Automatic Analysis of Psychotherapy Sessions

Dalia Attas

Supervisors: Prof Heidi Christensen and Dr Chris Blackmore

Department of Computer Science
The University of Sheffield

This dissertation is submitted for the degree of
Doctor of Philosophy

Speech and Hearing Research Group June 2022

I would like to dedicate this thesis to my husband and children.

Declaration

I hereby declare that this dissertation is of my own work, except where I specifically referred to the works done by the other authors in the text. The contents of the study are original and have not been submitted for any other awards, qualifications, or degrees in universities. Parts of the findings of this study have already been published or planned to publish as a journal or a conference papers.

Dalia Attas
June 2022

Acknowledgements

First and foremost, I would like to thank Allah, without whom nothing is possible.

I would like to express my sincere graduate to my supervisors, Prof Heidi Christensen and Dr Chris Blackmore, for their guidance and support throughout this entire journey. I wish to extend my special thanks to Prof Heidi for her endless kindness and invaluable support.

Furthermore, I would also like to thank my panel members, Dr Nikolaos Aletras and Dr Mauricio Alvarez. It has been an honour to work with Dr Stephan Kellet, Niall Power and everyone who contributed to the work of recording the psychotherapy sessions.

I cannot begin to express my thanks to my parents, who have been supportive and encouraging through this journey. I would like to extend my deepest gratitude to my sisters, who have been my best friends and for being always there for me.

This project would not have been possible without my husband's support, patience, understanding and encouragement. I'm extremely grateful to have such lovely children who filled this journey with smiles and consolation.

I would like to pay special appreciation to my colleagues, Dr Bahman Mirheidari, Dr Lubna Alhinti, Dr Rabab Algadhy, Fatimah Alzahrani, Gerardo Roa Dabike, Jack Deadman, Jisi Zhang, Lucy Skidmore, Dr Mashael AlSaleh, Megan Thomas, Yilin Pan, Zehai Tu, Zhengjun Yue, Nathan Pevy, Samuel Hollands, Wanli Sun and Juan Jose Giraldo. I am very thankful to my friends in this journey, Dr Areej Alokaili, Dr Amal Alharbi and Tarfah Alrashid.

Finally, I thankfully acknowledge the financial funding for this work from Umm-Alqura University and the Saudi Ministry of Education.

Abstract

Psychotherapy is known to be beneficial to people struggling with mental health disorders. It can assist them in regaining their normal life and recovery. Psychotherapy is considered one of the most common therapies that are emerging nowadays. According to the National Health Service (NHS), 1.46 million people were referred to psychotherapy in 2020-21. Therapists diagnose and treat the patients' mental illness through a defined number of conversational sessions to discuss and negotiate the patients' feelings. Premature therapy termination or patient's drop-out of treatment is one of the most critical problems in therapy, leading to poor treatment outcomes. It has been found that patients with a weaker therapeutic relationship *therapeutic alliance* with their therapists have more drop-out cases. This could relate to the therapist's judged **competence**. This project aims to investigate the feasibility of automatically analysing psychotherapy sessions to improve the therapist's work by detecting and tracking positive signs of therapeutic alliance and thereby helping to minimise the number of patient drop-out cases. Several patient and therapist behaviours, expressed during sessions, could reveal a positive therapeutic alliance and competence such as the emotional state, mood, synchrony, and empathy. Those behaviours have been investigated according to their impact on acoustic and linguistic cues in the speakers' speech. The main contributions of this project are developing an automatic system for analysing psychotherapy sessions using actual patient-therapist conversations. This has encompassed successfully implementing an automatic speech recognition system, developing an automatic system for detecting a patient's mood (depression and anxiety), developing an automatic system for predicting a therapist's competence score, and developing an automatic time-continuous recognition for a patient's emotional speech. The results indicate that it is feasible to build a successful automatic system for analysing psychotherapy sessions, considering different patient's and therapist's behaviours detected or tracked within a single treatment session.

Table of contents

List of figures

List of tables

List of Acronyms / Abbreviations

3RS	Rupture Resolution Rating System
ADU	Affective Dictionary Ulm
AFF	Affect Dictionary
AN	Affect Negative
AP	Affect Positive
AS	Affect Sum
ASD	Autism Spectrum Disorder
ASR	Automatic Speech Recognition
AUC	Area Under the Curve
AVEC	Audio/Visual Emotion Challenge
AZ	Neutral Affect
BAI	Beck Anxiety Inventory
BDI	Beck Depression Index
BERT	Bidirectional Encoder Representations from Transformers
BLSTM-NN	Bidirectional Long Short-Term Memory Recurrent Neural Network
BoAW	Bag of Audio Words
BSP	Behavioural Signal Processing

CASS Chinese Annotated Spontaneous Speech corpus

CAT-GSH Cognitive Analytical Therapy Guided Self Help

CAT Cognitive Analytical Therapy

CBT-GSH Cognitive Behavioural Therapy Guided Self Help

CBT Cognitive Behaviour Therapy

CCC Concordance Correlation Coefficient

CES Cross-cultural Emotion Sub-challenge

CNN Convolutional Neural Network

COVID-19 Coronavirus Disease 2019

DAAP Discourse Attributes Analysis Program

DBN Bayesian Network

DF Disfluency Dictionary

DNN Deep Neural Network

DPD Depressive Pseudo-Dementia

E2E End-to-End

eGeMAPS Geneva Minimalistic Acoustic Parameter Set

EWE Evaluator Weighted Estimator

F0 Fundamental frequency

FMD Functional Memory Disorder

GAD-7 Generalised Anxiety Disorder

GES Gold-standard Emotion Sub-challenge

GLM Generalised Linear Model

GMM Gaussian Mixture Model

GP General Practitioner

GSH Guided Self Help

HighP High Proportion score

HMM Hidden Markov Model

HNR Harmonics-to-Noise Ratio

IAPT Improving Access to Psychological Therapies

IVA Intelligent Virtual Agent

LDA Linear Discriminate Analysis

LI-CBT Low-Intensity CBT

LIWC Linguistic Inquiry and Word Count

LLD Low-Level Descriptors

LLR Locally Linear Reconstruction

LME Linear Mixed Effect Model

LSTM-RNN Long Short-Term Memory Recurrent Neural Network

MAE Mean Absolute Error

MDD Major Depressive Disorder

MFCC Mel Frequency Spectrum Coefficient

MHigh Mean High

MI Motivational Interviewing

MINI Mini International Neuropsychiatric Interview

MISC Motivational Interviewing Skill Code

MLLT Maximum Likelihood Linear Transform

ML Machine Learning

ND Dementia

Neg Negation Dictionary

NHS National Health Service

NICE National Institute for Health and Clinical Excellence

PANSS Positive and Negative Syndrome Scale

PCA Principal Component Analysis

PCL-C PTSD Checklist-Civilian version

PD Parkinson's Disease

PHQ-9 Patient Health Questionnaire

PTSD Post-Traumatic Stress Disorder

PWP Psychological Wellbeing Practitioner

RA Referential Activity

RBF Radial Basis Function

RECOLA REmote COLlaborative and Affective interaction

REF Reflection Dictionary

RFECV Recursive Feature Elimination Cross-validation

RFE Recursive Feature Elimination

RMSE Root Mean Squared Error

R Pearson Correlation Coefficient

RR Robust Regression

SAL Sensitive Artificial Listener Database

SenS Sensory Somatic Dictionary

SER Speech Emotion Recognition

SEWA Sentiment Analysis in the Wild Database

SFS Sequential Forward Selection

SRILM SRI Language Modeling

SVM Support Vector Machine

SVR Support Vector Regressor

TDNN Time-Delay Neural Network

TF-IDF Term Frequency-Inverse Document Frequency

THEPS Psychotherapy Session's Recordings

VAD Voice Activity Detection

VAM Vera Am Mittag Database

WAI Working Alliance Inventory

WER Word Error Rate

WFST Weighted Finite State Transducer

WRAD Weighted Referential Activity Dictionary

WRRL Weighted Reflection/Recognising List

WRSL Weighted Arousal List

WSAS Work and Social Adjustment Scale

WSJ Wall Street Journal

ZCR Zero-Crossing Rate

Chapter 1

Introduction

Mental health involves many disorders, such as anxiety and depression. These disorders may lead to problems with health and daily activities (World Health Organization, 2017). Psychotherapy has been demonstrated to be useful for many people with a variety of mental disorders, assisting them in regaining their everyday lives and improving their mental health. According to the Improving Access to Psychological Therapies (IAPT) programme of the National Health Service (NHS) in the UK, 1.46 million people were referred to psychotherapy in 2020-21. Of those, 51.4% of the people who completed the treatment, recovered in 2020-21 which is up 0.3% from 2019-20 (NHS, 2021). Wolberg (1988) defined psychotherapy as:

> The treatment, by psychological means, of problems of an emotional nature in which a trained person deliberately establishes a professional relationship with the patient with the object of (1) removing, modifying, or retarding existing symptoms, (2) mediating disturbed patterns of behavior and (3) promoting positive personality growth and development.

Psychotherapy is commonly referred to as talk therapy because it depends on conversational sessions between the therapist and the patient as a psychological treatment for mental disorders. According to Lambert and Bergin (1994), half of the patients needed five to 20 sessions to return to normal life, while 25% of the other patients required 30 to 50 sessions to achieve the same improvements. The person who receives treatment in the therapeutic session is referred to as a patient or a client. The professional relationship, referred to in the psychotherapy definition, is at the heart of the treatment and it is organised and promoted by the therapist (Wolberg, 1988). Maintaining a positive and reliable relationship between the patient and the therapist is one of the fundamental points defining the quality of therapy. However, it is not always easy for the therapist to judge the quality of the relationship across

sessions and therefore they are not always able to effectively repair a negative relationship. For this reason, there is a need for an automated system that can observe and help the therapist promote a positive relationship in the therapeutic treatment.

1.1 Motivation

Patient engagement is an important factor that can support the achievement of optimal results in psychotherapy. Unfortunately, if patients *drop-out* without finishing a complete sequence of sessions, they do not effectively receive the psychological treatment needed. In an analytical study conducted by Wierzbicki and Pekarik (1993) on 125 studies on psychotherapy drop-out, the average patient drop-out rate was 46.86%. This drop-out may relate to different factors such as, the patient's attitude toward the treatment, transportation difficulties preventing the patient from attending the sessions, the therapeutic relationship between the therapist and the patient (*Therapeutic Alliance*) and the patient's social class (Beckham, 1992). Although patient drop-out may affect the patient's successful outcomes, other parties can also be negatively affected by this. Mental health agencies can be affected economically in that a single patient drop-out can burden the agency with financial charges due to revenue loss, staff salary and overhead costs. Furthermore, this may lead to the staff experiencing low morale and high turnover (Barrett et al., 2008). Additionally, early termination of therapy can lead the therapists to feel negative emotions that could conflict with their work efficiency. Others might feel disappointed from their failure to complete the sessions (Piselli et al., 2011). In cases of patient drop-out, the therapist will usually ask the patient for feedback or try to understand the underlying reasons.

Several techniques can be used to minimise patient drop-out scenarios, such as educating patients about the benefits of therapy, keeping track of the patients' progress, focusing on the therapeutic alliance and informing the patients about their progress (Barrett et al., 2008). Although these techniques are easy to implement, it can be difficult for the therapists to keep track of all these signs throughout the entire therapy journey, which can span several weeks or months. For this reason, to avoid patient drop-out cases, there is a need for an automatic tool that can assist the therapist in improving their work by keeping track of the therapeutic alliance signs.

Therapeutic sessions are mainly conversations between the therapist and the patient. They are usually recorded and analysed to assess different traits. The recorded speech can give information about what the speaker is *saying* (verbally using language) or *expressing* (non-verbally using speech acoustics). Some studies have investigated the speech or transcript of a series of sessions to discover signs that can enhance the psychotherapy journey, especially

acoustic and language-based cues (features) of the speech (Hoffman et al., 2013; Imel et al., 2014). Although people usually communicate verbally with others, their internal states cannot be fully discovered using only their verbal expressions. Non-verbal communication in a conversation conveys important information, such as the emotions of the speaker. It would be useful to use an automatic method to analyse speaker's verbal and non-verbal information. The automatic analysis of the speaker's acoustic and language-based properties could reveal their mental and emotional states during treatment.

As mentioned earlier, one of the factors related to patient drop-out is the therapeutic alliance. In other words, developing a good relationship between the therapist and the patient can prevent patients from dropping out of sessions. This can also produce a better environment for the therapist to address the patient's concerns and achieve more optimal treatment results. The quality of the therapeutic alliance is positively correlated with the results obtained from different psychotherapy treatments (Horvath et al., 2011). It is known that there are various patients' and therapists' behaviours that can indicate a positive or disrupted therapeutic alliance, such as the synchrony between the patient and the therapist, the therapist's level of competence, the patient's emotions, the therapist's empathy and the patient's emotional state. The automatic detection of those signs (behaviours) could help maintain a positive therapeutic alliance in the session. These behaviours are discussed from a clinical viewpoint in Chapter 2.

There are several dyadic dynamics that can occur between the therapist and the patient in the session. Those dynamics could include several changeable patient or therapist behaviours over time, such as patient's and therapist's emotions. Many therapist's and patient's dynamics have been shown to be positively correlated with patient's treatment outcomes (Li and Kivlighan, 2020). Thus, automatic tracking of the important behaviours in patient's and therapist's speech over time would be beneficial for therapy. The tracking process could even help therapists regulate their emotions and enhance the patient's treatment outcomes.

Taking into consideration that direct automatic assessment of the therapeutic alliance might require high order cognitive and affective models for the individuals (Martinez et al., 2019), in this thesis, an automatic system for analysing recordings of psychotherapy sessions is proposed. This system will be able to detect and track several patients' and therapists' behaviours using acoustic and language-based features, which have been found to be positively correlated with the therapeutic alliance.

1.2 Thesis aim

Psychotherapy sessions are the core of mental health treatments. However, there are a high number of patients dropping out of sessions. Analysing psychotherapy sessions could help determine the common signs of a positive therapeutic alliance to avoid patient drop-out cases, which can negatively affect the economic and therapeutic factors of treatment. This project aims to build an automatic system for analysing real (genuine) psychotherapy sessions and to investigate approaches to automatically analyse psychotherapy sessions through detecting and tracking several signs relating to the positive therapeutic alliance.

In order to achieve the aims of the study, the following research questions should be addressed:

RQ1: To which degree is it possible to automatically *detect* any of the signs (behaviours) that may align with the positive therapeutic alliance, including any behaviours caused by the patient's emotional state, the therapist's competence, the therapist's empathy, the synchrony between the patient and the therapist and the patient's mood?

RQ2: To which degree is it possible to automatically *track* any of the signs (patient's emotions, the synchrony between the patient and the therapist and the patient's mood) in a single session?

This will include investigating several acoustic and language-based features to detect any of the therapist or patient behaviours mentioned in the research questions. Furthermore, an Automatic Speech Recognition (ASR) system will be established in order to achieve the full automation process.

1.3 Contributions

This section presents the contributions achieved in this thesis along with the papers published or prepared to publish under each contribution.

Contribution 1: Automatic System for Analysing Psychotherapy Sessions

From the literature, several patients' and therapists' behaviours were found to be positively correlated with the occurrence of the therapeutic alliance in the session. It would be beneficial to investigate the automatic detection of these behaviours to determine a positive therapeutic alliance. These behaviours could be one of the following but not limited to: the patient's emotional state, the therapist's competence, the therapist's empathy, the synchrony between the patient and the therapist and the patient's mood. These behaviours

can change quickly across time within a session. For this reason, the ability to track some of these behaviours automatically could be beneficial for time-continues prediction of the aforementioned behaviours.

As noticed in the literature, there is a lack of systems that address several patients' and therapists' behaviours related to the therapeutic alliance in a single automatic system. Chapter 3 addressed the ability to construct such a system by analysing psychotherapy session recordings with an overall aim to detect whether a positive therapeutic alliance is present in the session. This project investigated several important speaker behaviours that, as mentioned earlier, correlate with the therapeutic alliance positively. Chapter 9 highlighted the findings gained after applying the sub-modules in the proposed system. Furthermore, an important finding is introduced after constructing the automatic system, which is the ability to distinguish different speakers' behaviours.

This is novel because as far as we know, this is the first automatic system that aims to detect a positive therapeutic alliance by detecting patients' and therapists' behaviours in the session. The proposed system for investigating the automatic analysis of psychotherapy sessions, will be introduced in Chapter 3. This contribution address the research questions RQ1 and RQ2.

Contribution 2: Real Conversational Psychotherapy Sessions

Real recordings of conversational psychotherapy sessions would enable the capture of true and more authentic signs of the therapeutic alliance and achieve a high degree of transparency. Furthermore, it would expose the true technical challenges in real psychotherapy sessions, such as background noise and microphone echo. The microphone echo is the sound impulse that arises from a reflection after the direct sound with a strength and time delay, which is sensed as a repetition, , that is also called reverberation. The dataset used for this thesis came from a study by Kellett et al. (2020) that compared the efficiency and clinical durability of treatments for anxiety disorders. This study included recordings of psychotherapy or guided self sessions for multiple patient-therapist conversations, where each recording is for distinct patients. Those pre-existing recordings were utilised in this thesis and named as the psychotherapy session's recordings dataset (THEPS). The dataset was accompanied by depression and anxiety outcome measures, competency ratings and patients' demographics. The most interesting aspect about the data is that it included ratings of a novel measure (*competency*) to assess the therapist's competence required in the delivery of anxiety and depression treatments (Kellett et al., 2021a). The data was recorded for two types of anxiety and depression treatments: Cognitive Analytical Therapy (CAT) and Cognitive Behavioural Therapy (CBT), which are considered a novelty in the dataset. In addition, as the data were

recorded between 2019 and 2020, some of the data were recorded during the Coronavirus Disease 2019 (Covid-19) pandemic and lockdown periods in the United Kingdom and some patients mentioned therapeutic struggles related to the pandemic. Furthermore, the study related to the recordings was conducted under the IAPT services, an NHS initiative that provides additional psychotherapy to the general England population. This thesis is the first to apply automatic methods to analyse this dataset. As such, all the experimental chapters (5, 6, 7, and 8) investigated this dataset for the first time either to analyse or evaluate the automatic detection systems which are part of the proposed system in Chapter 3.

This type of real data is considered novel because it is hard to collect such data due to the confidentiality and the moral principles related to psychotherapy patients' session recordings. Most of the existing speech databases concerning counselling or therapeutic recordings are not available for public use. Some legal and ethical issues may prevent researchers from sharing such databases, such as privacy and copyright while using natural speech databases. Thus, there are very few benchmark databases that can be shared between researchers to study the automatic detection of mood or emotional state (Kaya et al., 2019; Ringeval et al., 2018). These databases include acted or induced recordings under a controlled experimental setting despite the fact that those recordings may reflect exaggeration in the perceived human behaviours (El Ayadi et al., 2011). Recordings with acted emotions might not accurately match the speaker's true feelings, which could minimise natural emotions' recognition rates. Acted datasets do not reveal the speakers' true emotions, which is opposite to the case in psychotherapy sessions dataset that captures real-life emotions.

Contribution 3: Automatic Detection of Depression and Anxiety

This contribution is focused on investigating automatic methods for detecting depression and anxiety by predicting mood outcome measures in psychotherapy sessions in particular. Predicting these measures could help therapists differentiate between several mental disorders and ensure a more reliable prediction of the patient's diagnosis. Due to the relationship between depression and Dementia (Muliyala and Varghese, 2010), the work was evaluated on recordings of consultations related to Dementia as described in Chapter 6. This is novel because this is the first study to detect depression and anxiety using Dementia consultations' recordings by predicting depression and anxiety outcome measures. The results gained presented the applicability of the automatic system to detect dementia-related diagnostics with the support of acoustic features. Further evaluation was established to detect depression and anxiety for the patients in the psychotherapy session recordings dataset described in Contribution 2 as presented in Chapter 6. This evaluation showed that detecting depression and anxiety automatically in real-life psychotherapy session recordings using mood outcome

measures is a promising area, which currently relay on mostly manual diagnostics measures. This contribution addresses the research question RQ1. The details of this contribution will be presented in Chapter 6. The work on the Dementia dataset was published as:

⋆ Attas, D., Mirheidari, B., Blackburn, D., Venneri, A., Walker, T., Harkness, K., Reuber, M., Blackmore, C., and Christensen, H. (2021). Predicting levels of depression and anxiety in people with neurodegenerative memory complaints presenting with confounding symptoms. In *Intelligent Computing*, pages 58–69. Springer.

Contribution 4: Automatic Prediction of Competency Measures

Investigating the automatic prediction of therapist's competency ratings could confirm the occurrence of the therapeutic alliance in the session and validate therapist's knowledge and skill needed to deliver the therapy treatment. Therapists that were assigned a high level of competence, could be regarded as better able to ensure a more successful progress of patient's treatment. This is considered the first study to automatically predict the Low-Intensity CBT (LI-CBT) treatment competency ratings described in Chapter 2 using the THEPS dataset. This type of rating is novel in that it was not used before for any automatic detection or prediction system. The final implemented system described in Chapter 8 revealed the advantage of using both acoustic and language-based features in predicting the therapist's competency measures along with the promising features in the prediction system. This work in under preparation for publication in a journal as:

⋆ Attas, D., Kellett, S., Blackmore, C., and Christensen, H. (2022a). Automated detection of competency in psychotherapy sessions via speech and language processing. *Applied Sciences, Applications of Speech and Language Technologies in Healthcare (in preparation)*.

