice* 42.6 (2011), p. 505.

[136] Pankaj Malhotra et al. "Long short term memory networks for anomaly detection in time series". In: *Proceedings*. Presses universitaires de Louvain. 2015, p. 89.

[137] Soroosh Mariooryad and Carlos Busso. "Compensating for speaker or lexical variabilities in speech for emotion recognition". In: *Speech Communication* 57 (2014), pp. 1–12.

[138] Katie Matton, Melvin G McInnis, and Emily Mower Provost. "Into the Wild: Transitioning from Recognizing Mood in Clinical Interactions to Personal Conversations for Individuals with Bipolar Disorder". In: *Interspeech*. 2019.

[139] Melvin G McInnis et al. "Cohort Profile: the Heinz C. Prechter longitudinal study of bipolar disorder". In: *International journal of epidemiology* 47.1 (2018), p. 28.

[140] Kathleen R Merikangas et al. "Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative". In: *Archives of general psychiatry* 68.3 (2011), pp. 241–251.

[141] David J Miklowitz and Michael J Gitlin. *Clinician's Guide to Bipolar Disorder*. Guilford Publications, 2015.

[142] Vikramjit Mitra et al. "The SRI AVEC-2014 evaluation system". In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM. 2014, pp. 93–101.

[143] H Zare Moayedi and MA Masnadi-Shirazi. "Arima model for network traffic prediction and anomaly detection". In: *2008 International Symposium on Information Technology*. Vol. 4. IEEE. 2008, pp. 1–6.

[144] E. Moore et al. "Critical analysis of the impact of glottal features in the classification of clinical depression in speech". In: *IEEE Transactions on Biomedical Engineering* 55.1 (2008), pp. 96–107.

[145] Paul J Moore et al. "Mood dynamics in bipolar disorder". In: *International journal of bipolar disorders* 2.1 (2014), p. 11.

[146] Richard Morriss et al. "Interventions for helping people recognise early signs of recurrence in bipolar disorder". In: *The Cochrane Library* (2007).

[147] Natalia B Mota et al. "Speech graphs provide a quantitative measure of thought disorder in psychosis". In: *PloS one* 7.4 (2012), e34928.

[148] Lili Mou et al. "How Transferable are Neural Networks in NLP Applications?" In: *arXiv:1603.06111* (2016).

[149] Amir Muaremi et al. "Assessing bipolar episodes using speech cues derived from phone calls". In: *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer. 2014, pp. 103–114.

[150] J. Mundt et al. "Vocal acoustic biomarkers of depression severity and treatment response". In: *Biological psychiatry* 72.7 (2012), pp. 580–587.

[151] J. Mundt et al. "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology". In: *Journal of neurolinguistics* 20.1 (2007), pp. 50–64.

[152] FC Murphy et al. "Emotional bias and inhibitory control processes in mania and depression". In: *Psychological medicine* 29.6 (1999), pp. 1307–1321.

[153] Masafumi Nakano, Akihiko Takahashi, and Soichiro Takahashi. "Generalized exponential moving average (EMA) model with particle filtering and anomaly detection". In: *Expert Systems with App.* 73 (2017), pp. 187–200.

[154] National Institute of Mental Health. *Bipolar Disorder in Adults*. `http://www.nimh.nih.gov/health/publications/bipolar-disorder-in-adults/Bipolar_Disorder_Adults_CL508_144295.pdf`. Accessed: September - 2015.

[155] Hong-Wei Ng et al. "Deep learning for emotion recognition on small datasets using transfer learning". In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM. 2015, pp. 443–449.

[156] V. Osmani et al. "Monitoring activity of patients with bipolar disorder using smart phones". In: *Proceedings of the International Conference on Advances in Mobile Computing & Multimedia*. ACM. 2013, p. 85.

[157] Sinno Jialin Pan, Qiang Yang, et al. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2010), pp. 1345–1359.

[158] Yixiong Pan, Peipei Shen, and Liping Shen. "Speech emotion recognition using support vector machine". In: *International Journal of Smart Home* 6.2 (2012), pp. 101–108.

[159] Zhongde Pan et al. "Detecting manic state of bipolar disorder based on support vector machine and Gaussian mixture model using spontaneous speech". In: *Psychiatry investigation* 15.7 (2018), p. 695.

[160] *Combining Kernel and Model Based Learning for HIV Therapy Selection*. 2016.

[161] Srinivas Parthasarathy and Carlos Busso. "Jointly predicting arousal, valence and dominance with multi-task learning". In: *Interspeech, Stockholm, Sweden* (2017).

[162] Adam Paszke et al. “Automatic differentiation in pytorch”. In: (2017).

[163] Jason Pelecanos and Sridha Sridharan. “Feature warping for robust speaker verification”. In: *2001: A Speaker Odyssey. The Speaker Recognition Workshop*. International Speech Communication Association (ISCA), 2001.