Contribution 5: Automatic Speech Recognition for Conversational Psychotherapy Sessions

Psychotherapy sessions are considered a challenging environment in which to achieve a high performance ASR. Through the literature, it was clear that using a conventional automatic speech recogniser in such a sparse data domain could not ensure high accuracy rates (Xiao et al., 2015a; Chen et al., 2021). Conducting ASR systems for psychotherapy sessions is considered challenging and would require a large effort to achieve decent results. As mentioned earlier, psychotherapy sessions are considered highly confidential and many patients would not agree easily to share this type of data. Therefore, collecting this type

of recording for many sessions could be challenging. In addition, ASR systems require a large amount of data to achieve reliable recognition results. For that reason, transfer learning, which is a state-of-art technique, was investigated to ensure more reliable results were obtained with a small amount of data. This is novel because this study is the first to perform ASR on the THEPS dataset presented in Contribution 2. Like many healthcare domain datasets, the THEPS dataset is a relatively small dataset recorded in a challenging but realistic environment. The details will be presented in Chapter 5. The findings in Chapter 5 showed the applicability of deploying an ASR system with the use of a small amount of data which is mostly the case in the medical domain, and it can act as a baseline for the other ASR systems that are based on the larger dataset. This contribution addresses research question RQ1. This work is under preparation for publication in a journal as:

⋆ Attas, D., Kellett, S., Blackmore, C., and Christensen, H. (2022a). Automated detection of competency in psychotherapy sessions via speech and language processing. *Applied Sciences, Applications of Speech and Language Technologies in Healthcare (in preparation)*.

Contribution 6: Automatic Time-Continuous Recognition of Emotional Dimensions

Patient's and therapist's emotions are considered one of the behaviours that are highly changeable in the session. Automatically tracking those emotions in the session could help therapist's identify moments of a positive or disrupted therapeutic alliance. Systems aimed to extract detailed representations of emotion in a continuous domain and the temporal dynamics of emotion are referred to as emotion tracking or continuous-time representation of the affective content (Malandrakis et al., 2011). Patient's and therapist's emotions are considered a dynamic behaviour and predicting those emotions using dimensions could help capture the small variations in the emotion. Those small variations could help therapists to regulate their emotions in the session, determine the appropriate treatment plan for the patient and guide the patient toward the treatment. Also, this could assist in achieving the study's overall aim. Due to the unavailability of the emotional labels for this thesis's main dataset, an available dimensional emotional dataset was used as benchmark data for training. The findings showed that utilising a benchmark database labelled with dimensional emotions is an advantage when predicting the patient's and therapist's emotions in real-life psychotherapy session recordings. This is the first study to track the emotions using the dimensional model based on real recordings of guided self help sessions used to assess therapist's competence. This contribution addresses the research question RQ2. The results gained from achieving this contribution revealed other behaviours intended to be detected in

RQ1 such as the therapist's empathy and the synchrony between therapists and patients. The work will be introduced in Chapter 7. It was published as:

⋆ Attas, D., Kellett, S., Blackmore, C., and Christensen, H. (2022b). Automatic time-continuous prediction of emotional dimensions during guided self help for anxiety disorders. *FRIAS Junior Researcher Conference: Human Perspectives on Spoken Human-Machine Interaction (SpoHuMa21)*.

Contribution 7: Using Referential Activity Measures as Language-based Features

The Referential Activity (RA) model estimates the degree to which a person's language is associated with emotions. It was mainly created to analyse the language of therapeutic sessions and its association with clinical evaluation (Bucci and Maskit, 2005). The model consists of several dictionaries mainly related to the person's use of emotional language. It has been demonstrated that the use of those dictionaries as linguistic measures indicates the correlation between the emotional elaboration and the therapeutic alliance (Atta et al., 2019).

This is the first study that investigated the RA measures as language-based features extracted from the therapist's and patient's speech to predict the therapist's competency ratings, including a preliminary measure introduced by Tocatly et al. (2019). The project first introduced a mapping for the manual competency rating items to the related automatic language-based dictionary based on the RA model. The research to date has investigated this preliminary measure only for the patient's language and it has not before been applied to the therapist's language in psychotherapy sessions, as confirmed by the authors [1]. This contribution addresses the research question RQ1. The results gained from this work revealed other behaviours intended to be detected in RQ1, such as the therapist's empathy. Furthermore, the findings demonstrated that selecting the RA as a language-based feature for predicting the therapist's competence is a reliable candidate for deploying the automatic system. The details will be presented in Chapter 8.

1.4 Thesis Structure

The thesis consists of 10 chapters, Figure 1.1 presents the organisation of the chapters in the thesis. The content of the remaining chapters is summarised as follows:

Chapter 2: Overview of Current Psychotherapy Practise. This chapter reviews the background on the current psychotherapy practise involving the common keywords

[1] (W. Bucci and K. Tocatly, personal communication, November 24, 2021)

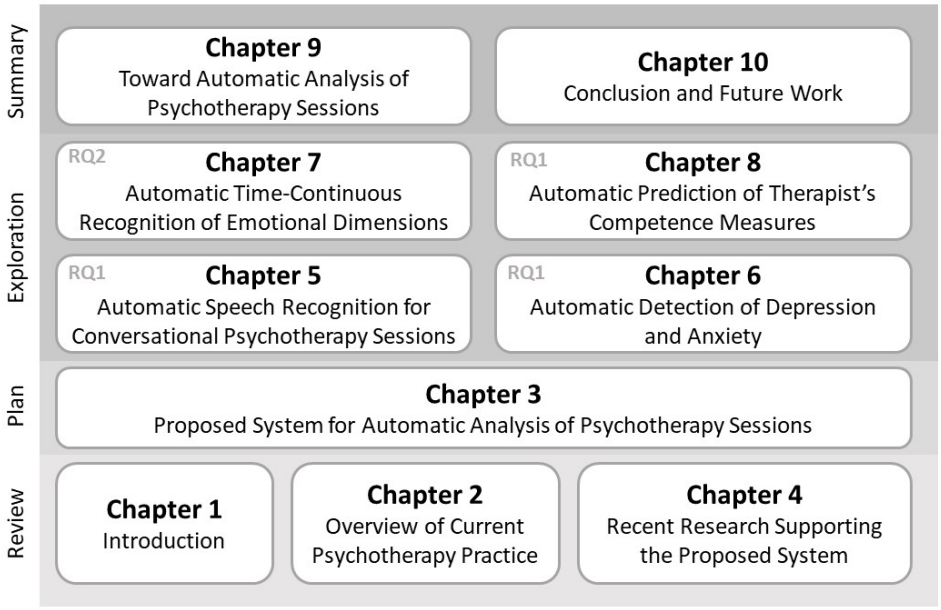


Fig. 1.1 Organisation of the thesis chapters highlighting the research questions addressed.

mentioned in the study. An overview of the nature of psychotherapy sessions is introduced alongside the common treatment types in therapy.

Chapter 3: Proposed System for Automatic Analysis of Psychotherapy Sessions. This chapter includes a description of the proposed system for automatic analysis of psychotherapy sessions and a full description of the main dataset used in this thesis.

Chapter 4: Recent Research Supporting the Proposed System. This chapter reviews the literature underpinning the structure and processing of the proposed system. This includes previous research on psychotherapy sessions, automatic detection of behavioural and emotional cues and automatic tracking of behavioural and emotional cues.

Chapter 5: Automatic Speech Recognition System for Conversational Psychotherapy Sessions. The work in this chapter demonstrates the system for ASR of psychotherapy sessions, including a section on related work, the data explored, the system architecture and the system experiments and results.

Chapter 6: Automatic Detection of Depression and Anxiety. The work in this chapter describes and validates a system used for detecting depression and anxiety using

mood outcome measures. Furthermore, it illustrates the exploration and evaluation process of the system on several datasets and the evaluation results.

Chapter 7: Automatic Time-Continuous Recognition of Emotional Dimensions. This chapter describes the experiment for predicting the dimensional emotions in a continuous manner for the patient and the therapist. It illustrates the related work, the system validation and the system analysis results.

Chapter 8: Automatic Prediction of Competency Measures. This chapter describes the experiments conducted for automatic prediction of competency measures. It facilities a literature review on the related systems, an analysis of the data used in the experiment and the experimental work for classifying and predicting the therapist's competency ratings. Finally, the results gained from the experiment are described in the chapter.

Chapter 9: Toward Automatic Analysis of Psychotherapy Sessions. In this chapter, a qualitative study of the experimental chapters mentioned earlier is presented. It includes an overall discussion of the current automatic experiments and results. Also, the connections between the results gained from the automatic systems and the literature studies' findings are discussed.

Chapter 10: Conclusion and Future Work. This chapter concludes the thesis work by including a summary of the PhD conclusions and suggestions for future work.

Chapter 2

Overview of Current Psychotherapy Practice

In order to reach the aims of this thesis presented in previous chapter, it is important first to understand the clinical practice of psychotherapy and the main attributes that determine the quality of therapy. This chapter will illustrate the common practice of psychotherapy, the common use of treatment sessions, several therapeutic interventions for mental and emotional problems and one of the most important attributes defining the quality of therapy, namely the *therapeutic alliance* determined by the relationship between the therapist and the patient. Ensuring a positive therapeutic alliance can enhance therapy treatment outcomes and minimise patient drop-out rates. On the other hand, having difficulties within this relationship could lead to contradictory therapeutic gains.

This chapter provides an overview of the current psychotherapy practice. Section 2.1 presents a brief introduction to psychiatry and Section 2.2 describes psychotherapy, which is one of the psychiatric specialities. Section 2.3 illustrates two related cognitive therapy interventions, namely Cognitive Analytical Therapy (CAT) and Cognitive Behaviour Therapy (CBT), highlighting the concept of therapeutic alliance. Finally, section 2.4 presents a summary of the chapter.

2.1 Psychiatry

Psychiatry is a medical field related to diagnosing, treating and preventing mental health disorders. Psychiatrists are medically qualified doctors who can prescribe medicines and recommend any form of treatment. They can also cooperate with a variety of other healthcare experts, such as mental health nurses, social workers and psychotherapists. There are

numerous psychiatric specialities, such as adult psychiatry, child and adolescent psychiatry, psychiatry of learning disability and psychotherapy (NHS, 2018b).

As reported by the National Health Service (NHS), the diagnosis of a mental disorder is usually obtained by a psychiatrist investigating both the mental and the physical health of the patient during the first session. The psychiatrist may initially conduct a clinical interview to ask the patient about their reasons for seeking help, their life and their thoughts. These questions are based on descriptive diagnosis manuals. Furthermore, other information about the patient may be obtained from a physical examination, relatives and social workers to eventually determine the proper treatment. Likewise, there are several clinical questionnaires (*mood outcome measures*) that assist in diagnosing and evaluating disorders. Examples include the Patient Health Questionnaire (PHQ-9) for depression, the Generalised Anxiety Disorder (GAD-7) for anxiety and the Positive and Negative Syndrome Scale (PANSS) for schizophrenia (to name but a few). The patients usually fill out those clinical questionnaires prior to each treatment session. The resulting mood outcome measures such as PHQ-9 and GAD-7 are changeable depending on the patient's mood and whether they are feeling either depressed or anxious while filling out those questionnaires. Predicting the mood outcome measures (PHQ-9 and GAD-7) can automatically assist the therapist by providing estimations of the patient's mood before the session (NHS, 2021; Newson, 2018; NHS, 2018b).

The PHQ-9 consists of nine depression modules, each scored from zero to three, summing up to 27 points. The scores range from zero ('not at all') to three ('nearly every day'). A total score is given at the end of the questionnaire to assess the implications of the problems mentioned in the questions on the patient's daily life routine. The PHQ-9 has proved its validity for use as a reliable tool to identify the severity of depression based on data from two studies with 6000 participants (Kroenke et al., 2001). A sample of the PHQ-9 is provided in Appendix A. In addition, GAD-7 is an assessment measure that can be used to detect generalised anxiety disorder. It contains seven questions, each scored from zero to three, with a total score of 21. It was validated in a large study and found to be a reliable tool for identifying anxiety severity scores (Spitzer et al., 2006). A sample of GAD-7 is provided in Appendix B. After the psychiatrist has assessed the patient's condition, a treatment plan may be prescribed for the patient to include medications or other clinical treatment sessions led by a psychiatrist or psychotherapist (therapist for short) specialised in the patient's condition (NHS, 2018b).

2.2 Psychotherapy

Psychotherapy is one of the psychiatry components that enables a patient to understand their difficulties, worries, abilities and motivations. Furthermore, it can assist the patient by investigating difficult and painful emotions and experiences or highlighting the behaviours or habits that could disturb them in life. It is the process whereby psychological difficulties are improved by a professional relationship between the patient and the therapist based on therapeutic regulations and techniques (Herkov, 2018; NHS, 2018a).

Patients with various mental disorders may benefit from psychotherapy as a treatment plan for conditions such as depression, anxiety, personality disorders, psychotic disorders, eating disorders and complex trauma disorders. Therapists can assess the patient's applicability for psychotherapy as a treatment and agree with the patient on the appropriate therapy plan for the following step. The therapist might consult other healthcare experts for advice on managing more complex situations. A number of interventions are used in psychotherapy, including psychodynamic psychotherapy, CBT, CAT, systemic therapy and trauma-focused therapy (NHS, 2018a). CAT and CBT are discussed later in this chapter, since both interventions are related to the main dataset used in this thesis.

2.2.1 Psychotherapy Sessions

Psychotherapy comprises conversations with a therapist to assist the patient with being comfortable with and able to change any suffering that makes life difficult. Treatment usually requires the therapist to work competently towards achieving the patient's treatment goals in a scheduled meeting or session, instead of providing a medical prescription. The patient expresses previous and recent experiences, emotions, reactions, relationships, reasoning and attitude in these sessions. The therapist's role is to recommend the best solution and plan the appropriate course to go beyond the patient's barriers or at least to help them live peacefully in life. The whole treatment procedure requires time and it would not be achieved in only one session (Milton, 2004).

The psychotherapy session begins when the patient arrives at the therapist's clinic, usually at a pre-arranged appointment and the sessions typically take 50 minutes which is a fixed duration for each session. Therapist tries to concentrate on the patient's inner feelings during the sessions without interruption from the outside world. Therapist's aim to remain in the background during the sessions to give patients the freedom to express themselves, even though sometimes the therapist's personality may come across in session. For that reason, therapists try not to talk about themselves and avoid normal social chat. The patient may consequently find it difficult to be friendly with the therapist. However, the therapist's

essential mission is to help the patient release feelings and not become friendly with them (Milton et al., 2011). Eventually, a large number of different emotions emerge from the patient during the session, either to express inner feelings or due to awkward engagement with the therapist.

2.3 Cognitive Therapy Interventions

CAT and CBT are psychological interventions for mental and emotional problems such as depression and anxiety. CAT provides a brief, structured and collaborative therapy that truly speaks to later life in the context of the patient's whole life story and the recognised importance of the conversation under a relieving and remedying therapeutic relationship. CAT can offer a coherent way of linking past and present and may be convenient for adulthood due to its attention to the interpersonal and need to find shared meaning and understanding in therapy through generational and cultural boundaries (Hepple, 2004).

According to the Department of Health and Social Care in the UK, CBT is considered one of the psychotherapies most commonly practised in the NHS (British Psychological Society, 2001). The Improving Access to Psychological Therapies (IAPT) program, which is the NHS's primary care mental health program, collects every session outcome, including a varied range of psychological practices. Furthermore, the National Institute for Health and Clinical Excellence (NICE) has strongly supported the use of CBT since a systematic review confirmed its effectiveness for depression and anxiety disorders (Clark, 2011). Additionally, CBT is considered a short-term therapy compared with other traditional psychoanalytic therapies and open-ended supportive psychotherapy (O'Donohue and Fisher, 2012).

In the IAPT program, high-intensity therapy denotes the standard CBT interventions delivered by a qualified mental health practitioner, usually weekly for 12-20 sessions. Due to the need for more efficient and effective interventions to meet the emergent need for mental health treatment, low-intensity psychological therapies have become more common over the past decade and this is especially the case for low-intensity CBT. Low-intensity CBT involves six hours or less of contact time, with each session typically lasting 30 minutes or less. Furthermore, the treatment utilises self-help materials and is delivered by trained practitioners or supporters. If the patient does not recover after low-intensity therapy, they are referred for high-intensity therapy (Shafran et al., 2021).

CBT empowers in patients awareness of their thoughts and emotions by identifying how situations, thoughts and behaviours influence emotions and improving feelings by altering dysfunctional thoughts and behaviours (Cully and Teten, 2008). As such, emotions are considered an important factor in CBT. The essential goal of treatment is symptom relief and

exemption of the patient's disorder. This can be achieved by acknowledging the patient's emotions and abstaining from challenging or disturbing them. Furthermore, the therapist tries to increase the patient's positive emotions through talk therapy, such as by examining positive events and memories (Beck, 2011).

CBT tries to conceptualise the patient's problem continually through sessions to understand their experiences and point of view. This is especially the case for specific situations when the patient's underlying beliefs evolve into particular thoughts that will affect their emotions and behaviours. There should be a realistic connection between the patient's thoughts, emotions and behaviours. Most patients can easily and correctly identify their emotions, although some may express emotions with less vocabulary or have difficulty labelling them. For that reason, it is helpful to encourage patients to link their emotional reactions in explicit situations to their labels. The emotion chart shown in Table 2.1 can assist patients in learning an efficient method for connecting their emotional reactions to their matching labels (Beck, 2011).

Table 2.1 Sample Emotion Chart from (Beck, 2011)

Angry	**Sad**	**Anxious**
1. Brother cancels plans with me.	1. Mom doesn't return phone call.	1. Seeing how low my bank account is.
2. Friend doesn't return my gym bag.	2. Not enough money to go away on vacation.	2. Hearing that we might have a tornado.
3. Carpool driver plays music too loudly.	3. Nothing to do on Saturday.	3. Finding a bump on my neck.

Another aspect for the therapist to consider is the degree of each patient's emotion, as patients may have flawed beliefs about experiencing emotions. Rating the intensity of an emotion can help patients to test their inner beliefs. This will also assist the therapist in determining if cognition requires further intervention (Beck, 2011). Eventually, the therapist should obtain a clear view of the patient's distressful situations and help them to distinguish their thoughts from their emotions. The therapist should be empathetic towards the patient's emotions throughout the therapy and assist them in evaluating the flawed thinking that has affected their mood (Beck, 2011). For that reason, the automatic detection and tracking of emotions in psychotherapy sessions would be an interesting task to investigate due to the large and changeable amount of different emotions.

The structure of the therapy within a session of CBT usually follows a specific guideline arranged by the therapist to improve a patient's mood and functionality during the week. This structure is based on the overarching goal of treatment, therapy notes and previous homework assignments. This arrangement may allow the patient to be more familiar with what to expect from therapy and to understand how the therapy will proceed. In the first part of a single session, the therapist aims to establish a positive relationship with the patient and prioritise the session agenda in a collaborative manner. The second part of the session is mainly about discussing the problems on the agenda and trying to solve them. The therapist seeks to teach the patient relevant cognitive, problem-solving, behavioural and other skills. This teaching phase is repeated during the session until the patient can summarise their new understandable thoughts. A homework assignment may be given in the session to remind the patient of new, more realistic ways of thinking about their problem and to apply solutions during the week. In the final part of the session, the therapist extracts the patient's views on the most important points of the session and writes them down, reviews the assignment and modifies it if needed and recalls the patient's feedback for the session and responds to it (Beck, 2011). From first contact between the therapist and the patient, it is important that trust and rapport are built. Research shows that positive relationships between the therapist and the patient, therapeutic alliance, are associated with positive outcomes (Horvath and Greenberg, 1994).

Therapeutic Alliance

The concept of the therapeutic alliance was introduced by Freud (1958) to consider the possibility of gaining benefits from the attachment between the patient and the therapist. Several authors have since addressed the concept of therapeutic alliance and its influence on therapy. Bordin (1979) introduced a definition of the therapeutic alliance to state that the patient and the therapist should consider the agreement on the following: tasks, treatment goals and establishing a genuine human relationship between them. They proposed that the therapeutic alliance should influence the results of the treatment positively because the patient should collaborate with the therapist in achieving their therapy goals. An analytical study conducted by Martin et al. (2000) was performed on 79 studies, indicating that there is a moderate positive correlation between the therapeutic alliance and the treatment outcome. Based on Bordin's theory, several measures were constructed to estimate the level of alliance and their relationship with the treatment outcome, including the Working Alliance Inventory (WAI) and Rupture Resolution Rating System (3RS). The WAI is a form containing 36 items that is filled in by the patient; each item is scaled on a seven-point range. The 3RS is a system designed to observe any disturbance in the alliance (rupture) and how to resolve it (resolution).

Some studies have investigated the relationship between the therapeutic alliance and patients who drop out of sessions, such as a comparative study done on 11 studies by Sharf et al. (2010) which found a high correlation between the therapeutic alliance and drop-out cases. A higher number of drop-out cases was seen among patients with a weaker therapeutic alliance with their therapist. A study by Cournoyer et al. (2007) used the therapeutic alliance measure as a process variable to estimate positive outcomes from treatment. The less involved patients in the sessions, which have lower therapeutic alliance measures, are less consistent in attending the sessions. The study showed that the therapeutic alliance could be used to predict patient drop-out.

Positive Therapeutic Alliance

There are different characteristics in therapy that assist in gaining a positive therapeutic alliance. In study by Fiedler (1950), therapists from different schools described the ideal therapeutic relationship to include an empathetic relationship. Real therapist's empathy was described by Greenberg et al. (2001) as not imitating or repeating the content of the patient's words, but the therapist should also understand the patient's goals and experiences. Furthermore, the therapist should capture the subtle variations and inductions behind the patient's words and reflect them. Eventually, if the therapist can introduce empathetic attitudes toward the patient, this would ensure a smoother relationship between them. Automatic detection of empathy can be used as an assisting tool for indicating a positive therapeutic alliance.

Other terms are used to refer to the relationship between the patient and the therapist, such as patient rapport. Such rapport is defined as "the relative harmony and smoothness of relations between people" (Spencer-Oatey, 2005). Tickle-Degnen and Rosenthal (1990) described three essential components that represent rapport. The first refers to mutual attentiveness, or the mutual interest that one may experience while interacting with others in any relationship. Positivity is the second element, which refers to a sense of friendship and caring for others. Finally, the most important component is coordination, which is often also termed as being 'in sync'. Synchronisation is one of the signals that regularly occurs in sessions. Scheflen (1963) noticed that synchrony can exist in a session if its participants act perfectly in rhythm, as if they had previously introduced to each other. For that reason, automatic synchrony detection can be used to detect if there is a positive therapeutic alliance or rapport in the sessions.

As mentioned earlier, therapeutic sessions, most of the time, convey many patient emotions that are either positive or negative. Detecting a patient's emotions automatically can assist the therapist in determining an appropriate diagnosis. Furthermore, it can aid in cap-

turing any signs during sessions that indicate a positive or negative alignment between the therapist and the patient. Sexton et al. (1996) conducted a study to investigate the relationship between the therapist and the patient on a moment-to-moment basis that would lead to the construction of the therapeutic alliance. They found that the patients who were more emotionally involved in the sessions maintained a high alliance (positive) and responded less depressingly when interacting with the therapist. Additionally, the verbal content of the low alliance patients showed more negative emotions than those of high alliance patients.

The therapist's ability to introduce trust and confidence within the therapeutic frame is considered essential for therapeutic success. Their ability to connect with the patient and have a sufficient level of competence can help patients achieve more reliable therapeutic gains. Additionally, the therapist's attributes similar to benevolence, responsiveness and dependability are predicted to be connected to the development and maintenance of a positive therapeutic alliance (Ackerman and Hilsenroth, 2003). The therapist should have the necessary competence to guide the patient, provide adequate help and ensure positive results (Bachelor, 1995). *Therapist's competence* can be defined as the level of skill shown by the therapist in delivering treatment and comprises the therapist's response and consideration towards the relevant contextual variables (Waltz et al., 1993). In a study by Weck et al. (2015) to evaluate the therapists' adherence, therapist's competence and therapeutic alliance, they found that the therapist's competence and therapeutic alliance to be important for the therapy outcome in the treatment of health anxiety. The process of assessing the therapist's competence involves analysing their capacity to provide treatment to acceptable standards, which requires evaluation of the therapist's knowledge, usability and implementation of the treatment. There are several measures to assess the therapist's competence that require treatment sessions (recordings of them) to be evaluated by a rater for the presence and quality of certain therapist-determined features referenced in the content of relevant treatment manuals. A score is generated based on the total ratings and if it is above a specified threshold, the session is judged to have been delivered adequately (Fairburn and Cooper, 2011). One of those measures is the Low-Intensity CBT (LI-CBT) treatment competency scale provided in Appendix C (Kellett et al., 2021a). The LI-CBT treatment competency scale is a scaled measure used for rating therapists during treatment for patients with mild-moderate depression and anxiety disorders. The measure depends on the COM-B model (Michie et al., 2014), which is a model of behaviour and behaviour change to conceptualise the patient's problem behaviour as resulting from the interaction of three factors: (a) capability to perform behaviour change, (b) the motivation for behaviour change, (c) the opportunity to carry out necessary behaviour change. This model is used to inform, guide and influence therapists

(practitioners) on treatment delivery. The LI-CBT competency scale consists of six items that enable treatment session raters to examine a range of competencies (Kellett et al., 2021b):

- Focusing the session
- Continued engagement competencies
- Interpersonal competencies
- Information gathering: specific to change
- Within session self-help change method
- Planning and shared decision making competencies

Each item in the scaled measure is described in a documented manual for the raters or clinical supervisors to ensure an adequate rating based on each specified competency. The competencies are rated on a scale from zero to six based on a competence level, as described in Figure 2.1 for each competency item. The total score may range from zero to 36.

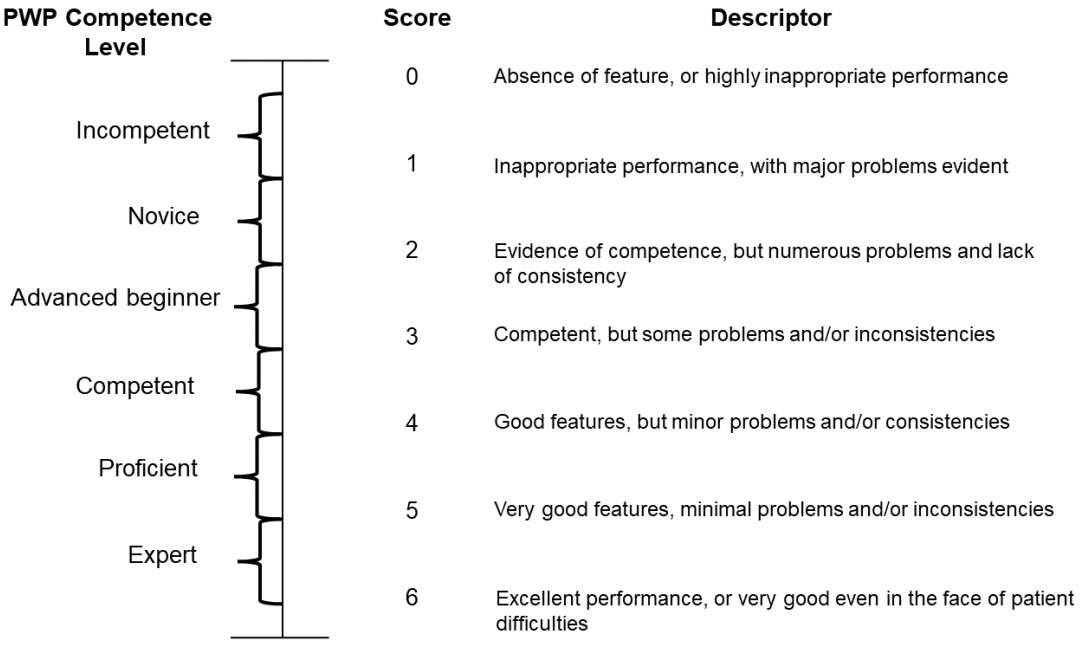


Fig. 2.1 Levels of competence as scored in the LI-CBT treatment competency scale

The *focusing the session* item concentrates on rating the therapist's ability to develop and subsequently adhere to an agenda for the treatment session in a competent manner. The

continued engagement competencies item assesses the therapist's competence in continuous engagement of the patient in the process of change in a collaborative way. The therapist should guarantee that the patient's progress is acknowledged by reflection and summaries. The *interpersonal competencies* item depends on rating the therapist's ability to develop interpersonal skills for maintaining therapeutic relationships with patients and for providing an empathetic and containing space for patients to proceed with their treatment. The *Information gathering: specific to change* item estimates the therapist's competence in gathering information from the patient concerning the transformation made and improvement achieved over the course of treatment in a positive and considerate manner. The *within session self-help change method* item evaluates the therapist's ability to select the appropriate treatment method for the patient based on their problem and following the treatment principles. The scale item *planning and shared decision making competencies* determines the therapist's competence in planning actions related to the patient's needs, session content and stage of treatment, taking into consideration the appropriate evidence base. Additionally, therapists should ensure that planning and associated decisions are made collaboratively in a patient-centred environment (Kellett et al., 2021b). The LI-CBT competency scale has been used for rating the main dataset used in this thesis.

Rupture in Therapeutic Alliance

According to the empirical importance of the therapeutic alliance in gaining better treatment outcomes, a study by Safran (1990a,b) tried to understand what can disturb this relationship. One of the frequent disturbances that may accrue in any therapeutic relationship is the *Alliance Rupture*. The alliance rupture is defined as any deterioration or alteration in the alliance quality among therapists and patients. Usually, the rupture can be observed in the session as a patient behaviour indicating crucial points to be explored more in therapy. Depending on the relationship between the therapist and the patient, ruptures can be different in duration, intensity and frequency. Sometimes, the therapist cannot detect if a rupture has occurred in the alliance or the patient is unaware of its existence. Usually, this would not interrupt therapy development. However, in severe cases, when the therapist does not detect the rupture, it might lead the patient dropping out of sessions or to treatment failure. On the other hand, if the rupture were discovered and resolved by the therapist, this would establish a significant opportunity for enhancing the treatment (Safran, 1993).

The 3RS system was built by Eubanks et al. (2015) to be used by human raters to observe and detect any rupture signs that might occur in the sessions and any attempts by the therapist to resolve the existing rupture. The method of attempting to resolve the rupture is called *Rupture Resolution*. The rupture signs might include pressure and an absence of

participation between the therapist and the patient. The 3RS is an observational manual system, meaning that it requires a third person (judge) to observe the existence of rupture during the therapy, while other systems that detect rupture usually include self-reported questionnaires. It has been proven that the observational systems that detect rupture can be used to enhance clinical practice and guide new therapists in the field. Coutinho et al. (2014) compared observational and self-reported systems when detecting rupture and noticed that self-reported methods do not precisely detect rupture because it was the patients who assessed the quality of the therapeutic alliance and not an independent judge, as was used in the observational methods. Furthermore, more rupture was identified using observational methods. Observational methods can detect rupture moment by moment in a session, but self-reported methods detect rupture once in a single session. A study by Eubanks et al. (2019) examined and validated the 3RS by comparing it with other rupture detection systems. The study proved that 3RS can be a reliable manual tool for exploring the therapy process and predicting drop-out cases. The study showed that the system can be a supportive tool for understanding how rupture can be established, recognised and resolved in therapy. Detecting rupture signs in therapy can assist the therapist to determine weak points in the therapeutic alliance that they might not be aware of and this could eventually prevent the patient from dropping out of the subsequent sessions.

2.4 Summary

In conclusion, psychiatry is an essential medical field for improving patients' mental health. It includes a variety of specialities, of which psychotherapy is one. Psychotherapy is more related to assisting patients to overcome their stress, emotions and disturbing habits. Among the common psychotherapy practises in the NHS is CBT, which depends mainly on exploring and linking thoughts, emotions and behaviours. Emotions can be an essential factor in CBT, in that the therapist should during sessions assist the patient to distinguish between emotions and thoughts, label emotions correctly and rate the severity of an emotion.

Furthermore, the therapist should build a stable and consistent relationship with the patient to reach optimal therapy results in an empathetic and understandable manner. It is vital to automatically detect the signs that can indicate a positive relationship between the therapist and the patient. For that reason, a proposed system that will address all the important signs previously mentioned will be presented in the next chapter, including the automatic detection of the patient's mood outcome scores, the therapist's empathy, the therapist's competence, the patient's emotions and the synchrony between the patient and the therapist.

Chapter 3

Proposed System for Automatic Analysis of Psychotherapy Sessions

The previous chapter discussed the various signs that might be observed in a session to indicate if a therapeutic relationship is positive or at risk of rupture. There was suggestive evidence that synchrony between the patient and the therapist plays an important role in establishing a positive therapeutic alliance. Another cue found to be related to the positive therapeutic alliance is the patients' emotions, as because of the nature of the psychotherapy sessions, several dynamic emotions could be observed during them. As discussed in Chapter 2, the scores gained from the mood outcome measures can indicate levels of the patient's mood, for example, whether the patient is depressed or anxious. A positive therapeutic alliance was also found to correlate with high levels of the therapist's competence. Taking these indications into consideration, it is clear that the therapists' and patients' behaviours during the therapy treatment could indicate a positive therapeutic alliance in term of the patient's mood, the patient's emotions, the therapist's competence, the therapist's empathy and the synchrony between the patient and the therapist. Developing an automatic system for detecting and tracking any of the mentioned behaviours can help indicate positivity in the therapeutic alliance and eventually lead to a successful treatment therapy. Such a system should take into consideration that direct automatic assessment of therapeutic alliance may require high order cognitive and affective models for the individuals, as cited by Martinez et al. (2019).

Studying rupture as a sign of deterioration in the quality of the relationship between the therapist and the patient could give insights into the opposite directions of a positive therapeutic alliance. Section 3.1 of this chapter examines the automatic methods for detecting rupture markers. Section 3.2 presents proposed system for automatically analysing psychotherapy sessions by detecting and tracking speakers' signs (behaviours), as discussed earlier. Section

3.3 explains the psychotherapy session's recordings dataset (THEPS dataset), which is the main dataset used in this thesis for developing the proposed system. Finally, Section 3.4 presents a summary of this chapter.

3.1 Automatic Rupture Marker Detection

The therapeutic alliance is considered to be among the essential elements that could indicate if the patient is going to drop out of therapy sessions. One of the signs that are important to detect is the possibility of a disturbance occurring in the relationship between the therapist and the patient, indicating a possible alliance rupture. Several automatic methods could be applied to automate the procedure of detecting rupture using observational methods. Observational methods usually require a human to observe a number of behaviours that could exist in a session and lead to a rupture. Several manual systems have been established to identify rupture in the therapeutic alliance, such as the Rupture Resolution Rating System (3RS) described in Section 2.3. The 3RS system has been adopted to automatically detect rupture moments in the therapeutic alliance.

3.1.1 Rupture Markers

Rupture can be accrued several times in a session, but not every minute throughout the session. If there is no rupture accrued in the session, the following characteristics should exist in the session, as reported in Eubanks et al. (2015) :

- Agreement between patient and therapist on the tasks of treatment.
- Agreement on the goals of treatment.
- A personal, affective bond between the patient and the therapist.

If a rupture occurred during a session, it could indicate deterioration in the alliance, demonstrating a reduction in tasks or goals collaboration or tension in the emotional bond between the therapist and the patient (Eubanks et al., 2015). For example, the patient and the therapist are not working together collaboratively and productively, or they seem distant from each other. There are two types of rupture, the first is the *withdrawal rupture*. In that type, the patient moves *away* from the therapist or the therapy work, or the patient could move *toward* the therapist if they deny their own experiences and therefore withdrawn from the work of therapy. The other type of rupture is called *confrontation*, which means the patient moves *against* the therapist either by expressing dissatisfaction or anger through a non-collaborative attitude or attempting to control or pressure the therapist.

Every type of rupture contain numerous markers defined in the manual of the 3RS system (Eubanks et al., 2015). Withdrawal rupture markers are: "Denial, Minimal response, Abstract communication, Avoidant storytelling and/or shifting topic, Deferential and appeasing, Content/affect split and Self-criticism and/or hopelessness". The Confrontation rupture markers are: "Complaints/concerns about the therapist, Patient rejects therapist intervention, Complaints/concerns about the activities of therapy, Complaints /concerns about the parameters of therapy, Complaints/concerns about progress in therapy, Patient defends self against therapist and Efforts to control/pressure therapist".

The manual rating is done every five minutes in the session by human judges and again after the session ends. The evaluation is based on each type of rupture marker using a scale from one (no rupture) to five (high rupture). The rating scale depends on the significance of the rupture and not the duration or frequency. Finally, a group rating should be performed for the overall existence of rupture depending on the type itself (Eubanks et al., 2015).

3.1.2 Mapping the Manual Rupture Markers to Automatic Characteristics

To detect rupture markers automatically, there is a need to translate or map each rupture maker's attributes to an automated one. Each rupture marker is described in a manual that provides human raters with a clear view of each marker based on a specific description (Eubanks et al., 2015). After reviewing the rupture marker descriptions in the related manual, the patient's and therapist's language in the session would be a good candidate for discovering the rupture type. For this reason, the automatic extraction of characteristics related to the manual rupture markers may require dictionary-based models for automatic descriptions such as Affective Dictionary Ulm (ADU) and Referential Activity (RA), which will be presented in next chapter. The RA estimates the degree to which a person's language is associated with emotions. It contains several dictionaries that reflect certain behaviours in the spoken language, including the Weighted Referential Activity Dictionary (WRAD), Disfluency Dictionary (DF), Affect Dictionary (AFF), Negation Dictionary (Neg), Sensory Somatic Dictionary (SenS) and Reflection Dictionary (REF). The ADU relies on quantitative measures of words related to their emotional implications, such as contentment, anxiety, depression, anger and fear.

Table 3.1 Mapping of rupture markers' manual descriptors to automatic characteristics

Type	Rupture Marker	Manual Rating Descriptors	Proposed Automatic Mapping Characteristics
withdrawal	Denial	The patient denies a feeling state.	Detect WRAD measure in the RA model.
	Minimal Response	* The patient exits the session before the end. * The patient answers a phone call without explanation. * The therapist talks more than the patient and the patient does not engage. * The patient pauses followed by minimal response or by a change in topic.	1. Count the number of words in patient's response. 2. Count the number of words in therapist's response. 3. Identify the topic of conversation.
	Abstract Communication	* The patient repeats words or makes global statements. * The patient feels conflicted.	Count the repetitive words in the patient's speech.
	Avoidant storytelling and/or shifting topic	* The patient talks about others (they, other people, all people, others). * The patient would change the topic of the conversation unless it was a useful topic. * The patient feels stressed.	1. Count the number of anxiety words from ADU in the patient's speech. 2. Identify the topic of conversation. 3. Count the number of disfluency words in the patient's speech from DF.
	Deferential and appeasing	The patient's language is appeasing the therapist.	Count the contentment words in the patient's speech from ADU.
	Content/affect split	The patient's affect language does not match the content.	Identify the DF words in the patient's speech from the RA model and match it with the patient's emotions.
	Self-criticism and/or hopelessness	The patient feels depressed and hopeless.	Count the depression words in the patient's speech from ADU.
confrontation	Complaints or concerns about the therapist	The patient feels anger and hate toward the therapist.	1. Count the anger words in the patient's speech from ADU.
	patient rejects therapist's intervention	The patient rejects the therapist's views or interpretations.	2. Identify the Neg words in the patient's speech from the RA model.
	Complaints or concerns about the activities of therapy	The patient's language confirms that they did not do the task, homework, or activity unless they faced obstacles.	3. Measure the SenS language in the patient's speech from the RA model.
	Complaints or concerns about the parameters of therapy	The patient conveys concerns or complaints about the parameters of treatment.	4. Count the fear words in the patient's speech from ADU.
	Complaints or concerns about progress in therapy	The patient conveys complaints, concerns, or doubts about the therapy progress.	5. Detect WRAD measure in the RA model.
	patient defends self against therapist	The patient uses defensive language to protect themself against the judgments.	6. Identify disfluency language in the patient's speech from the DF.
	Efforts to control/pressure therapist	The patient interacts in a hostile manner trying to control the therapist.	7. Identify REF words in the patient's speech from the RA model.

Table 3.1 illustrates the proposed methods for the latter dictionary models to automatically map the rupture markers' rating descriptions to automatic characteristics. After reviewing the manual of the rupture markers, the main human rating descriptions that could indicate each type of rupture marker have been selected, as indicated in Table 3.1. Afterwards, an investigation was conducted to find a suitable dictionary model for each rupture marker type, each representing the most important characteristics of that dictionary model.

Due to the unavailability of the rupture marker labels in the THEPS dataset, the process of automating the detection of the rupture markers has been suggested for future work. With the existence of the therapist's competence ratings in the acquired dataset, the Low-Intensity CBT (LI-CBT) treatment competency scale mentioned in Section 2.3 has been mapped to automatic characteristics.

3.1.3 Mapping the Manual LI-CBT Treatment Competency Scale to Automatic Characteristics

The LI-CBT treatment competency scale consists of several items that raters should take into consideration as they assess a therapist's competence. Each item is described in a documented manual with a detailed description of the key attributes that the therapist should demonstrate in the session. Each of these manual rating item descriptions has been mapped to automatic characteristics as a first step towards setting up an automated system. The ADU and RA have been adopted to provide the appropriate alignment between the manual rating descriptors and the automatic characteristics. The automatic system for detecting the therapist's competence using the LI-CBT treatment competency scale is described in Chapter 8. Table 3.2 presents the mapping of the competency scale items presented by the human description of each item as discussed in the related manual to automatic characteristics based on the RA and ADU models. An investigation was implemented to map those human descriptors to the best matching dictionary model and the corresponding measure of each dictionary model, as clarified in Table 3.2.

3.2 Proposed System

To achieve the system goals, there is a need to detect different signs that may improve the experience of both the patient and the therapist during therapy. For that, it is essential to assess the relationship between the therapist and the patient throughout the entire therapy journey and determine if there is a rupture in this relationship. Additionally, the automatic detection of some mental disorders that affect the patient's mood or thinking can help the

Table 3.2 Mapping of the manual competency scale descriptors to automatic characteristics

Rating item	Manual Rating Descriptors	Proposed Automatic Mapping Characteristics
Focusing the session	The treatment session's introduction provides an opportunity for the therapist to re-engage with the patient and highlight the planned and agreed content of the treatment.	Measure the REF words in the therapist's speech and SenS language in the patient's speech from the RA model.
Continued engagement competencies	The patient should feel positive and confident about the changes they have made/ are making or that their problems are being approached in a collaborative way. The therapist should confirm that progress is acknowledged by reflection and summaries.	Count the contentment words in the patient's speech from ADU.
Interpersonal competencies	The therapist should be able to sustain a trusting and containing therapeutic relationship with the patient. The therapist should ensure the patient has time to talk through any issues.	Detect the WRAD measure in the RA model.
Information gathering: specific to change	The therapist should inspire the patient to feel that they are an active participant in the treatment process and to feel positive through the changes they have made between sessions.	Identify REF words in the therapist's speech and WRAD in the patient's from the RA model.
Within session self-help change method	Cognitive restructuring is a way of influencing mood by targeting unhelpful thoughts via a process of identification and challenge.	Identify REF words in the therapist's and patient's speech from the RA model.
Planning and shared decision-making competencies	The therapist should present a patient-centered, flexible approach when delivering self-help and show a good understanding of the delivery of the treatment methods themselves.	Detect the WRAD measure in the RA model.

therapist build an overview of the patient's mental health state, such as depression and anxiety. Furthermore, detecting and tracking patients' emotions automatically in a session can help estimate accurate patient diagnosis and optimal treatment. Detecting some of the therapist's characteristics automatically, including competence and empathy, can deliver an objective assessment of the therapist's skills and professional behaviour. Finally, the automatic detection of the synchrony between the patient and the therapist could determine the level of the therapeutic alliance conducted between the two parties. Figure 3.1 shows block diagram of the proposed system for the automatic analysis of psychotherapy sessions.