[164] James Pennebaker et al. *The Development and Psychometric Properties of LIWC2007*. 2007.

[165] Brandon Pincombe. “Anomaly detection in time series of graphs using arma processes”. In: *Asor Bulletin* 24.4 (2005), p. 2.

[166] M. Pogue-Geile and T. Oltmanns. “Sentence perception and distractibility in schizophrenic, manic, and depressed patients.” In: *Journal of Abnormal Psychology* 89.2 (1980), p. 115.

[167] Daniel Povey et al. *The Kaldi speech recognition toolkit*. Tech. rep. IEEE Sig. Proc. Soc., 2011.

[168] Emily Mower Provost. “Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, pp. 3682–3686.

[169] T. F. Quatieri and N. Malyska. “Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity.” In: *Proceedings from the Conference of the International Speech Communication Association (Interspeech)*. 2012.

[170] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: `http://www.R-project.org/`.

[171] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: `http://www.R-project.org/`.

[172] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks”. In: *arXiv:1511.06434* (2015).

[173] Srinivasan Ramakrishnan. “Recognition of emotion from speech: A review”. In: *Speech Enhancement, Modeling and recognition–algorithms and Applications* 7 (2012), pp. 121–137.

[174] H. Resnick and T. Oltmanns. “Hesitation patterns in the speech of thought-disordered schizophrenic and manic patients.” In: *Journal of Abnormal Psychology* 93.1 (1984), p. 80.

[175] James A Russell. "Core affect and the psychological construction of emotion." In: *Psychological review* 110.1 (2003), p. 145.

[176] Andrei A Rusu et al. "Progressive neural networks". In: *arXiv:1606.04671* (2016).

[177] Andrei A Rusu et al. "Sim-to-real robot learning from pixels with progressive nets". In: *arXiv:1610.04286* (2016).

[178] S. O. Sadjadi and J. Hansen. "Unsupervised speech activity detection using voicing measures and perceptual spectral flux". In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 197–200.

[179] Saurabh Sahu, Rahul Gupta, and Carol Espy-Wilson. "On Enhancing Speech Emotion Recognition using Generative Adversarial Networks". In: *arXiv:1806.06626* (2018).

[180] Saurabh Sahu et al. "Adversarial Auto-encoders for Speech Based Emotion Recognition". In: *arXiv:1806.02146* (2018).

[181] Stefan Scherer et al. "Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 4789–4793.

[182] S. Scherer et al. "Investigating voice quality as a speaker-independent indicator of depression and PTSD." In: *Proceedings from the Conference of the International Speech Communication Association (Interspeech)*. 2013, pp. 847–851.

[183] Björn W Schuller, Stefan Steidl, Anton Batliner, et al. "The Interspeech 2009 emotion challenge." In: *Interspeech*. Vol. 2009. 2009, pp. 312–315.

[184] Björn Schuller, Gerhard Rigoll, and Manfred Lang. "Hidden Markov model-based speech emotion recognition". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. Vol. 2. IEEE. 2003, pp. II–1.

[185] Bjorn Schuller et al. "Cross-corpus acoustic emotion recognition: Variances and strategies". In: *IEEE Transactions on Affective Computing* 1.2 (2010), pp. 119–131.

[186] Björn Schuller et al. "Paralinguistics in speech and language—State-of-the-art and the challenge". In: *Computer Speech & Language* 27.1 (2013), pp. 4–39.

[187] Björn Schuller et al. "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization". In: *Proc. Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel*. Citeseer. 2011.

[188] Björn Schuller et al. "Using multiple databases for training in emotion recognition: To unite or to vote?" In: *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[189] Mohammed Senoussaoui et al. "Model fusion for multimodal depression classification and level detection". In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM. 2014, pp. 57–63.

[190] Ryan J Shaw et al. "Mobile health devices: will patients actually use them?" In: *J. Am. Med. Inform. Assoc.* 23.3 (2016), pp. 462–466.

[191] Maxim Sidorov, Stefan Ultes, and Alexander Schmitt. "Comparison of Gender- and Speaker-adaptive Emotion Recognition." In: *LREC*. 2014, pp. 3476–3480.

[192] M. Śmieja. "Weighted approach to general entropy function". In: *IMA Journal of Mathematical Control and Information* (2014), dnt044.

[193] A. Smith. "Smartphone ownership–2013 update". In: *Pew Research Center: Washington DC* (2013).

[194] Peng Song. "Transfer linear subspace learning for cross-corpus speech emotion recognition". In: *IEEE Transactions on Affective Computing* (2017).

[195] Peng Song et al. "Speech emotion recognition using transfer learning". In: *IEICE TRANSACTIONS on Information and Systems* 97.9 (2014), pp. 2530–2532.

[196] *SoX, Sound eXchange (v14.4.1)*. `http://sox.sourceforge.net/`.