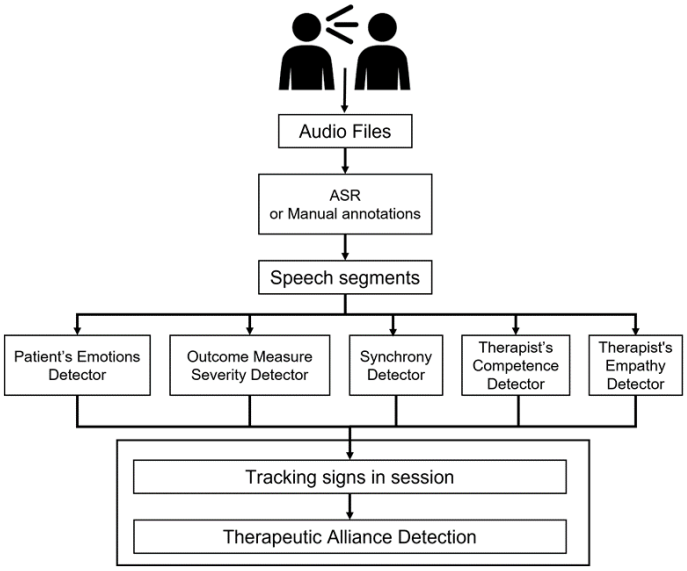


Fig. 3.1 Block diagram of the proposed system for automatic analysis of psychotherapy sessions

The proposed system starts by collecting recordings of several patients' sessions to be pre-processed. Initially, the audio files of the session recordings should be pre-processed using Automatic Speech Recognition (ASR) to attain transcriptions (what is said) and timestamp annotations (when it is said) for each speaking turn in the dataset. ASR provides an automated process for recognising spoken utterances, requiring the dataset recordings to be transcribed and annotated for training the ASR systems. The transcription is carried out by a team of expert transcribers trained explicitly for the task. Another option would be to use manual transcriptions and annotations in the pre-processing phase (Sell et al., 2018; Young and Mihailidis, 2010). The ASR system implemented, trained and evaluated on THEPS dataset is presented in Chapter 5.

After applying ASR on the audio recordings, each speaking turn is segmented to form a separate recording based on the timestamp annotations gained from the ASR. This results in separate speech segments that include each speaker in the session separately. For clarification, the speaker turns are referred to as speech *segments*. These speech segments are then passed to the next level of processing modules in the model. These each help extract various signs that have been identified in the literature (Chapter 2) to align with therapeutic alliance. The patient's speech segments are then used to implement the mood outcome measure detection module, as presented in Chapter 6. Furthermore, the patient's speech segments are employed for the time-continuous emotion detection module, as discussed in Chapter 7. The therapist's segments are adopted for the competence prediction module, as presented in Chapter 8. In addition, the therapist's segments can be explored for the empathy detection module. Both the therapist's and the patient's segments can be used to detect synchrony between the two parties.

Due to the interrelations between the several modules mentioned in the proposed system, some speakers' behaviours can be attained as a result of other modules, as described in Chapters 7 and 8. Tracking the patient's and the therapist's behaviours across the modules can assist in determining the moments in therapy that help establish a positive therapeutic alliance. Finally, the results gained from the detection and tracking modules can be integrated to form a stable system that aligns with the thesis aims and research questions. The discussion of the final proposed system is presented in Chapter 9.

3.3 Psychotherapy Sessions Dataset (THEPS Dataset)

This thesis focuses on predicting specific signs or cues to help therapists achieve a positive therapeutic alliance and eventually minimise patients' drop-out rates. To achieve those aims, it is essential to acquire psychotherapy session recordings to investigate the study aims and predict the important signs mentioned in the research questions. The recordings of the psychotherapy sessions used in this thesis are from research by Kellett et al. (2020) aimed at comparing the efficiency and clinical durability of two treatments for anxiety disorders. A total of 93 pre-existing session recordings were collected for the study, each including a conversation between a therapist and a patient during therapy treatment. Those recordings will be used to develop and evaluate the key modules of the proposed system, as described in Chapters 5, 6, 7 and 8. This section presents a brief introduction about the study purpose relating to the THEPS dataset, the clinical setting and selected outcome measures specified in the obtained dataset study, and the data collection and pre-processing.

In terms of using these recordings for automatic processing of the speech, several challenges should be noted. Those challenges are highlighted in the following points:

- The therapist and the patient may experience several inner states or behaviours during the session that would be reflected in their speech acoustics, such as the therapist's empathetic state or the patient's sad feelings.
- During episodes of extreme emotional states, the ability to hear some words would be challenging, such as if the patient were crying.
- Some parts of the session may include overlapping speech between the therapist and the patient. This would be difficult for most humans to hear and is extremely difficult to process with a speech recogniser.
- Differing from the acted databases (using human actors to record pre-defined behaviours in an ideal environmental settings), the sessions' recorded in a real world environment could include background noises such as doors slamming, chairs moving, clocks ticking, therapists typing, patients coughing and room echoes. Such background noises make it challenging to recognise words.
- Due to the Coronavirus Disease 2019 (Covid-19) pandemic, some sessions were recorded over a mobile phone. In some sessions, recognising speech in mobile phone sessions would be challenging, since the line might sometimes drop during the call and various background noises are captured from both the patient's and the therapist's side, such as a baby crying or a dog barking.

3.3.1 Background and motivation of the Clinical Study

As mentioned earlier, the audio recordings dataset used in this thesis is from research by Kellett et al. (2020) conducted to compare the efficiency and clinical durability of two anxiety disorders' treatments. The following section will describe the organisation responsible for conducting the study and the purpose of this study for the clinical community.

The Improving Access to Psychological Therapies (IAPT) programme is the National Health Service's (NHS) primary care mental health programme. It provides a systematic approach to organising and enhancing the NHS's psychological therapies delivery and access. The IAPT offers treatment for a variety of mental health problems, including depression, anxiety, panic and specific phobias. The treatments adopted by IAPT are developed by the National Institute for Health and Care Excellence (NICE). They are offered as a course of multiple sessions by a Psychological Wellbeing Practitioner (PWP). In this thesis, the name

therapist will be used instead of PWP for consistency. The IAPT services provide several NICE recommended therapies, such as Guided Self Help (GSH) (NHS, 2021), GSH is a type of psychological therapy combined with a self-help workbook to help solve a patient's mental problems with guidance from a therapist in a flexible time environment. It can be used along with Cognitive Analytical Therapy as (CAT-GSH) or Cognitive Behavioural Therapy as (CBT-GSH) (NHS, 2018c). As discussed in Chapter 2, CAT focuses on the reasons for the problem, challenges, how it started in the past and mainly how patients relate to themselves and others. CBT is concerned with links between actions, feelings and thoughts. It targets thoughts and actions in the now and here and especially how a change of thoughts and actions can change the patient's feelings (Jacobson, 2015). Although CAT and CBT are different types of therapy treatments, they use the same limited amount of sessions and goals.

The IAPT programme mentioned earlier adopted the treatment stepped-care principles for depression and anxiety. The stepped-care principals provide the least intensive and the least restrictive treatments based on evidence for the patient. In case a patient requires ongoing care, they are offered more intensive and costly interventions (Bower and Gilbody, 2005). Based on the stepped-care principles, the first step in therapy consists of an initial assessment by a general therapist, the second step includes low-intensity GSH treatment sessions and the third step comprises more in-depth psychological therapies. According to the clinical guidelines, there is a need for more patient treatment choices, mostly that each patient may prefer different interventions. The third step in IAPT services already includes a range of therapies (Perfect et al., 2016). However, the second step lacks this variety of therapies, especially as the currently available therapies in step two are based on CBT principles and conducted within the structure of group-based CBT, individual low-intensity CBT or computerised CBT. To acknowledge this issue, a study by Meadows and Kellett (2017) created a version of CAT-GSH to deliver treatment sessions for anxiety disorders. Implementing CAT-GSH has shown a reliable degree of GSH principles, low drop-out rates, suitable to be delivered in step two and effective implementation in the clinic with lasting short term results. Due to the minimal evidence that CAT therapy can be a treatment option for common mental disorders especially anxiety disorder, a study by Kellett et al. (2020) was conducted to compare the efficiency and durability of CAT-GSH and CBT-GSH for anxiety disorders performed at step two of IAPT service. The pre-existing session recordings collected from the latter research will be used as the thesis experiments' main dataset.

3.3.2 Setting and Selected Outcome Measures for Clinical Study

This section describes the study setting and selected outcome measures related to the aforementioned study by Kellett et al. (2020), which is the clinical source for the main thesis dataset.

When patients were referred to the IAPT service by the General Practitioners (GP), they were offered the choice to participate in the study trial. If they were interested, they would be given a trial eligibility interview to identify their suitability for GSH. The possible participants were screened using a short version of the Mini International Neuropsychiatric Interview (MINI) to perform a verification of an anxiety disorder (Sheehan et al., 1998). The patients selected a type of GSH therapy based on a detailed information sheet describing each type of GSH. A random selection would be considered if a patient did not select a specific GSH therapy. The therapy followed the IAPT structured six to eight session treatment protocols and was linked with patient workbooks for anxiety disorder. After the treatment sessions' termination, the patient were offered eight weeks and 24 weeks of follow up interviews. The Beck Anxiety Inventory (BAI), a well-known mood outcome measure for anxiety, was measured during the eligibility screening interview and the follow-up interviews. Several mood outcome measures were collected at the service screening, treatment sessions and follow up interviews, such as Patient Health Questionnaire(PHQ-9), Generalised Anxiety Disorder(GAD-7) and Work and Social Adjustment Scale (WSAS) (Kellett et al., 2020). PHQ-9 and GAD-7 are described in detail in Section 2.1. Those scales are useful for conducting automatic detection of depression and anxiety outcome measures. Also, the sessions were rated using the LI-CBT treatment competency scale described in Section 2.3 to assess the therapist's competence (Kellett et al., 2021a). The therapist's competence is defined as the degree of the therapist's knowledge and skill needed to deliver treatment efficiently to achieve its expected effects (Fairburn and Cooper, 2011). The competency rating scale empowers the raters to inspect a range of competencies consisting of six items: "(1) introducing the assessment session; (2) establishing engagement; (3) interpersonal skills; (4) information gathering relevant to the problem; (5) information giving relevant to the problem; and (6) shared planning and decision making." (Kellett et al., 2021a). This score is useful for conducting the automatic detection of the therapist's competence. Figure 3.2 shows the flow diagram of the study enrollment process, including referral, screening and allocation of patients.

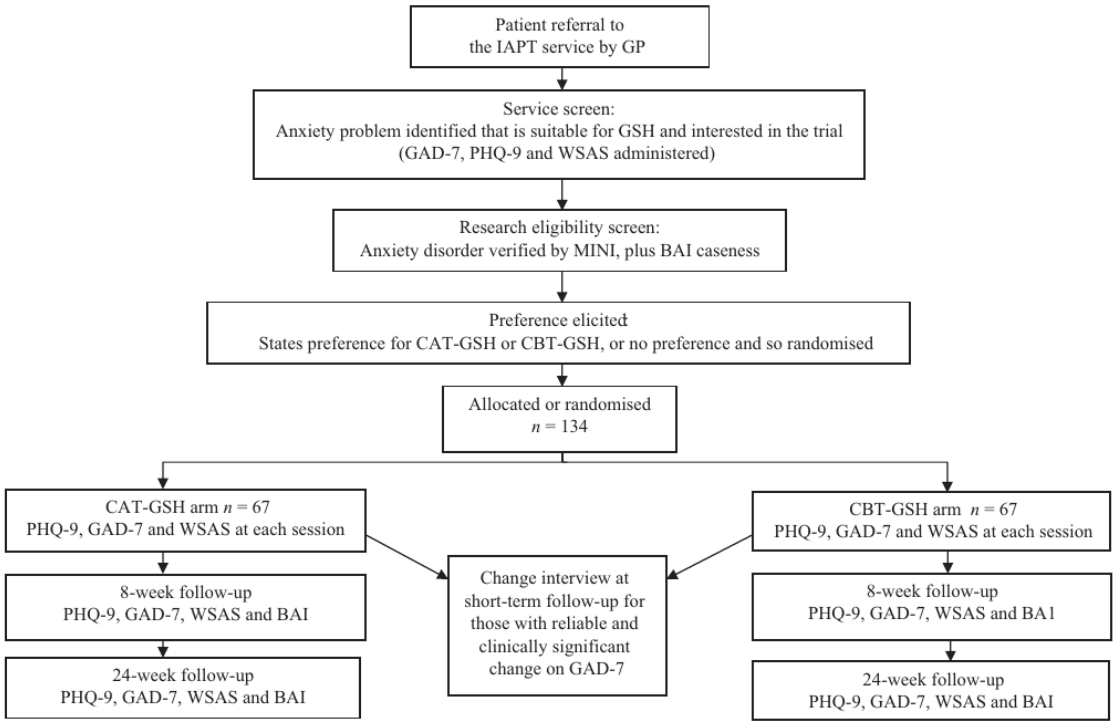


Fig. 3.2 Flow diagram showing the study enrollment process (Kellett et al., 2020)

3.3.3 Data Collection and Pre-Processing

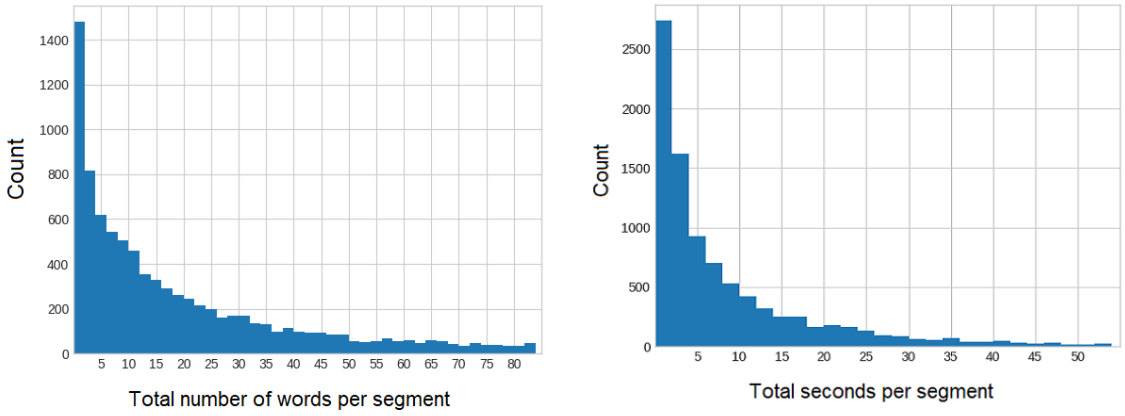
The total pre-existing session recordings from Kellett et al. (2020) study are 93 recordings, each including a conversation between a therapist and a patient as part of their therapy treatment. The therapists deliver low-intensity interventions for mild to moderate anxiety. They *guide* the patients through treatment in contrast to traditional therapists (Firth et al., 2015). The sessions could vary from 30 minutes to 40 minutes. The sessions were recorded using telephone recorders, audio recorders, or laptop recorders. The sampling rate for the recordings is 44100 Hz.

Some recordings was found to be corrupted and therefore eliminated from the study due to several reasons, such as being a duplicate, containing a high microphone echo or being very short or completely empty. The total number of recordings after eliminations is 84. The recordings included unnecessary parts in the conversation, such as scheduling the next session date and time. Those parts were trimmed from the recordings, which decreased the length of sessions to around 20 to 35 minutes. After implementing this, the total number of hours in the dataset is approximately 45 hours and 36 minutes. As mentioned earlier, some sessions were conducted over a mobile phone due to the Covid-19 pandemic. The total

Table 3.3 Patient demographics and therapy session information for the full dataset (Transcribed + Un-transcribed)

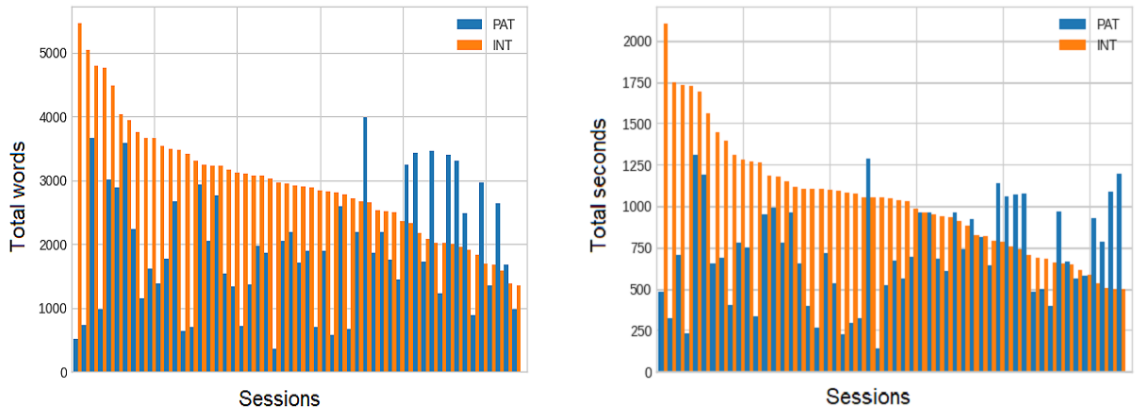
Patient demographics	**Total (all sessions)**	**Average**	**Min**	**Max**
Number of patients	84	-	-	-
Female	64 %	-	-	-
Age	-	37	16	74
In-person sessions	41 %	-	-	-

number of mobile phone sessions is 43 sessions, while the number of face-to-face sessions is 41. Table 3.3 describes the demographics of the dataset and therapy session information. The sessions were sent for transcription using a third-party agency. The number of transcribed sessions is 54 sessions with approximately 27 hours, 9271 total segments and 264,069 total number of words. This part of the dataset will be used to develop the ASR system described in detail in Chapter 5. The following figures are for the transcribed part of the dataset. Each speaker turn in the conversation was labelled in the transcription with the speaker name (PAT for patient and INT for therapist). Furthermore, the dataset has been annotated and adjusted based on the common filler words from several transcribers, with the finalised filler words being, umm, oh, aa, um, ah, hmm and mm. The transcriptions were also annotated with the start and end time for each speaker turn in minutes. The time alignment annotations for each segment's start and end times have been reviewed and adjusted manually by the thesis's author to produce more accurate timestamps for each segment in the dataset using the 'Transcriber' software (Barras et al., 2001). It is important to distinguish the therapist's from the patient's speaking turn due to differences in the acoustics and language-based characteristics expressed by each speaker in the sessions. Furthermore, other modules in the main proposed system design may require either the patient's or the therapist's segments separately to extract their specific characteristics. Some patients suffer from high depression or anxiety and that could affect on their use of words. Figure 3.3 shows a histogram of the total words and seconds in the transcribed part of the dataset. The figure shows that most of the time the speakers used few words around 20 words during less than 10 seconds per each segment. Figure 3.3 is only showing a sample window of the highest frequencies, due to the existence of a considerable number of low frequencies in comparison to the highest ones. Figure 3.4 shows the total seconds and number of words per session for each speaker. It is clear from the figure that most of the time the therapist would mostly speak more than the patient in terms of the number of words and time.



(a) Number of words (b) Total seconds

Fig. 3.3 Dataset histograms for the transcribed part of the dataset (A sample of the highest frequencies)



(a) Number of words (b) Total seconds

Fig. 3.4 Dataset details per session for each speaker (PAT: patient, INT: therapist)

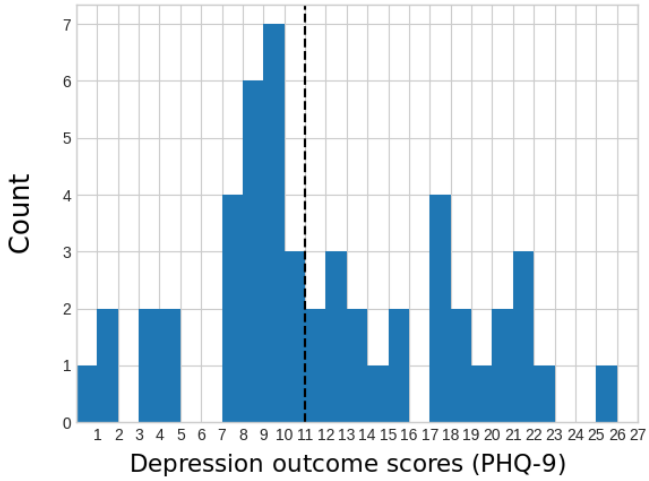


Fig. 3.5 Histogram of depression outcome scores (PHQ-9) presenting the depression severity cut-point in a dashed line, the higher the score means, the more severe state of depression.

As mentioned earlier, the dataset is labelled with the PHQ-9, GAD-7 and the competency ratings. The sessions that are labelled with PHQ-9 and GAD-7 are 51 out of 54 transcribed sessions. The 51 transcribed sessions labelled with the competency ratings intersected with the group labelled with PHQ-9 and GAD-7 scores. Figures 3.5 and 3.6 present the histograms of the PHQ-9 and GAD-7 values in the dataset. Figure 3.7 shows the histogram of the competency measure. The latter figures include the severity cut-off score in a dashed line to highlight the distribution of the low and high ends of the scores in the dataset which means that the line represents the dividing line between the low scores and high scores of each measure; this will be discussed further in Chapter 7. The figures show that most of the PHQ-9 and GAD-7 score ranges exist in the dataset, but that the available scores are not equally distributed in the dataset. Figure 3.7 shows the competency ratings presented in the dataset, which range from 13 to 27, in that it does not limit the distribution of the scores based on the several levels of competence.

3.3.4 Data Samples

This section introduces samples of the THEPS dataset and their corresponding transcription, highlighting some of the challenges recognised in the session recordings. The *spoken noise* transcription label denotes any noise related to the speaker's articulations, such as laughing and crying. The *noise* transcription label denotes any noise that exists in the recording

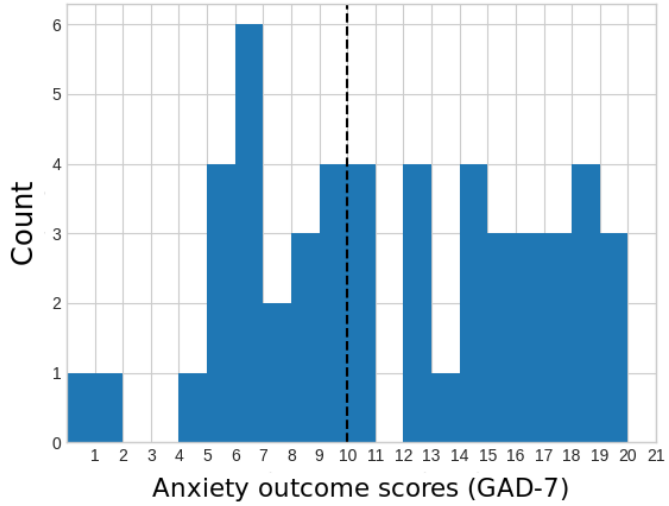


Fig. 3.6 Histogram of anxiety outcome scores (GAD-7) presenting the anxiety severity cut-point as a dashed line, the higher the score means, the more severe state of anxiety.

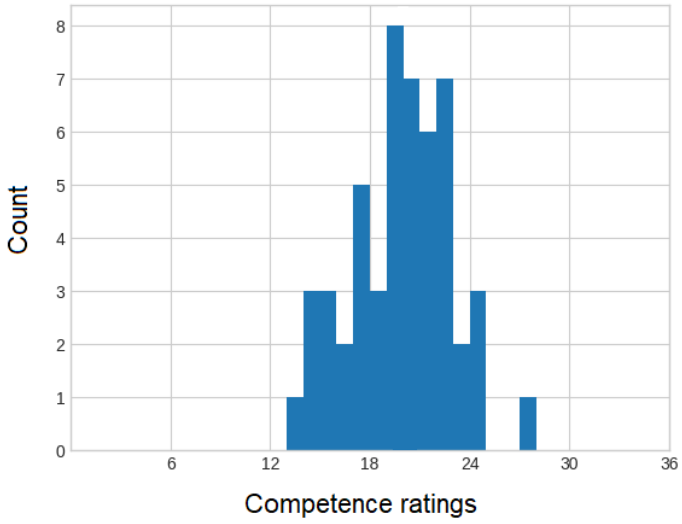


Fig. 3.7 Histogram of competency measures presenting the competency ratings cut-point as a dashed line, the higher the score means, the more the therapist's state of competence.

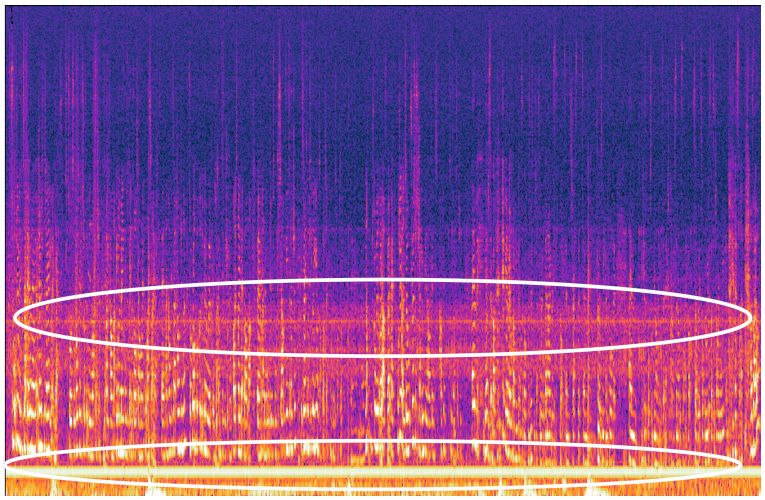


Fig. 3.8 Spectrogram of a sample from the THEPS dataset displaying types of electrical noise.

environment, such as a door slamming and fire alarm sounds. The *unknown* transcription label denotes the words that the transcribers could not recognise.

Some of the challenges in the THEPS dataset related to the electrical noise in the recorded signal chain, such as hum and buzz. These challenges are shown in Figure 3.8 where the hum is usually heard as a low-frequency tone and can be seen as a series of horizontal lines with a bright line at lower frequencies. In comparison, it can be seen that the buzz extends to higher frequencies and appears as a thin horizontal line.

The manual transcription of the audio sample shown in the spectrogram Figure 3.8 as follows, highlighting the use of the transcription label (spoken noise) due to the difficulty for the transcriber to hear the spoken words while the electrical noise existed in the recording:

> Therapist: So what we'll do here if we set a bit of an agenda for today anyway.
> Patient: Yes.
> Therapist: So also we go through the stuff that we did in the last session, how you found that. I think one of the main things was getting the thoughts down, weren't it, for the last time?
> Patient: Yeah.
> Therapist: And how we can start challenging them. So throughout the booklet that we're going to work through, so it's cognitive restructuring; I'll explain a bit about that and then what we'll do is we'll just start with the intervention. Is there anything in particular you want to talk about today, anything that's cropped up?
> Patient: No.
> Therapist: No? And just go with it like that?

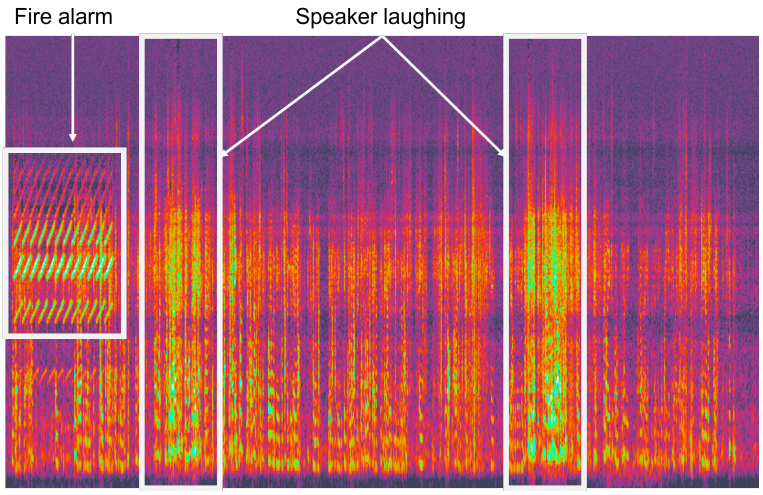


Fig. 3.9 Spectrogram of a sample from the THEPS dataset displaying fire alarm sound and speaker laughs

> Patient: Yes, yeah.
> Therapist: Right, OK, we'll do that then. So first just to check in with yer how things have been since our last session.
> Patient: Yeah, just that, well (spoken noise) we had a bit of a bad weekend, I suppose, cos we've been stressed at work but then I was, got stressed that, I know it was the weekend before but because I'd booked something on and it was at an extra price and I was like got so worked up and we just had such a horrible weekend (spoken noise) it's like.

Another challenge in the THEPS dataset is the short impulse noises such as fire alarm and door slamming sounds. Those challenges made it difficult for the transcribers to transcribe the words spoken while the noise existed in the recordings, as presented in the transcriptions below. Also, more spoken noise was noticed related to the speakers, such as laughs and filler words that would affect the manual transcription process. Figure 3.9 and 3.10 highlight those challenges in the spectrogram images of samples from the THEPS dataset.

The manual transcription of the audio sample shown in the spectrogram Figure 3.9 as follows highlighting the use of the transcription labels (spoken noise, noise, and unknown) due to the difficulties mentioned earlier in the recording:

> Therapist: I suppose, and ultimately they feel (noise) sorry (unknown) ultimately (unknown) feeling abandoned.
> Patient: Yeah.
> Therapist: so that is (unknown) (noise) in the first place, they're still feeling alone, OK. Do you want to have a go at.

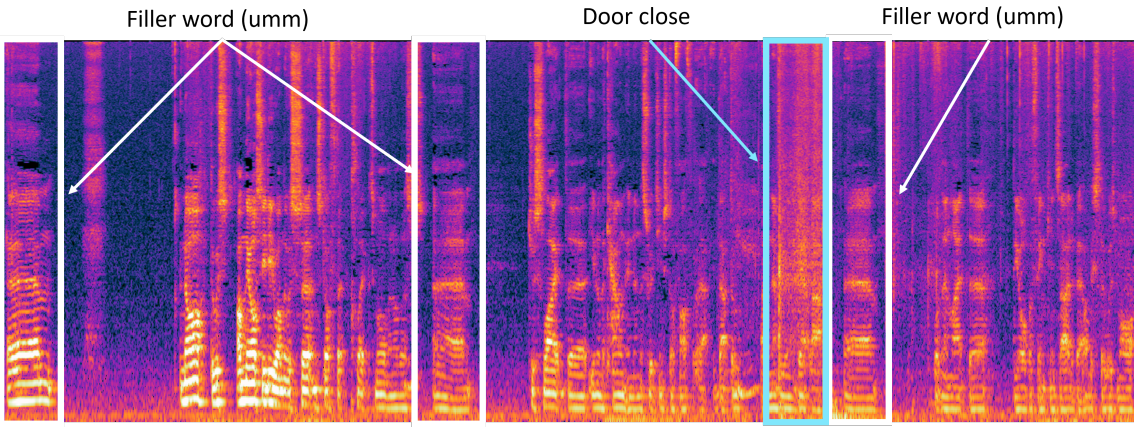


Fig. 3.10 Spectrogram of a sample from the THEPS dataset displaying door close sound and speaker filler words

Patient: Yeah, at least try.
Therapist: mapping out yours? (spoken noise) Find one and then.
Patient: Yeah. (spoken noise)
Therapist: Yeah, like I say, I think, because we spent quite a long time last session going over I suppose the traps from, from, from early on and then looking at the, the past, the past experiences to what's going on now we've prob, we've probably, probably started to get a bit of a, an idea from, from some of the things that we wrote down.
Patient: Yeah.
Therapist: So we've got, I, I'll write it on paper, right, don't like writing (unknown) (noise) so this is me there not wanting to make a mistake.
Patient: Yeah. (joint spoken noise) I hope you write that (unknown).
Therapist: (Unknown) (overlapping) I will. (spoken noise) So we've got the reciprocal role, your dad is the (unknown) area, and we said that he can, he could be quite judging.
Patient: Yeah.
Therapist: those high expectations (noise) and you feeling judged, not good enough?
Patient: Yeah, yeah.

The manual transcription of the audio sample shown in the spectrogram Figure 3.10 is:

Patient: Umm No, there was, on the OCD one there was certain stuff that clicked more than others, Umm just like the, you know, the counting and the switching stuff off; that, that don't, I never really get that. (door closed) Umm But then

like the, the overthinking side of it obviously was more, I guess, tailored to me, or suited me better.

3.4 Summary

Psychotherapy sessions are the main focus of this project. A proposed system for analysing psychotherapy session recordings has been introduced in this chapter to investigate the behaviours of therapists and patients in treatment sessions. For that reason, the THEPS dataset has been obtained and pre-possessed to implement the experiments of the thesis's proposed system. The obtained sessions were originally collected for a comparison of two types of guided self therapy for anxiety disorder treatments. The comparison was primarily conducted for efficiency and clinical durability. The THEPS dataset will be used to develop and evaluate the key modules of the proposed system. It is evident in the dataset that the speakers mostly spoke few words in a very short amount of time in each speaking turn. Furthermore, the therapist spent more time speaking in the session, which might relate to their need to guide their patients through the session and implement the treatment plan. To implement the automatic system proposed in this chapter, previous research on psychotherapy sessions and research studies concerning each module in the proposed system will be reviewed in the next chapter.

Chapter 4

Recent Research Supporting the Proposed System

In Chapter 3, the proposed system for exploring automatic methods for analysing psychotherapy sessions by detecting and tracking several patients' and therapists' behaviours across sessions was introduced. The important behaviours that the proposed system is focusing on in order to explore the positive therapeutic alliance in the treatment session are the patient's emotions, the patient's mood, the therapist's empathy, the therapist's competence and the synchrony between the patient and the therapist. To start investigating the applicable ways for implementing the proposed system, it is essential to review previous research that supports the detection of any of the aforementioned behaviours. Furthermore, it is of interest to review existing research aiming at tracking the behavioural changes represented in sessions.

4.1 Introduction

Recently, several clinical psychology and psychiatry research studies have emphasised the use of Machine learning (ML) approaches for learning statistical functions from multidimensional datasets to generally predict symptoms or behaviours about individuals. ML is commonly known as the computational strategy that automatically determines or learns methods and parameters to gain an optimal solution for a problem instead of being programmed by a human to present a fixed solution (Dwyer et al., 2018). ML algorithms are now integrated into users' daily life in internet searches, translation services, product recommendations and speech recognition services (Jordan and Mitchell, 2015). Furthermore, several ML techniques have been adopted to address multiple challenges investigated in clinical psychological and psychiatric research, such as the diagnosis, prognosis, treatment prediction and detection

and monitoring of potential biomarkers. The predicted results could be in the form of classification, such as if a person will benefit from a specific treatment or not, or as a regression framework to deliver continuous estimations, such as if a patient will benefit from a specific dose of a particular type of medicine (Dwyer et al., 2018). The current applications of ML in the field of psychotherapy research have demonstrated the ability of ML-based approaches to be informative about the treatment process, therapist's skill and treatment response prediction (Doorn et al., 2021). Furthermore, several studies have shown that it is possible to predict therapy outcomes and personalise psychiatric treatment using ML models, such as if a depressed patient will respond to specific psycho-therapeutic techniques (Chekroud et al., 2021).

ML works by fitting mathematical models to data to derive insights or make predictions. Those models typically take features as input that are a numeric representation of an aspect of the raw data. The features stand between data and models in the ML pipeline (Zheng and Casari, 2018). To deploy an ML pipeline in the proposed system presented in Chapter 3, it is important to understand the applicable features to represent the psychotherapy session recordings as input for the ML models. During psychotherapy sessions, several dynamic patients' and therapists' behaviours may be reflected in their speech. Those behaviours could be represented automatically using vocal cues (present in the acoustic signal) and verbal cues (as seen in words and sentences) as acoustic and language-based features. These features could be an input for ML models to predict an aspect of a speakers' behaviour, for example, detection of clinical outcome measures or assessment of the speaker's skills. This chapter will start by reviewing the previous research studies that employed automated approaches in the field of psychotherapy. Furthermore, the chapter will present a review of the appropriate features for deploying the proposed system, along with the research studies that adopted those features in the automatic detection of any of the human behaviours suggested in the proposed system from a literature review perspective.

The remaining of this chapter is organised as follows: Section 4.2 presents a review of the previous research on employing automatic approaches concerning psychotherapy sessions. Section 4.3 review the automatic methods of detecting several human behavioural and emotional cues highlighting the use of the language and acoustic features. Section 4.4 reviews previous studies on automatic methods for tracking human behavioural and emotional cues. Section 4.5 presents a list of the most common databases in the domain, 4.6 presents fundamentals in ML that were highlighted through the thesis and Section 4.7 summarises the chapter.

4.2 Previous Research on Psychotherapy Sessions

Several research studies in the psychotherapy field tend to investigate automatic approaches that could enhance the quality of therapy and therapeutic outcomes. One factor contributing to positive therapeutic outcomes is the therapeutic alliance. A study by Martinez et al. (2019) aimed to analyse therapeutic outcomes from the stories told during psychotherapy sessions. They proposed to model the alliance as a function based on the interaction between certain types of therapists with certain types of patients. Those types were automatically discovered from the sessions' transcripts by identifying the shared attributes between the therapist's and patient's characters in the stories told throughout the sessions. The character attributes were attained based on a *Personae* model. Personae (character archetypes) are classes of characters with similar traits, behaviours and motivations. The patient's and therapist's characters used automatic identification of their personae by an unsupervised model (personae model) for narrative understanding using a topic modelling technique. The data used in the study consisted of 1,235 recorded sessions with an average duration of around 50 minutes and labelled with a patient-reported alliance. For evaluation, they first trained the personae model with automatically transcribed sessions using an Automatic Speech Recognition (ASR) tool. Then, they analysed the relationship between the characters' personae and the therapeutic alliance to show that the discovered personae helped estimate the alliance. They used Linear Mixed Effect Models (LMEs) to measure the effect of characters' personae on the therapeutic alliance. Then, they trained a Support Vector Regressor (SVR) with a linear kernel and cross-validation fashion to capture the relationship between the alliance and the assigned personae. The results showed that using only therapist or patient text, the LME models performed poorly compared to using both therapist and patient information. Based on earlier research by the authors, the noted poor performance supported that an individual speaker's language use does not capture alliance information. Furthermore, the personae model was found to perform better than the supervised model that considers both therapist and patient text. They found that models trained with the therapist character from the patient's text and the patient character from the therapist's text achieved the best results compared to any other possible combination of characters. This study is interesting because it relates to the therapeutic alliance that is the main focus of this thesis, especially that it handle the use of automatic transcriptions.

Another related research domain has concentrated on investigating Motivational Interviewing (MI), that is a counselling style used in the field of psychotherapy to help people with addiction problems to resolve conflicts and aims at changing addictive behaviours. In MI, the therapist uses empathetic listening to understand the patient's perspective and minimise resistance (Rollnick and Miller, 1995). The therapist's empathy is commonly observed as a

quality index of psychotherapy sessions. Xiao et al. (2015b) investigated an automatic system for evaluating the quality of psychotherapy sessions by classifying therapist's empathy in 200 MI sessions for drug and alcohol counselling with human ratings of counsellor empathy. ASR was used to transcribe the sessions such that the resulting transcriptions were used in a text-based predictive model of empathy. For the ASR, two supporting datasets were used to help define the common language used in the psychotherapy sessions. Statistical N-gram language models were used to capture the differences in therapist empathy as detected by the human raters. Two separate models were investigated to classify empathy, either high versus low empathy or empathy scoring, using the automatic transcriptions generated from the ASR system. They used ASR (Kaldi adaptation) to transcribe sessions and the resulting words were used in a text-based predictive model of empathy. For training the ASR, they used 1200 therapy transcripts to help define the typical vocabulary and language use. The empathy scores were generated from a predictive model based on a scoring algorithm trained using human generated transcripts provided input words for an output of an empathy score. The results showed that both models were highly accurate against human-based ratings. Considering that the ASR results gained a mean Word Error Rate (WER) of 43.1%, the empathy prediction system based on human transcriptions as input yielded a slight increase in prediction accuracy (85.0%) compared to the system based on the automatic transcriptions with an accuracy of (82.0%). This study was complimented by a study analysing speech rate entrainment and investigating its relationship to perceived empathy (Xiao et al., 2015a). The degree of entrainment was measured by the averaged absolute differences of turn level speech rates for both patient and therapist. Furthermore, the researchers investigated the correlation of ratio and duration statistics of speech, pause and gap segments with the therapist's empathy ratings. The results showed that the degree of entrainment was correlated with the therapist's empathy rating. In addition, the speech rate, inter-word pause and inter-turn gap revealed knowledgeable information supporting previous prosodic cues detected for empathy modelling (Xiao et al., 2014).

These research studies highlight the good performance shown by automatic systems approaches to detecting therapeutic cues using psychotherapy session recordings.

4.3 Automatic Detection of Behavioural and Emotional Cues

As mentioned earlier, the speakers' behaviours expressed in the sessions could be revealed in their speech (acoustics) or language use. Those characteristics could be represented as

acoustic and language-based features for modelling several human behaviours. The acoustic features represent the physical characteristics of the speech signal, such as the loudness, frequency and amplitude of the signal (Wickramasinghe and Geisler, 2008). The extracted acoustic features are a sequence of acoustic feature vectors and each vector represents the information in a small time window of the signal (Jurafsky and Martin, 2013). The language features are related to the speakers' use of language. They are extracted from the text transcripts and can be lexical, semantic or syntactic in nature (Stegmann et al., 2020)

The ML models deployed for detecting human behavioural and emotional cues usually depend on features extracted from several modalities, such as speech. The detection of those behaviours could support therapists in decision making. This section reviews the different types of features from which could be extracted emotional and behavioural cues. Based on the research presented in literature, the previous studies that adopted those features to detect several human behaviours in a session will be reviewed in this section.

4.3.1 Language Features

Approaches developed for aiding language-based analysis of psychotherapy sessions typically involves predefined dictionaries developed to capture particular aspects of a patient's and therapist's language. Several dictionaries have been deployed to understand the language use of humans in conversational interactions. These dictionaries usually consist of words that reflect specific human behaviours by verbal expressions. Texts that are scored using these dictionaries by calculating mathematical functions capture the flow of the human behaviours intended to be reflected by those dictionaries. The calculated scores may act as an appropriate candidate for language-based features. This section will review two of the most related dictionary models in the literature that could be used as a base for language features in this thesis experiments.

Emotions represented in patient language can play an important role in assisting therapists in discovering patients' behaviours in the session. Some models have been proposed to distinguish emotions from each other using language. One model is Referential Activity (RA) which estimates the degree to which a person's language is associated with emotions. It was built by Dr Wilma Bucci and her team to analyse the language of therapeutic sessions and how it is connected to clinical evaluation. They proposed a set of functional stages that discuss the general process of bringing non-verbal material that occurred both outside of and within awareness into a form that could be translated to language (Bucci and Maskit, 2007; Mergenthaler and Bucci, 1999). The model incorporated three major phases: *Arousal*, *Symbolising* and *Reflection/Recognising*. The three phases often occur in order, although there might be re-occurrences of the phases, especially for the Symbolising and Recognising

phases. The Arousal phase involves activating a person's experience that is outside or on the edge of awareness. Those experiences could exist within the vast expanse of the person's personality, memories and knowledge base. Through the process of Symbolising, some discrete elements that exist in the flow of the experience description, such as episodes and images, are described in language. The person in that phase speaks within the experience of a specific event or image, as if they were there, in time and place. The next phase is the Reflection/Recognising phase, which is when the person retains the experience of the event or image; capable of reflecting on it from outside and finding new emotional meaning in the event (Bucci, 1997). The three phases are distinct psychological processes, such that each person is involved in a variable degree at each point of the conversation. RA has been observed in psychotherapy and other spoken and written conversation contexts (Maskit, 2021).