[197] Brian Stasak et al. "An Investigation of Emotional Speech in Depression Classification." In: *Interspeech*. 2016, pp. 485–489.

[198] Emelyn Sue Qing Tan and Yuen Jien Soo. "Creating Apps: A Non-IT Educator's Journey Within a Higher Education Landscape". In: *Mobile Learning in Higher Education in the Asia-Pacific Region*. Springer, 2017, pp. 213–238.

[199] M. Taylor, R. Reed, and S. Berenbaum. "Patterns of speech disorders in schizophrenia and mania." In: *The Journal of Nervous and Mental Disease* 182.6 (1994), pp. 319–326.

[200] P. Thomas et al. "Speech and language in first onset psychosis differences between people with schizophrenia, mania, and controls." In: *The British Journal of Psychiatry* 168.3 (1996), pp. 337–343.

[201] S. Tilsen and A. Arvaniti. "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages". In: *The Journal of the Acoustical Society of America* 134.1 (2013), pp. 628–639.

[202] George Trigeorgis et al. “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network”. In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2016, pp. 5200–5204.

[203] Eric Tzeng et al. “Adversarial discriminative domain adaptation”. In: *Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2. 2017, p. 4.

[204] Michel Valstar et al. “Avec 2014: 3d dimensional affect and depression recognition challenge”. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM. 2014, pp. 3–10.

[205] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9.2579-2605 (2008), p. 85.

[206] N. Vanello et al. “Speech analysis for mood state characterization in bipolar patients”. In: *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2012, pp. 2104–2107.

[207] Dimitrios Ververidis and Constantine Kotropoulos. “Automatic speech classification to five emotional states based on gender information”. In: *Signal Processing Conference, 2004 12th European*. IEEE. 2004, pp. 341–344.

[208] Olli Viikki and Kari Laurila. “Cepstral domain segmental feature vector normalization for noise robust speech recognition”. In: *Speech Communication* 25.1 (1998), pp. 133–147.

[209] Thurid Vogt and Elisabeth André. “Improving automatic emotion recognition from speech via gender differentiation”. In: *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*. 2006.

[210] David Watson and Lee Anna Clark. *The PANAS-X: Manual for the positive and negative affect schedule-expanded form*. 1999.

[211] J. Williamson et al. “Vocal biomarkers of depression based on motor incoordination”. In: *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM. 2013, pp. 41–48.

[212] Tracy K Witte et al. “Naturalistic evaluation of suicidal ideation: variability and relation to attempt status”. In: *Behaviour Research and Therapy* 44.7 (2006), pp. 1029–1040.

[213] Tracy K Witte et al. “Variability in suicidal ideation: a better predictor of suicide attempts than intensity or duration of ideation?” In: *Journal of affective disorders* 88.2 (2005), pp. 131–136.

[214] Rui Xia and Yang Liu. "A multi-task learning framework for emotion recognition using 2D continuous space". In: *IEEE Transactions on Affective Computing* (2015).

[215] Bing Xu et al. "Empirical evaluation of rectified activations in convolutional network". In: *arXiv:1505.00853* (2015).

[216] Le Yang et al. "Multimodal measurement of depression using deep learning models". In: *AVEC*. ACM. 2017, pp. 53–59.

[217] Shirley Yen et al. "Borderline personality disorder criteria associated with prospectively observed suicidal behavior". In: *American Journal of Psychiatry* 161.7 (2004), pp. 1296–1298.

[218] R. Young et al. "A rating scale for mania: reliability, validity and sensitivity." In: *The British Journal of Psychiatry* 133.5 (1978), pp. 429–435.

[219] Houssam Zenati et al. "Efficient gan-based anomaly detection". In: *arXiv:1802.06222* (2018).

[220] Biqiao Zhang, Georg Essl, and Emily Mower Provost. "Predicting the distribution of emotion perception: capturing inter-rater variability". In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM. 2017, pp. 51–59.

[221] Biqiao Zhang, Soheil Khorram, and Emily Mower Provost. "Exploiting Acoustic and Lexical Properties of Phonemes to Recognize Valence from Speech". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 5871–5875.

[222] Biqiao Zhang, Emily Mower Provost, and Georg Essi. "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach". In: *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE. 2016, pp. 5805–5809.

[223] Jing Zhang et al. "Analysis on speech signal features of manic patients". In: *Journal of psychiatric research* 98 (2018), pp. 59–63.

[224] Zixing Zhang et al. "Unsupervised learning in cross-corpus acoustic emotion recognition". In: *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE. 2011, pp. 523–528.

[225] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *arXiv preprint* (2017).

[226] Xinyue Zhu et al. "Emotion Classification with Data Augmentation Using Generative Adversarial Networks". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2018, pp. 349–360.

[227] Yuan Zong et al. “Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis”. In: *Journal on Multimodal User Interfaces* 10.2 (2016), pp. 163–172.