There are several dictionaries that were developed through a process of modelling the frequency with which words are presented in texts at several phases of RA as scored by human raters, yielding weighted dictionaries. The weighted dictionaries have weights lying between -1 and + 1, with 0, the weight received by words not in the dictionary, as a neutral value. The weighted dictionaries that relate to the RA process are as follows (Bucci and Maskit, 2007; Maskit, 2021):

- The *Weighted Referential Activity Dictionary (WRAD)* is a RA measure that consists of words identifying moments in language when a speaker is immersed in the narrative such that high WRAD indicates the Symbolising phase in the RA process.

- The *Weighted Reflection/Recognising List (WRRL)* is a measure that assesses the Reflection/Recognising phase in RA. It measures the degree to which a speaker is attempting to recognise and understand the emotional implications of an event or set of events in their own or someone else's life or in a dream or fantasy.

- The *Weighted Arousal List (WRSL)* is a preliminary measure for modelling the Arousal phase in the RA process. Based on a preliminary clinical validation of the WRSL dictionary, the results showed that the measure could distinguish between the step that can reflect moments going toward subsequent Symbolising and Reflection/Recognising phases as opposed to moments of avoiding that do not lead to such an RA process (Tocatly et al., 2019).

There are other dictionaries for the RA process, including unweighted dictionaries. These are lists of words with a common theme regularly used in RA analyses. These content-based dictionaries contain several words developed based on human raters picking words derived

from association with a target content category. The unweighted dictionaries comprise several related dictionaries (Maskit, 2021; Bucci and Maskit, 2007):

- The *Affect Dictionary (AFF)* measures how a person feels and communicates using feelings. The dictionary contains many emotional words, such as sad, happy and angry. Furthermore, the dictionary contains words related to the arousal affected by emotions, such as cried and screams. The words in the dictionary are classified as Affect Positive (AP), Affect Negative (AN), Neutral Affect (AZ) or Affect Sum (AS). AN words denote negative affect, AP words denote positive affect, AZ denote words without valence and AS is the union of AN, AP and AZ

- The *Reflection Dictionary (REF)* concerns how a person thinks and communicates through thinking. It includes words relating to logic, such as if and but. It also contains words referring to logical and cognitive activities, logical entities, failures in logical activities, difficult communicative actions and mental functioning.

- The *Disfluency Dictionary (DF)* includes items that people use in situations when they cannot describe their experiences in verbal form. The dictionary comprises exactly five items, kind, like, know, mean and filled pauses. The pauses are transcribed as uhm or uh.

- The *Negation Dictionary (Neg)* includes words denoting negating in communication, for example, no, not and never.

- The *Sensory Somatic Dictionary (SenS)* includes words denoting bodily and/or sensory experiences, for example, dizzy, eye, face and listen.

The Discourse Attributes Analysis Program (DAAP) is a computer-based text analysis system designed based on the RA model. It is designed to analyse any type of text, including written texts and transcripts of verbal language with any number of speakers. The DAAP reads the words in each speaking turn (segment), compares each word with the considered dictionary and assigns a number called the *dictionary value* to every word in the turn. The dictionary value is +1 if the word is in the dictionary and 0 otherwise for the unweighted dictionary. For the weighted dictionary, if the word matches an item in the dictionary, the dictionary value is the linear transformation of the corresponding dictionary weight, which lies between 0 and 1 (Maskit, 2012).

The RA model was created by Dr Wilma Bucci and her team in the beginning to assist therapists in recognising crucial areas in treatment. Hoffman et al. (2013) applied an automated text analysis on the treatment notes that are usually written by a therapist in a session.

They measured two essential variables: the Mean High (MHigh) of the WRAD, which is the mean of the referential activity when it reaches above the midpoint and the REF mean, which is the average number of words used in the REF Dictionary. The MHigh of WRAD indicates the intensity of the speaker's emotional engagement, while REF demonstrates that the speaker is distancing from the gained emotional experience. These two measures can play an important role in indicating if the speaker is emotionally engaged in therapy. If the MHigh of WRAD is higher and the REF is lower, that shows that the speaker is emotionally involved, living the experience and not moving back to reflect on it, while if the REF is higher than the MHigh of WRAD, the patient is less engaged in the therapy and stepping back from the contents. The researchers found that the linguistic analysis of process notes using the MHigh of WRAD and REF could find out points that the therapist missed in the clinical review and that would mean a difference in the recovery and progression of the patient. Furthermore, they suggested studying how these measures can help in predicting therapeutic failures and patient drop-out cases (Hoffman et al., 2013). Furthermore, a study by Bucci et al. (2012) found that high WRAD in therapists' process notes, denoted by greater emotional engagement for the therapist, was associated with better treatment outcomes. Due to the usefulness of deploying such a dictionary as a language-based feature in the thesis's proposed system, the DAAP has been adopted to analyse therapist's and patient's language as described in Chapter 8.

As mentioned earlier, the psychotherapy journey contains two important speakers, the therapist and the patient. It is important to diagnose each party's emotions either toward him/her self or to the other party in the therapy. One of the dictionaries that addresses this dimension is the affective dictionary Ulm (ADU), which is a computer model for investigating affective vocabularies. It is based on Dahl's emotion theory (Dahl, 1978) and was created for the German language (the English language edition is called ADUE in that E refers to English). It relies on quantitative measures of words related to the words' emotional implications. It consists of eight categories to classify emotions. It classifies emotions into three dimensions: Orientation, Valence and Activity. Orientation relates to the personal focus of attention, meaning that the person is focusing on an object or his/her internal state. Valence indicates if a person's emotions are positive or negative. Activity relates to the personal focus of control, whether active or passive. Table 4.1 shows the eight categories of ADUE (Hölzer et al., 1997). The mentioned emotional language models can be applied to capture several signs of a positive therapeutic alliance as proposed in Chapter 3.

Table 4.1 The eight categories of ADUE modified from (Hölzer et al., 1997)

	Positive	Negative
OBJECT	active-positive-object 'Love' (Affectionate, esteem, love, pity, sympathetic, tolerant, tender,...) 1	active-negative-object 'Anger' (Agressive, anger, cruel, dislike, furious, hate, envious, rage,...) 5
OBJECT	passive-positive-object 'Surprise' (Amazed, amused, astonished, fascinated, impressed, surprise,...) 2	passive-negative-object 'Fear' (Afraid, aversion, dominated, fear, humiliated, scared, shocked,...) 6
SELF	passive-positive-self 'Contentment' (Calm, contented, pleasant, quiet, safe, satisfaction, secure,...) 3	passive-negative,self 'Depression' (Alone, bad, despair, depression, helpless, lonely, miserable, sad,...) 7
SELF	active-positive-self 'Joy' (Adventurous, bold, courageous, elated, optimism, vigorous,...) 4	active-negative-self 'Anxiety' (Anguished, anxiety, frustrated, nervous, panicky, troubled,...) 8

4.3.2 Acoustic Features

Usually therapists record sessions with patients using an audio recorder or a videotape. These recordings are useful for evaluation and training purposes. Psychotherapy treatment depends on the therapeutic diagnosis of communicative behaviours established in the sessions. Some of these treatment decisions can be made in the sessions, while others can be reviewed in the recorded sessions. The therapist could search for different signs in the conversation, such as acoustic and linguistic cues. This procedure might be costly to do manually and it would be difficult for the therapist to observe some acoustic signs, such as tracking a patient's pitch or speaking rate. Detecting these types of cues using automatic speech processing would be more efficient and accurate. The methods of signal processing and ML algorithms are widely applied to recognise human-centred information based on features extracted from audio signals (acoustic features) (Ringeval et al., 2018; Schuller et al., 2011a; Liang et al., 2019). This section will review the different types of acoustic features discussed in the literature to serve as input for detecting speakers' behaviours in psychotherapy session.

Two common methods are used for extracting features from continuous speech signals: global or local features. Global features (or long-term features) represent the overall minimum, maximum, mean and standard deviation statistics. In some studies, global features are called functionals. Local features (or short-term features) represents the signal temporal dynamic information extracted from each utterance segment to estimate a stationary state.

The local features could be described as Low-Level Descriptors (LLD). The stationary state in the local features has been considered an important factor in Speech Emotion Recognition (SER) where the emotions' labels are dynamic in the speech signal and capturing this change using the local features could enhance the performance of the SER system (Rao et al., 2013).

The local and global acoustic features of any system are examined based on the following features: prosodic features, spectral features and voice quality features. Prosodic features could be identified by the human auditory system, such as rhythm and intonations. The most used prosodic features are the Fundamental frequency (F0), energy and duration. The F0 is established by the vibrations in the vocal cord. F0 generates the tonal and rhythmic characteristics of speech and it is the lowest frequency of the speech signal, perceived as pitch. The energy of the speech signal (intensity) maintains a representation reflecting the amplitude variation of the speech signal over time. Scientists have found that high arousal emotions like anger or happiness return increased energy, while sadness and disgust are expressed with decreased energy. Duration is the amount of time needed to construct words, vowels and related constructs conveyed in speech (Zeng et al., 2008; Lin et al., 2011). The spectral features are extracted by transforming the time domain signal to a frequency-domain signal. The frequency-domain forms a good representation of the vocal tract characteristics determining the shape of the vocal tract and the sound that is produced (Koolagudi and Rao, 2012). The most commonly used spectral features are the Mel Frequency Spectrum Coefficients (MFCCs). The MFCCs are the coefficients acquired by computing a spectrum of the log-magnitude Mel-spectrum of the audio signal. Lower coefficients illustrate the vocal tract filter, while higher coefficients illustrate periodic vocal fold sources (Low et al., 2020). The voice quality features are driven by the physical properties of the vocal tract. There is a strong relationship between voice quality and the emotional content of speech that could be captured using speech properties such as jitter, shimmer, formants and Harmonics-to-Noise Ratio (HNR). Jitter is the variations in the F0 between successive vibratory cycles. On the other hand, shimmer is the change of amplitude. The difference between jitter and shimmer is that the former measures frequency instability while the latter measures amplitude instability. HNR is the amount of the relative level of noise in the frequency spectrum of vowels (Cowie et al., 2001). The formants are the peaks that are determined within the spectrum envelope (Abhang et al., 2016). It is important to highlight that the categorisation of the acoustic features mentioned earlier is the common categorisation used in the literature, although some studies use a different categorisation pattern.

There are a number of acoustic features that were found in the literature to be useful in assessing psychiatric disorders. For Major Depressive Disorder (MDD) or depression for shortness, the acoustic features that were found to contribute significantly to reflect the

monotonous speech often seen in depression were jitter, shimmer and F0. These features were founded to increase with depression severity and psychomotor retardation, for example, slowing of thought, physical movement and reaction times, which could eventually affect motor control precision and laryngeal muscle tension (responsible for controlling sound production) (Horwitz et al., 2013; Kiss and Vicsi, 2017; Quatieri and Malyska, 2012). Several studies found a high increase in the mean F0 in generalised anxiety disorders, in addition to jitter and shimmer, which were found to be high in anxious patients (Özseven et al., 2018; Silber-Varod et al., 2016). There was also evidence that patients with anxiety used more filled pauses and their silent pauses were longer than the healthy control group (Hofmann et al., 1997). A growing body of literature has found an association between different acoustic markers and emotions. Whiteside (1998) reported that anger could be characterised by an increase in mean F0, high frequency energy, falling F0 contours and increased articulation rate, especially in hot anger, while cold anger can be characterised by an increase in mean F0, an increase in high frequency energy and falling F0 contours. The fear was found to be characterised by an increase in high frequency energy and falling F0 contours. On the other hand, sadness was characterised by a decrease in mean F0 and F0 range, a decrease in mean intensity, falling F0 contours and decreased high frequency energy and articulation rate. They found that joy is characterised by an increase in the following: mean F0 and F0 range, F0 variability, mean intensity and in some situations an increase in high frequency energy and articulation rate. Common literature on classifying discrete emotions has highlighted that happiness and anger are close to each other in the prosodic dimensions and therefore classifiers are often confused for one another. This is also the case for sadness and boredom (Yacoub et al., 2003).

While a wide range of speech alterations is attributed to the people's mental states and emotions, the clinical utility of many of the prosodic, voice quality and spectral features previously discussed is potentially limited. Most features were not specifically designed for the task of capturing emotional and mental states in speech and are also sensitive to the phonetic variability within an utterance, as well as differences in speakers' ages, genders, ethnicities and emotional and behavioural attributes (Cummins et al., 2015). According to the literature, there can be no guarantees about the presence and level of speakers' mental states on any one of the speech features discussed. One approach approved by the literature that was adopted to mitigate this is to use large feature sets, for example, those adopted by the Audio/Visual Emotion Challenge (AVEC) baseline systems such as the AVEC 2014 feature set and the eGeMAPS feature set (Valstar et al., 2014, 2013; Ringeval et al., 2018). Another example is the COMPARE feature set adopted by the INTERSPEECH 2013 Computational Paralinguistics Challenge (Schuller et al., 2013). Those feature sets include several acoustic

features related to low-level descriptors (LLDs). Further investigations on the AVEC 2014 and eGeMAPS feature sets will be presented in Chapters 6, 7 and 8.

4.3.3 Automatic Detection of Emotions

As mentioned in Chapter 3, emotions act as one of the human behaviours that could be reflected in the psychotherapy sessions. This section will review the methods used for modelling emotions in the literature and the previous research that adopted those methods to detect emotions.

There are various methods of modelling emotions related to the emotional theories that exist in the literature. The common techniques used in SER are the discrete and dimensional models. Discrete emotion theory defines six categories of basic emotions: sadness, happiness, fear, disgust, anger and surprise (Ekman and Oster, 1979; Ekman et al., 2013). The dimensional emotional model uses several controlled dimensions to represent emotions, such as arousal, valence, control and power (Russell and Mehrabian, 1977; Watson et al., 1988). These dimensions are categorical and universal aspects of emotion. One of the most adopted dimensional models in SER is a two-dimensional model consisting of arousal versus valence. The arousal dimension describes the strength of the felt emotion. It may range from excited to apathetic. On the other hand, the valence dimension illustrates whether an emotion is positive, such as happy and calm, or negative, such as angry and depressed (Nicolaou et al., 2011). The arousal and valence are predicted based on a scale from +1 to -1. Figure 4.1 shows the two dimensional valence and arousal space (Yu et al., 2016). However, some emotions have been considered challenging to predict due to their resemblance to one another, such as fear and anger. Furthermore, some emotions cannot be categorised because they could have positive or negative valence depending on the context such as surprise (Akçay and Oğuz, 2020).

Many studies have investigated the acoustic features in speech to assist in recognising emotions using discrete emotional models. A study was conducted by Crangle et al. (2019) to recognise emotions in natural speech from audio recordings of couples' therapy sessions. Video and audio recordings were collected for three couples, providing over 18 hours of therapy session recordings. The discrete emotions that were detected in the speech were anger, sadness, joy, tension and neutral. The researchers investigated the following acoustic features of speech: energy, frequency and various voice quality features. To classify the four emotions plus neutral, a five class classification problem was introduced with a random forest classification model. The results showed that the energy features gave the best recognition rates for recognising emotions for each couple separately. The recognition for tension is similar for female and male but for the other emotions the recognition results for the female

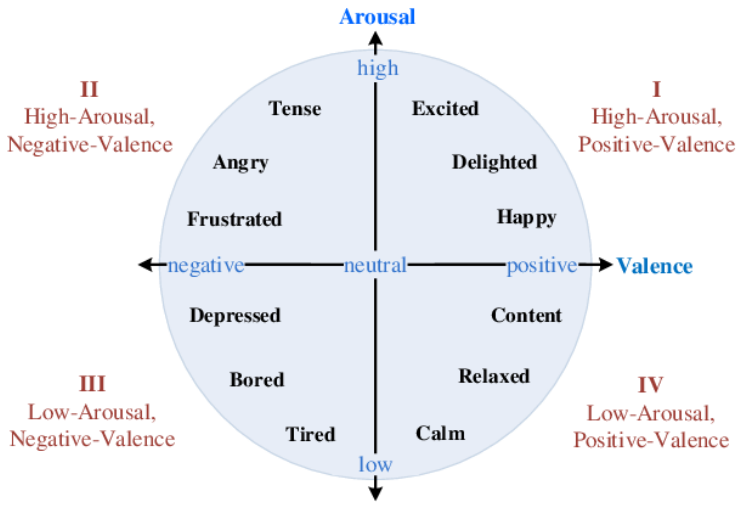


Fig. 4.1 The two dimensional valence and arousal space (Yu et al., 2016)

are lower than those for the male. From their points of view, it would be useful for the therapist to know when an emotion started to accrue in the patient's speech using acoustics because the patient's language does not always express emotions. The results show that the therapy sessions were full of emotions that could be detected in natural speech. A study by (Black et al., 2013) investigated the observational methods for a study of human behaviours in 71 married couple problem-solving interactions. Each couple received a total of 26 sessions of therapy over a one year course. Those dyadic interactions were manually coded with multiple session-level behavioural observations related to the level of acceptance towards the other spouse, level of blame, global positive affect/emotion, global negative affect/emotion, level of sadness and use of humour. They used acoustic features to automatically classify extreme instances for the six selected behavioural codes, for example, low versus high blame. The acoustic features extracted were prosodic, spectral and voice quality features to acquire global acoustic properties for each partner and trained gender-specific and gender independent classifiers. They experimented two linear classifiers: Support Vector Machines (SVM) with linear kernel and logistic regression (LR), and two types of regularisation. The best overall automatic system gained 74.1% in overall classification accuracy, for the wife instances ranged from 67% to 85%, while the best classification performance for the husband instances ranged from 60% to 86%.

Rather than detecting absolute emotion automatically, other studies investigated the dimensional model of emotion. Emotional state can be represented by a specific value for each one of the dimensions (activation, valence and dominance). The Activation dimension is another term used for the arousal dimension. Dominance describes the control level of a person in an emotional experience, either weak, strong or natural (Russell and Mehrabian,

1977). Grimm et al. (2007) analysed the recordings of 47 speakers in a German talk show on TV. The data were labelled by continuous value estimations for valence, activation and dominance. They used acoustic features extracted from the prosody and the spectrum of spontaneous speech signals. The total number of the features was 137; however, they applied feature selection methods such as Sequential Forward Selection (SFS) thereby reducing the number of features to 20. To estimate the emotion primitives, they used SVR, Fuzzy Logic and Fuzzy k-Nearest Neighbour methods using ten folds cross-validation. The SVR using Radial Basis Function (RBF) performed better than the other methods, with correlation coefficients of 0.46, 0.82 and 0.79 for valence, activation and dominance, respectively. This implies that the recognition results for activation and dominance were moderately better than the recognition results for the valence dimension.

4.3.4 Automatic Detection of Empathy and Synchrony

This section will review previous research described in the literature related to the automatic detection of therapist's empathy and the synchrony between patient and therapist, taking into consideration that these come from the important cues related to the proposed system suggested in Chapter 3.

There are several characteristics that therapists should use to enhance the therapeutic progress of patients such as empathy and competency. Empathy could be noticed in the clinic from a person's vocal expressions and language use. A study by Aziz-Zadeh et al. (2010) focused on investigating the role of prosodic ability and its neural processes, especially exploring the relationship between prosody and empathy. They found that the individuals who scored high on the measures of empathy produced more activity in the perception of emotional prosody. To emphasise the relational aspects of psychotherapy communication, a study by Weiste and Peräkylä (2014) considered the prosodic cues of therapist's speech in relation to the prosody of patients. They examined two therapist expressions towards patients, either validating patient's emotions or challenging them. In the validation expressions, the therapists paraphrased the patient's speech about emotions to expose their understanding of the patient's emotions. Based on the study results, there were no differences in the lexical representations of each expression and the therapists named the patient's emotions in common psychological vocabulary. On the other hand, the prosodic cues showed whether the therapist's expressions were intended to validate or challenge. In the validation expressions, the therapist's speech was characterised by *prosodic continuity* in that the prosody of the therapist's formulations continued the prosodic cues of the patient's previous turns. They investigated if the therapists continued the rhythms of the patient's last turns, in a lower and a more quiet voice and with a narrow pitch span. The patients considered that a sign

of approval and permission to go with the feeling. The therapist's expressions to challenge patient's feelings were represented by discontinuation in intonation and rhythm between the parties' turns, the pitch span of the therapist's turns was wider and the therapist's voices were higher and louder than the patient's previous turns. The patients rejected those expressions and considered them as problematic or, in some cases, confirmed them partially (Weiste and Peräkylä, 2014). Thus, past studies indicates the relationship between prosodic cues and empathy.

Many studies have explored automatic methods of detecting empathy in speech, either using acoustic or language features. One of the studies that investigated the acoustic features, was a study by Xiao et al. (2013) that aimed to understand the relationship between empathy and vocal entrainment in drug addiction counselling sessions. They extracted MFCCs and pitch features from the patient's and the therapist's speech. They defined the entrainment through turn-based differences in weighted pitch between speakers. The results showed that the extracted cues significantly correlated with the empathy ratings of humans and confirmed the link between empathy and entrainment. This work was expanded to include prosodic cues and their relationship to empathy in drug addiction counselling. The prosodic features included pitch, energy, jitter, shimmer and utterance duration from the audio signal. The study started by extracting the features, then normalising and quantising the extracted features to determine the distribution of the common prosodic patterns. They used linear SVM as a classifier to classify therapist's empathy. A 75 % accuracy was achieved when classifying the therapist's empathy levels. The results showed that the therapist's empathy was negatively correlated with high pitch and energy. Furthermore, the quantisation step of prosodic features assisted in capturing the silent patterns that would influence the accurate assumptions of empathy (Xiao et al., 2014).

Alternatively, other studies have aimed to investigate empathy from linguistic perspectives. One study proposed a speech system that automated the evaluation of therapist empathy from audio recordings of psychotherapy interactions (Xiao et al., 2016). They built a speech processing system using Voice Activity Detection (VAD) and diarisation modules and an automatic speaker recogniser to extract the therapist's language cues. To categorise the therapist's high versus low empathetic language, they employed maximum likelihood language models to estimate therapy session empathy behavioural codes. They employed Gaussian Mixture Model- Hidden Markav Model (GMM-HMM) and Deep Neural Network (DNN) as classification models. The results showed that the fully automatic system achieved an accuracy of 81% in classifying high versus low empathy in comparison to 86% accuracy using manual transcripts. In the future, they suggested combining language empathy detection with a model built earlier by Xiao et al. (2013) that estimates the vocal similarity between

therapist and patient (Xiao et al., 2016). The previously mentioned study by Xiao et al. (2015a) explored the speech rate entrainment and its association with empathy during dyadic interactions in drug addiction counselling and telephone conversations. The speech rate is the number of words, syllables, or phonemes a human utters in a unit of time that could reflect many human internal states. The degree of entrainment was captured by the averaged absolute differences of turn level speech rates of the therapist and patient, correlated with the therapist's empathy rating. They used linear SVM as the classification model. The results showed that therapist empathy ratings highly correlated to differences in speech rates between therapist and patient and the silence durations. Furthermore, they found that silence and speech duration could correlate to empathy. They combined their model with an earlier, acoustic based model and found that speech rate cues provided complementary information to the previous model.

Many studies have investigated the synchrony between patient and therapist to enhance the relationship between them. One study focused on analysing the language between therapist and patient. It aimed to automatically study language style synchrony with respect to whether the patient and the therapist are using the same function word. They analysed 122 MI transcripts labelled with high and low empathy ratings based on a global rating scale. They used software that counted the semantic content of audio transcripts. They found that sessions rated with high empathetic language had greater language style synchrony than sessions with low empathy ratings (Lord et al., 2015). Other studies have focused on measuring vocal pitch synchrony. A study by Imel et al. (2014) computed the vocal acoustics synchrony of the therapist and standardised patients. Standardised patients are not real patients in that they participated only in conducting the research and were trained to respond flexibly to the therapists. The researchers applied speech signal processing methods to measure the synchrony in the mean of the F0 of speech in 89 MI sessions and its relationship to empathy using predefined codes. They observed an association between vocal synchrony and therapist empathy. However, the study was applied to standardised patients and it would be interesting to test the model on real patients.

It was observed in dyadic interactions that partner behaviour could influence a person. The detection of synchrony in emotions can benefit in detecting emotions in dyadic interactions. Another study aimed to investigate the emotional entrainment and the corresponding impact on the design of emotion recognition systems. The study showed that behaviours from one speaker maintain accompanying information about the emotional state of the other speaker (Mariooryad and Busso, 2013). Other studies have examined people's reactions when interacting with animated characters in a simulated dialogue system, especially in adapting the speaking rate of the system. The results concluded that the speakers adapted

to the speaking rate of the system and later on the speakers did not notice the adaptation (Bell et al., 2003). Lee et al. (2009) proposed a model using a Dynamic Bayesian Network to capture the progress of emotions in conversations using IEMOCAP database. The model was used to capture the conditional dependency between two interacting partner's emotion states in a dialogue using data from expressive dyadic spoken interactions. They aimed to model the dynamics and the cooperative influence of emotion states that could improve the classification of emotions. They focused on the automatic computing of the valence and activation emotion dimensions to continuously characterise the participant's emotion flow. The extracted acoustic features in the study are F0, HNR, energy, speech rate and MFCCs. The results showed that the proposed network improved the classification accuracy comparable to the GMM baseline. Bone et al. (2014) investigated the coupled dynamics of child and psychologist vocal arousal in Autism Spectrum Disorder (ASD) diagnostic interactions. They examined the child-psychologist affective synchrony in relation to ASD severity. Then, they model the arousal dynamics conditioned on other relevant social and conversational features such that the content captured by the model was evaluated through a classification task using maximum likelihood. The resulted model was capable to discriminate between sessions involving children with high and low ASD severity. Furthermore, the results showed that children with higher levels of ASD tended to have their arousal dynamics affected more, considering that the children could not be responsive to the psychologist's affective modulations, which relate to the internal state and external social and conversational factors. Their proposed vocal arousal model captured conversational signal relations, which distinguished between high and low ASD severity.

4.3.5 Automatic Detection of Mental State (Mood)

During the diagnostic process of a patient's mental state, therapists often use several mood outcome measures that could assist in defining the appropriate diagnosis for the patient and the severity level of the patient's mental state. These mood outcome measures are usually represented as questionnaires filled out by the patient. The results from the questionnaires are scores that estimate the severity of the mental state expressed by the patient. Classifying these score levels automatically could assist therapists by saving time and obtaining more accurate results. Furthermore, automatically predicting the scores of the mood outcome measures could enhance the experience of therapy for both the patient and the therapist. For that reason, this section will review previous research on automatic classification and prediction of human mental states, such as depression and anxiety.

Kraepelin (1921) stated that depressed patient's voices tend to be lower in pitch, lower in sound intensity, more monotonous, and lower in speech rate, with more hesitations, stuttering

and whispering. Accordingly, to detect depression, which is a common mental disorder that therapists assess in sessions, several studies have investigated the acoustic properties that might accrue in a depressed patient's speech and their relationship to the clinical subjective ratings of depression. A study by Cannizzaro et al. (2004) examined the acoustic properties of speech samples that were initially collected the administration and scoring of a depression severity score. They relied on early results reported by Ellgring and Scherer (1996) that showed that speech rate and pause duration indicate mood recovery in depressed patients. For that reason, they measured speaking rate, pitch variability and pause time. They found that the quantitative measures of acoustics aligned with the subjective measures of depression and mood state.

It would be relatively easy and acceptable to apply these types of measures in the clinic as it requires only to audio record the sessions, which is already a part of the therapist's responsibilities in therapy. Many studies have investigated the ability to classify speech automatically as a measure for depression based on patient's reported outcome measures. According to Cummins et al. (2015), these studies can be divided into three groups of problems: presence, severity or score level prediction of depression. Presence is considered a detection problem aimed at discovering whether signs of depression can be observed in speech or not. Severity is regarded as a categorical detection of depression based on multiple classes, each relating to a different score category. Score level prediction is the assignment of a continuous depression's score to an unknown speech sample.

Several previous studies have aimed to detect and classify the presence or absence of depression. An investigation by Cummins et al. (2011) intended to study the speech characterisation and speech segment selection methods that would affect the automatic classification of depressed speech. The system started by performing segment selection based on VAD, acoustic features extraction, feature normalisation and modelling of depressed speech. The extracted acoustic feature set consists of speech production, pitch, energy and formants features. Feature normalisation was applied to minimise the mismatch in feature distribution between the different speakers in the data set. They applied a GMM to model depressed speech. The classification accuracy achieved 80% using a combination of MFCC and formant based features. Also, they concluded that basing the analysis on voiced speech segments could be desirable as an initial step in the system. They suggested including more glottal features and improving the normalisation techniques. A study by Helfer et al. (2013) aimed to investigate the importance of vocal tract formant frequency features in the presence of depression in speech. They assessed the performance of GMM and SVM in classifying depression state from the extracted features. The data used was from Depression-severity telephone-based speech recordings. The performance was reported using area under the ROC

curve (AUC). They found that the first three formant frequencies could be used as primary features, while other features related to speech production could improve the classification accuracy.

Other studies have investigated the automatic classification of the severity of depression in speech based on identified classes. A study by Scherer et al. (2013) investigated the ability of an SVM to classify depressed speech based on predefined classes. The used database is the human-human Distress Assessment Interview Corpus (DAIC). They were interested in studying the voice quality features of speech on a breathy-to-tense dimension and the relationship between these features and depression disorders. The severity of depression was detected based on neutral, positive or negative classes of depression. They concluded that there is a difference between depressed participant's and control participant's speech, especially in voice quality. Also, the results showed that SVM could distinguish depression from no-depression with an accuracy of up to 75%.

With respect to score level prediction, there are different depression indicators collected in clinics before sessions start, such as a patient's reported questionnaires represented by the Beck Depression Index (BDI) and the Patient Health Questionnaire (PHQ-9). Several research studies have automatically predicted the score levels of depression. They focused on specific features or feature sets to accomplish the mission of classification. Much of this research has been based on leading challenges focusing on detecting emotions and depression, such as the AVEC Challenge. Each year the challenge baseline has different objectives, but they are mainly centred around identifying emotion and depression. The 2014 edition of the challenge baseline aimed to predict depression levels using BDI estimations for each experiment session. The challenge baseline provided the data for the participants and the depression levels. Furthermore, they described the initial set of acoustic features to be used by the participants, but the participants should upgrade the features to achieve better results. The 2016 version encouraged the participants to estimate the severity of depression using the PHQ-8 depression scores, which is the same as PHQ-9 but without the suicidal question for ethical purposes. The results were represented using RMSE and Mean Absolute Error (MAE). The AVEC 2014 challenge baseline results obtained for the development data partition are 8.934 and 11.521 for MAE and RMSE and 10.036 and 12.567 for MAE and RMSE of the test data partition (Valstar et al., 2014, 2016).

These studies provide an important insight into the efficacy of automating the detection of several human behaviours with the support of various acoustic and language-based (linguistic) features. Furthermore, they highlighted the applicability of exploring the automatic detection of human behaviours in the field of therapy.

4.4 Automatic Tracking of Behavioural and Emotional Cues

Any human internal state can be expressed and declared by behavioural and emotional cues. These cues include verbal and non-verbal expressions of conversational intentions and emotions through vocalisation, intonation, language and others. Analysing and measuring these human behaviours is considered the primary goal of Behavioural Signal Processing (BSP) in that language and speech are fundamental for measuring, modelling and tracking human behaviours (Narayanan and Georgiou, 2013). Tracking human behaviours is one of the aims of this thesis due to the changing nature of those behaviours in psychotherapy sessions. This section will review existing research studies related to tracking human behavioural and emotional cues, as mentioned in the thesis proposed system in Chapter 3.

Of the therapy domains that have frequently been the focus of BSP is couples' therapy due to the frequent amount of and high dynamic behaviours that might accrue from both couples within a session and over the whole journey of therapy. According to Xia et al. (2015), modelling the couples' behaviour at small intervals and studying the effect of the change in behaviour might benefit psychologists and enable them to better understand behavioural mechanisms. The researchers proposed an automatic dynamic behavioural model for capturing couples' behaviours using acoustic features to determine the behavioural state transitions that might eventually model the dyadic interaction. They extracted prosody, spectral and voice quality features. They built a static behaviour model to maintain the optimal frame size to use for the dynamic model. They employed SVM, Fisher Linear Discriminant Analysis (LDA), and Voted Perceptron with the Static Behavioral Models (SBM), and the results showed that Fisher LDA outperformed the other methods, albeit only by a small margin. Also, the results showed that the dynamic model achieved better results than the static model by 10%.

Other studies have investigated the ability to capture couples' behaviours using language features. A study by Tseng et al. (2016) implemented Long Short-Term Memory Recurrent Neural Networks (LSTMs-RNN) to model couples' behaviours, especially in larger context windows through pre-trained word representations. The corpus used in the study consisted of 134 couples from a couples' therapy research project. To improve the results, the researchers fused the results with a frame-level behaviour model using RBF kernel SVR. The obtained results achieved a high correlation with the human annotators' results.

A study by Nasir et al. (2017) explored the ability of acoustic features of distressed couples' speech to assess therapy outcomes. This assessment included the improvement of the couples' relationships and the changing levels of the relationship status. The research

aimed to study each acoustic feature independently and relate to each other as signs for the therapy outcome. The fusion of the machine predicted results and the human-annotated ones were investigated to improve prediction performance. An overview of this work is described in Figure 4.2. Several spectral, prosody and voice quality acoustic features were extracted from the audio recordings. Static and dynamic functionals were considered while extracting the features. The static functionals were extracted for each session per speaker without including any dynamics or any influences of the other speaker. On the other hand, the dynamic functionals were implemented to capture the dynamic patterns of behaviours occupied in speech. Two types of dynamic functionals were included in the study: short term dynamic functionals (turn level analysis to capture the interaction between the couples in one session) and long term dynamic functionals (indicating the change in the marital relationship between two different time intervals before therapy and after therapy, by connecting two sessions together). The researchers used SVM to classify the improvements in the couples' relationships as a binary classification and the level of this relationship as multiple binary classification problems. The results showed that the extracted acoustic features captured more related information than the manual ones constructed by human experts. Furthermore, they found dynamic functionals obtained better results than the static ones for outcome predictions (Nasir et al., 2017).

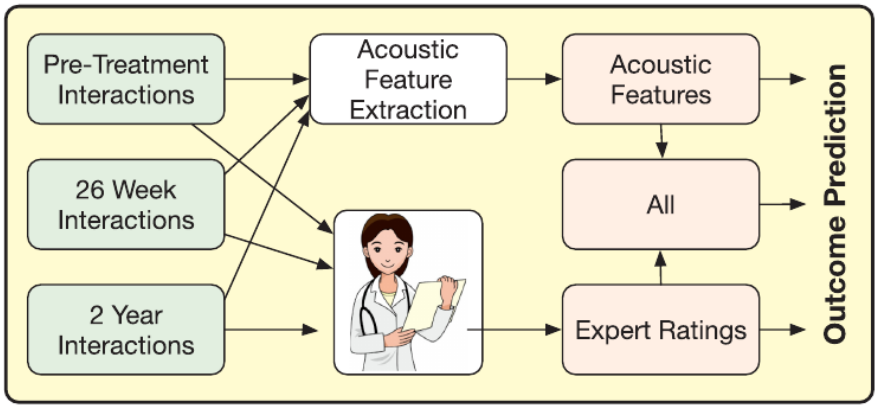


Fig. 4.2 An overview of the study by (Nasir et al., 2017)

A recent study investigated combining prosodic and the lexical features for predicting behaviours in psychotherapy sessions based on utterance levels. They trained LSTMs-RNN based on attention models that empowered models to observe a particular word based on input to predict the resulting classes using manual and automatic word transcriptions. The prosodic features were extracted based on word level. The data were coded for the therapist based on reflection, meaning that the therapist responded to the patient in reflective statements. Also

the therapist was coded by question that implies that the therapist asked either an open or closed question. Other therapist's codes included advice, affirm, confront and other. The patient behaviour was coded from three dimensions: a 'follow-natural turn' meaning there was no indication from the patient towards or away from the goal behaviour change, or it can be coded as 'positive' or 'negative' valence turn depending on the reflection direction processed by the patient toward the goal behaviour change. The results showed that the prosodic features defined information related to the behaviour mission and the integration with the lexical features improved the prediction results (Singla et al., 2018).

4.4.1 Automatic Tracking of Emotional Behaviours

Due to the continuous nature of emotions in time, a number of studies have investigated the continuous prediction of emotions using affect state dimensions. A study by Nicolaou et al. (2011) fused facial expression, shoulder gesture and audio cues to continuously predict emotions in valence and arousal (activation) space using the Sensitive Artificial Listener Database. They extracted multiple acoustic, facial expressions and shoulder features. In the acoustic features, they extracted MFCCs and prosody features, such as the energy of the signal and the pitch. For affect cue predictions, they used SVR and Bidirectional Long Short-Term Memory Neural Networks (BLSTM-NNs). The experiment results (correlation coefficients and Root Mean Square Error (RMSE)) showed that, on average, BLSTM-NNs performed better than the SVR because of the ability of BLSTM-NNs to learn past and future events. Furthermore, they found that arousal can be more easily predicted than valence using audio. The prediction of valence was found to perform better using facial expression and shoulder features. Others investigated the acoustics features to recognise emotions in continuous 3D space (activation, valence and time) using LSTM-RNNs (Wöllmer et al., 2008). They used a database called HUMAINE, which contained 25 audio-visual recordings from four speakers during a natural human-computer conversation. The speakers spoke to four virtual characters, each representing a different emotion. It was labelled continuously in real-time based on valence and activation. The full recordings were divided into utterances using VAD. The acoustic features used for emotion recognition included energy, spectral, pitch, formants and voice quality. The feature extraction phase was performed on each utterance. They modelled emotion history using LSTM-RNNs, which outperformed SVR. The results of recognising the activation dimension using the LSTM-RNNs outreached human performance highlighted in the results of the Recognition Rate (RR) and Mean Square Error (MSE). Additionally, the ability to track emotions in human speech while interacting with a virtual agent was studied by Wollmer et al. (2010). Their work combined the acoustic and linguistic features of speech to recognise emotions in an incremental manner. They

considered the dimensions of speech valence and activation using the HUMAINE database. They used a Bayesian Network (DBN) for linguistic keyword detection and SVM, LSTM and BLSTM-RNNs for modelling context and emotional background to predict the affect state of the speaker. BLSTM-RNNs achieved better F1 scores with 68.9 % for activation and 71.7 % for valence.

Other studies have investigated the ability to track emotions in a cross-cultural environment in which the training and testing sets are from different cultures. One of the AVEC 2018 challenges aimed to generalise the recognition of continuous emotions across different cultures using audio and visual features. This sub-challenge was implemented on a video chat database (SEWA database) advertising a water tap for up to three minutes and designed so that the participants should reveal their reaction and opinion about the advertisement. The video chats were annotated based on the arousal and valence dimensions. The languages introduced in the database were German and Hungarian. The baseline systems used the SVM and LSTM-RNNs models in the challenge. The network was trained on the German language and the generated predictions were made on the Hungarian language. The results presented using the averaged Concordance Correlation Coefficient (total CCC) higlighted that the Facial Action Units (FAUs) features outperformed other features (Ringeval et al., 2018).

Rather than segmenting the databases to utterances for continuous emotion recognition, other studies have segmented the databases based on the flow of body gestures, especially in dyadic performances. A study by Metallinou et al. (2013) examined features that describe a person's body language and speech information. The study aimed to track emotion through activation and arousal values based on body language and speech. The database used is the CreativeIT database, which is a collection of verbal and non-verbal affective interactions in dyadic theoretical performances. They extracted a variety of psychologically based body language features. For the vocal features, they extracted MFCCs, pitch and energy. To track emotions, they used the GMM and LSTM-RNNs at the frame and window levels. The results showed that the GMM outperformed the LSTM-RNNs given the same features for most of activation and dominance tasks, while the LSTM-RNNs outperformed the baseline results on both dimensions.

To conclude, human behaviours are essential components of any dyadic interaction. Tracking these behaviours could ensure a better understanding of the underlying internal states for each participant and the correlation between them. This section reviewed research studies that explored the procedure of detecting and tracking human behaviours automatically in interactions using acoustic and linguistic features from a therapeutic perspective as presented in Table 4.2. Although those studies explored the automatic detecting and tracking of a single

human behaviour in therapy, it is interesting to investigate those behaviours together in a fully automatic system to understand the interactions between them and provide the therapists with a full overview of the behavioural interactions in the session.

Due to the limitations in the thesis's main dataset (the THEPS dataset) as mentioned in Section 3.3, each patient has only one session reordered in the dataset, which limits the tracking of behaviours to within sessions and not across sessions.

4.5 Databases Used in This Domain

Table 4.4 lists all the databases used in the literature from the thesis domain based on the human behaviours mentioned in this chapter and Chapter 2, such as emotions, empathy, synchrony and alliance rupture. The table illustrates the databases' names with brief descriptions of the numbers and labels of each database, taking into consideration their availability. The table is organised based on the human behaviours reviewed in this chapter that relate directly to the proposed system in Chapter 3. The two databases from the table used in this thesis are the AVEC 2014 database described in Section 4.5.1 and the REmote COLlaborative and Affective interaction (RECOLA) database described in Section 4.5.2. They are used as benchmark databases for validating the baseline systems implemented in Chapters 6 and 7 for detecting depression and anxiety using mood outcome measures and predicting dimensional emotions. Furthermore, a number of corpora have become standard when training ASR systems, One of those is LibriSpeech described in Section 4.5.3.

An in-house dataset, the THEPS dataset, described in Section 3.3, has been used for evaluating the proposed automatic system due to the suitability of this dataset for the purpose of this project and the ethical challenges related to obtaining a database for psychotherapy session recordings. Some of the mentioned databases related to therapeutic factors are not available due to ethical permissions. In order to achieve the aims of this thesis, it is also important to obtain databases labelled with therapist's competence and alliance rupture. As far as is known, those types of databases are considered sensitive and limited to the research authors and there are no datasets that fit this requirement.

4.5.1 AVEC 2014 Database

The AVEC 2014 database, used in the challenge, was originally created for the 2013 challenge round (Valstar et al., 2013). This dataset consists of 150 videos (240 hours) of task-oriented depression data recorded in a computer interaction context using a microphone and a webcam.

Behaviours	Author (citation)	ML Method	Database	Evaluation	Results
Emotions	Crangle et al. (2019)	Random forest classification model	Couples therapy sessions labelled with four emotions database.	Accuracy and confusion matrix	The recognition for tension is similar for female and male but for the other emotions the recognition results for the female are lower than those for the male.
	Black et al. (2013)	SVM with linear kernel and LR and two types of regularisation	Married couple problem-solving interactions database	Average Accuracy	The best classification performance for the wife instances ranged from 67% to 85% , while the best classification performance for the husband instances ranged from 60% to 86%.
	Grimm et al. (2007)	SVR, Fuzzy Logic and Fuzzy k-Nearest Neighbour methods	German talk show on TV database	Correlation coefficients	The SVR using Radial Basis Function (RBF) performed better than the other methods, with correlation coefficients of 0.46, 0.82 and 0.79 for valence, activation and dominance, respectively.
	Xia et al. (2015)	SVM, Fisher LDA, and Voted Perceptron with the SBM	Couples' therapy sessions database	Accuracy	Fisher LDA outperformed the other methods, albeit only by a small margin.
	Nicolaou et al. (2011)	SVR and BLSTM-NNs	Sensitive Artificial Listener database	Correlation coefficients and RMSE	BLSTM-NNs performed better than the SVR because of the ability of BLSTM-NNs to learn past and future events.
	Wöllmer et al. (2008)	SVR and LSTM-RNNs	HUMAINE database	RR and MSE	The LSTM-RNNs outperformed SVR such that the results of recognising the activation dimension using the LSTM-RNNs outreached human performance.
	Wollmer et al. (2010)	DBN, SVM, LSTM and BLSTM-RNNs	HUMAINE database	Accuracy, precision, recall, and F1 measure	The BLSTM-RNNs achieved better F1 scores with 68.9 % for activation and 71.7 % for valence.
	Ringeval et al. (2018)	SVM and LSTM-RNNs	SEWA database	Averaged CCC	The results presented that the Facial Action Units (FAUs) features outperformed other features.
	Metallinou et al. (2013)	GMM and LSTM-RNNs	CreativeIT database	Correlation Coefficients	The results showed that the GMM outperformed the LSTM-RNNs given the same features for most of activation and dominance tasks, while the LSTM-RNNs outperformed the baseline results on both dimensions.
Empathy and Synchrony	Xiao et al. (2013)	SVM	Drug addiction counselling sessions database	Accuracy	A 75 % accuracy was achieved when classifying the therapist's empathy levels.
	Xiao et al. (2016)	GMM-HMM and DNN	Audio recordings of psychotherapy interactions database	Accuracy	The results showed that the fully automatic system achieved an accuracy of 81% in classifying high versus low empathy in comparison to 86% accuracy using manual transcripts.
	Xiao et al. (2015a)	linear SVM	drug addiction counselling and telephone conversations database	Accuracy	The prediction system based on human transcriptions yielded a slight increase in prediction accuracy (85.0%) compared to the system based on the automatic transcriptions with an accuracy of (82.0%).
	Lee et al. (2009)	Dynamic Bayesian Network and GMM	IEMOCAP database	Accuracy	The results showed that the proposed network improved the classification accuracy comparable to the GMM baseline.
	Bone et al. (2014)	Maximum likelihood model	ASD diagnostic interactions database	Cross-correlation	The resulted model was capable to discriminate between sessions involving children with high and low ASD severity.
Mental State	Cummins et al. (2011)	GMM model	Depressed patients' recordings database	Accuracy	The classification accuracy achieved 80% using a combination of MFCC and formant based features.
	Helfer et al. (2013)	GMM and SVM	Depression-severity telephone-based speech recordings database	AUC	They found that the first three formant frequencies could be used as primary features, while other features related to speech production could improve the classification accuracy.
	Scherer et al. (2013)	SVM	DAIC database	Accuracy	The results showed that SVM could distinguish depression from no-depression with an accuracy of up to 75%.
	Valstar et al. (2014)	SVR	AVEC 2014 database	RMSE and MAE	The results obtained for the development data partition are 8.934 and 11.521 for MAE and RMSE and 10.036 and 12.567 for MAE and RMSE of the test data partition.

Table 4.2 Research studies explored detecting and tracking human behaviours automatically in therapeutic interactions.

The total number of participants is 84 and there is only one person in each recording. Each participant has been recorded up to four times with a two-week period between each recording. The lengths of the recordings varied from 20 to 50 minutes. The participants performed different human-computer interacted tasks guided by a PowerPoint presentation. In the 2014 challenge, there were two main tasks to perform. The first asks the particpant to read aloud an excerpt of the fable 'The north wind and the sun' spoken in German stated as 'Northwind'. For the second task, the participant is asked to respond to several questions such as 'Discuss a sad childhood memory' in German stated as 'Freeform'. The challenge database was organised into three equal subsets: Training, Development and Testing. The age, gender and depression levels were considered when distributing the data over the subsets. The depression level was labelled in each recording using the Beck Depression Inventory (BDI) questionnaire. As reported in the AVEC 2014 challenge for depression classification, the use of the Northwind data was ignored because the classification results using the Freeform data showed superior performance. Thus, the validation was implemented only using the Freeform data (Valstar et al., 2014).

4.5.2 RECOLA Database

The RECOLA database was designed in collaboration with a team in informatics and psychology at the Université de Fribourg, Switzerland (Ringeval et al., 2013). It was recorded in order to study the socio-affective behaviours from multimodal data. The recordings have been emotionally annotated continuously for the dimensions arousal and valence using six annotators for every 0.4 seconds. The AVEC 2018 challenge baseline used a partition of the RECOLA database using 27 speakers, each talking for five minutes. The dataset was divided into training, development and test sets, having nine speaker recordings and a total of 45 minutes in each partition. The test set labels could not be obtained because they were hidden for challenge baseline purposes and have not been made publicly available.

4.5.3 LibriSpeech Database

LibriSpeech is a database of read English speech, which is famous for training and evaluating speech recognition systems. It consists of 1000 hours of audiobooks speech available to download. It was demonstrated that models trained with the Librispeech database achieve better on the standard Wall Street Journal (WSJ) test sets than models built on WSJ (Panayotov et al., 2015).

4.6 Machine Learning Related Background

Based on Samuel (1967) machine learning (ML) is the field of study that promotes computers the ability to learn without being expressly programmed. ML algorithms can be organised based on whether they are trained with supervision into supervised, unsupervised, semi-supervised, and reinforcement learning approaches. In *supervised learning*, the training data provided to the ML algorithms contain the desired solutions, and are called labels. A typical supervised learning task is classification, and a disease diagnostic system is a good example of classification which is trained with many variables or features along with their class. In this thesis, the classification of depression and anxiety levels described in Chapter 6 and the classification of competency levels described in Chapter 8 are examples of classification tasks. Another conventional task is to predict a target numeric value, such as depression and anxiety scores, given a set of features called predictors, as described in Chapter 6. Furthermore, competency scores have been predicated as described in Chapter 8. This type of task is called regression. The supervised ML algorithms commonly used for both classification and regression includes k-nearest neighbour (k-NN), naïve Bayes classifiers, SVMs, neural networks, decision trees, random forests, linear regression, and logistic regression.

In contrast, in *unsupervised learning*, the training dataset is unlabeled. The main tasks using unsupervised learning are clustering such as k-means and fuzzy c-means, and feature reduction algorithms such as Recursive Feature Elimination Cross-Validation (RFECV). RFECV aims to find the optimal set of features by fitting over the training folds and selecting the features that produce the smallest averaged error across all folds.

Semi-supervised learning is a class of supervised learning tasks and techniques that also employ unlabeled data for training such that the training dataset includes both unlabeled and labeled data. *Reinforcement learning* involves an agent that observes the environment in order to select and perform actions, and obtain rewards in return. Afterwards, the agent learns by itself what is the best strategy to get the most rewards over time. The following section focus on the supervised learning methods. Pecht and Kang (2018) described more details about ML methods.

4.6.1 Supervised Machine Learning Algorithms

In this section, the fundamentals of widely used supervised learning techniques for classification and regression are covered. Similarly, because deep learning algorithms have become increasingly popular for feature learning in various medical applications, this section provides the fundamentals of DNN, as the state of the art. Most of the mentioned approaches

are used for the experimental work in this thesis or cited in the related work sections. Further motivation for their use is presented in the experimental chapters.

K-Nearest Neighbour (KNN)

The KNN algorithm can be deployed for both regression problems and classification problems. It is based on a very simple concept of identifying K-nearest neighbours for a given data point based on the assumption that similar items are "closer" to each other. The concept of close is implemented as a distance measure that can be a simple Euclidean distance between two data points. The algorithm start by searching for the largest class of items that are close to the test data, then the investigated test data is considered to belong to that class. Actually, there is no learning involved in the algorithm, other than keeping track of all labeled data. For regression problems, the average of the label values for these K-nearest neighbours is considered, while, for classification problems, the majority is taken instead of the average (Churiwala, 2019).

Support Vector Machine (SVM)

SVMs are a supervised learning algorithm adopted for resolving binary classification and regression problems. If it is used in regression problems, it is called Support Vector Regression (SVR). The main idea of SVMs is to establish a hyperplane by which the margin of separation between the two classes is maximized. First, each of the data points in this algorithm is plotted as a data point in n-dimensional feature space (F). Then a hyperplane is constructed that maximizes the separation between two classes (Churiwala, 2019). The hyperplane can be represented as follows:

$$w^T x_i + b \tag{4.1}$$

where $w \in F$ is a weight vector, x is the input vector and b is a bias value. The hyperplane determines the margin between the classes such that the data instances for a class given the output of x is y $(y_i = -1)$ are on one side, whilst data instances for another class are on the other $(y_i = 1)$:

$$\begin{cases} w^T x_i + b \leqslant -1, & \text{if } y_i = -1 \\ w^T x_i + b \geqslant 1, & \text{if } y_i = 1 \end{cases}$$

Naïve Bayes

Naïve Bayes, is a supervised learning algorithm originated from the Bayes' theorem. It assumes independence between every pair of input instances as handled in the naïve assumption. The Naïve Bayes equation is:

$$\hat{y} = \underset{y}{argmax}\, P(y) \prod_{i=1}^{m} p(x_i \mid y) \tag{4.2}$$

To estimate P(y) and $P(x_i \mid y)$, a maximum a posteriori estimation method can be used such that P(y) is then the relative frequency of class y in the training dataset (Pecht and Kang, 2018).

Decision Trees

A Decision Tree is a machine learning model constructed using a series of decisions based on variable values to take one path or the other. A tree is an acyclic directed data structure with nodes and edges which connect nodes. It is a tree with nodes representing deterministic decisions derived from variables and edges representing the path to next node or a leaf node based on the decision. The leaf node or terminal node of the tree characterises a class label as output of prediction. The goal of the Decision Tree is to establish a model from training data by learning decision rules to predict class or value of target variable. Decision Tree creation is manageable and can be easily constructed for a small number of decision items by hand. The Decision Tree also can be constructed using bagging technique with the rules extracted from the large amount of data (Churiwala, 2019).

Random Forest

Random Forests are effective and easily constructed supervised learning models used in classification and regression problems. They are based on underlying Decision Tree structures. It is a collection of Decision Trees which improve the prediction over a single Decision Tree. It is a way of averaging multiple Decision Trees built from different parts of the training set with the aim of reducing over fitting by a single Decision Tree. It grows many Decision Trees by sampling from the input dataset

$$D$$

for a set of features instead of the training data samples. The process is also called feature bagging. At the end of bagging, the final results are

$$B$$

Decision Trees, each tree with reduced dataset and reduced feature set. For a new data point, the Random Forest prediction of its value is an aggregation of predicted values from the corresponding constructed trees. The prediction can be categorical or regression based on the Decision Tree construction (Churiwala, 2019). The final result is calculated based on the following equation:

$$f = \frac{1}{B}\sum_{n=1}^{B}(f_n(x)) \tag{4.3}$$

Where:
f is the final prediction of the Random Forest.
B is the number of trees constructed from the dataset.
n is the index of the Decision Tree constructed.
f_n is the result of the Decision Tree n.
x is the input sample that is in the test set, for which prediction is desired.

Boosting: AdaBoost

Adaptive boosting (AdaBoost) is a practical boosting algorithm and aims to transform a set of weak classifiers into a strong classifier sequentially. A training dataset containing

$$m$$

input-target pairs x_i, y_i is given, where x_i is an n-dimensional input instance (or feature vector) and $y_i = \{1, 1\}$. On each round, t = 1, 2, ..., T, a distribution D_t is computed over m training instances, and a given base weak classifier is applied to find a hypothesis $H_t : x \to -1, 1$, where the aim of the weak classifier is to find a weak hypothesis with low weighted error ε_t relative to D_t. The final hypothesis H computes the sign of a weighted combination of weak hypotheses:

$$H(x) = \sum_{t=1}^{T}\alpha_t H_t(x) \tag{4.4}$$

$H(x)$ is calculated as a weighted majority vote of the weak prediction H_t where each is assigned weight α_t (Pecht and Kang, 2018).

Linear Regression

Linear Regression supposes that the relationship between the variables can be expressed as a linear function such that for a single input function, this is a straight line. It is possible to convert nonlinear relationships into a pattern that allows Linear Regression to be constructed. A linear regression model develop a prediction $\hat{y}$ by simply calculating a weighted sum of the predictors $\{x1, x2, ..., x_n\}$, plus a constant called the bias b. The model can be mathematically expressed by:

$$\hat{y} = w_1 x_1 + w_2 x_2 + ... + w_n x_n + b = \sum_{i=1}^{n} w_i x_i + b \tag{4.5}$$

where $w = \{w1, w2, ..., w_n\}$ is a weight vector (Pecht and Kang, 2018).

Ridge Regression

Ridge regression is a regularised version of linear regression where adding the regularisation term to the linear regression forces ridge regression to fit the data and keep the model weights as small as possible. The cost function of ridge regression is defined as follows:

$$J_{RR}(w) = J_{LR}(w) + \alpha \sum_{i}^{n} w_i^2 \tag{4.6}$$

where $J_{RR}(w)$ is the cost function of linear regression and α controls how much is wanted to regularise the model. If $\alpha = 0$, ridge regression would be just linear regression (Pecht and Kang, 2018).

Lasso Regression

Least Absolute Shrinkage and Selection Operation (LASSO) regression is considered another regularised version of linear regression. LASSO regression depends on the l1 norm of the weight vector instead of the l2 norm in ridge regression as described:

$$J_{LASSO}(w) = J_{LR}(w) + \alpha \sum_{i}^{n} |w_i| \tag{4.7}$$

The regularisation concept in LASSO regression leads to penalising values which cause some of the parameter estimates to turn out exactly zero. The difference between previously mentioned regression techniques and LASSO regression is that it eliminates the weights of the least important predictors, such as features (Pecht and Kang, 2018).

Elastic Net Regression

Elastic net regression sits between ridge and LASSO regression such that the regularisation concept is a simple mix of both ridge and LASSO's regularisation terms (Pecht and Kang, 2018). It is controlled by the mix ratio γ as follows:

$$J_{ENR}(w) = J_{LR}(w) + \gamma\alpha\sum_{i}^{n}|w_i| + (1-\gamma)\alpha\sum_{i}^{n}w_i^2 \tag{4.8}$$

Deep Neural Network (DNN)

Deep learning has shown significant success with applications such as speech recognition, image processing, and language translation. It refers to neural networks with many layers and many neurons. The availability of computing power and a large amount of data made these large structures very effective in learning hidden features and data patterns. There are several types of DNN such as Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNN) (Churiwala, 2019).

Convolutional Neural Network (CNN)

A convolutional neural network is a group of deep learning methods that have become prominent in several computer tasks and is attracting attention across various domains, including emotion detection. It comprises multiple elementary units, that is convolution layers, pooling layers, and fully connected layers, which are designed to automatically and adaptively learn spatial hierarchies of features through a back propagation strategy. The convolution and pooling layers perform feature extraction, while the fully connected layer maps the extracted features into the final output, such as classification. The convolution layer plays a crucial role in CNN, composed of a stack of mathematical operations, such as convolution, a specialized type of linear operation (Yamashita et al., 2018).

Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are models of neural networks that are a dominant approach for processing sequential data. A series or sequence of data can be observed in multiple problem areas with time-series data or a sequence of data that has repetitive properties. An example of a sequence domain is speech recognition which requires modelling time series data with individual speech patterns. It solves the problem of sequence modelling by providing the ability to recall past data and process the current vector based on the inputs before or after the present vector in the sequence. They have been successful in solving

Table 4.3 A confusion matrix.

		Predicted	
		Positive	Negative
Actual	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

context recognition and repetitive pattern recognition problems. For example, in natural language processing, the recognition of word context requires looking at other words in the sentence (Churiwala, 2019).

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are presented to solve the problem of long-term dependency modelling in RNNs. LSTM networks address the problem of fading gradients by designing a memory cell approach and providing gates that control the impact of the previous sequence on the current state and output of the current sequence. LSTM networks have potently modelled NLP problems with perfect long-range dependency modelling. Language modelling for translating one language into another or suggesting the following word while typing and speech recognition is an examples of the applications of LSTM (Churiwala, 2019).

4.6.2 Performance Metrics

This section primarily reviews performance metrics used in data-driven diagnostics related to medical fields. Most of the mentioned metrics are used for the experimental work in this thesis or cited in the related work sections.

Confusion Matrix

To assess the performance of a classification model on a test dataset for which the true values (or classes) are known, a confusion matrix as presented in Table 4.3 is widely used. The confusion matrix consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The TP is the case that a test instance in the positive class is correctly identified as the positive class, TN is the case that a test instance in the negative class is correctly identified as the negative class, FP is the case that a test instance associated to the negative class is incorrectly recognised as the positive class, and FN is the case that a test instance associated to the positive class is incorrectly assigned to the negative class, respectively (Pecht and Kang, 2018).

Accuracy, Precision, and Recall

The common performance measures for medical diagnostics include accuracy, sensitivity (or recall), and specificity. These measures are calculated based on the number of TPs, TNs, FPs, and FNs, and defined as :

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{4.9}$$

$$Sensitivity(orrecall, ortruepositiverate) = \frac{TP}{(TP+FN)} \tag{4.10}$$

$$Specificity = \frac{TN}{(TN+FN)} \tag{4.11}$$

Accuracy calculates the proportion of true TP or TN assessments in a population. Sensitivity measures the proportion of TPs, the percentage of positive instances that are correctly identified as positive. Specificity measures the proportion of TNs, which is the percentage of negative instances correctly identified as negative. Both sensitivity and specificity are commonly used with accuracy as diagnostic metrics (Pecht and Kang, 2018).

Receiver Operating Characteristics (ROCs)

To assess classification performance, particularly for a binary classification problem, a famous method is Receiver Operating Characteristics (ROCs) analysis using the true positive rate (TPR), which also called sensitivity, against the False Positive Rate (FPR), where FPR can be measured as:

$$FPR(1-specificity) = \frac{FP}{FP+TN} \tag{4.12}$$

All possible combinations of TPR and FPR consist of an ROC space which is a location of a point in the ROC space can show the balance between sensitivity and specificity. For example, the increase in sensitivity is accompanied by a decrease in specificity. The location of the point in the space can demonstrate whether the binary classifier performs accurately or not.

The area under the receiver operating characteristic curve that is also known as AUC can be calculated to provide a way to measure the accuracy of a classifier:

$$AUC = \int_0^1 ROC(t)dt \tag{4.13}$$

such that t equals FPR, and ROC(t) is TPR. The larger the AUC, the more accurate is the classifier (Pecht and Kang, 2018).

Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)

The common prognostic metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The MAE is a quantity used to calculate how close the estimated performance degradation trend $\hat{y}$ or estimates is to the actual performance degradation trend y (Pecht and Kang, 2018), defined by:

$$MAE = \frac{1}{(t_{EOP} - t_p + 1)} \sum_{t=t_p}^{t_{EOP}} |\hat{y}(t) - y(t)| \tag{4.14}$$

The MSE known as the mean squared deviation is a measure of the average of the squares of the errors or deviations which is the difference between $\hat{y}$ and y expressed as:

$$MSE = \frac{1}{(t_{EOP} - t_p + 1)} \sum_{t=t_p}^{t_{EOP}} (\hat{y}(t) - y(t))^2 \tag{4.15}$$

The RMSE which is also called the root mean squared deviation which is a measure of the differences between the values predicted by a prediction model and the values actually observed, calculated as:

$$RMSE = \sqrt{MSE} \tag{4.16}$$

Correlation Analysis

Correlation analysis is a statistical approach to evaluate the power of the relationship between two numerical continuous variables. If correlation is established between two variables, it means that when there is a systematic change in one variable, there is also a systematic change in the other such that positive correlation exists if one variable increases at the same time with the other. While, negative correlation exists if one variable decreases when the other increases. The most commonly approach used in correlation analysis is the Pearson correlation coefficient (PCC). The PCC quantifies the direction and strength of the linear association between two variables such that the resulting value is between +1 and -1. The sign and magnitude of the PCC specifies the direction and strength of the correlation, respectively. Given two datasets $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_m\}$, each of which contains m observations, the PCC can be measured as:

$$PCC = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}} \tag{4.17}$$

Behaviours	Name	Description	Availability
Emotions	Stanford Suppes Brain Lab Psychotherapy database (Crangle et al., 2019)	Couples therapy sessions labelled with four emotions (Anger, Sadness, Joy and Tension).	Not available
	Vera am Mittag (VAM) corpus (Grimm et al., 2007)	Recordings from 47 speakers in German TV talk show annotated with three continuous values for emotion estimation valence, activation and dominance).	Available
	Couple therapy corpus (Nasir et al., 2017)	Recorded audio-visual interactions for couples in therapy sessions annotated with manually specified behaviours and marital outcome measures.	Not available
	Sensitive Artificial Listener Database (Nicolaou et al., 2011)	Audiovisual interactions between a human and an operator acting in four personalities. The data annotated continuously with arousal and valence.	Available
	HUMAINE database (Wöllmer et al., 2008; Wollmer et al., 2010)	25 recordings from 4 speakers interacted in human-computer conversations to let the users revel different emotional states. The data annotated with valence and activation every 10 ms.	Available
	SEWA database (Ringeval et al., 2018)	Audiovisual recordings of participants watching a set of commercials and discussing that in up to three minutes. The database continuously labelled with arousal, valence and liking dimensions.	Available
	CreativeIT (Metallinou et al., 2013)	A collection of verbal and non-verbal affective interactions in dyadic theoretical performances. The verbal interactions labelled continuously with activation, valence and dominance.	Available
	RECOLA database (Ringeval et al., 2013)	A multimodel database devoted for spontaneous collaborative and affective interactions. It consists of 46 speakers recorded in dyads during a video conference while completing a required task.	Available
Empathy and Synchrony	MI in drug addition counseling (Xiao et al., 2013, 2014, 2015a, 2016)	Recorded sessions for MI of changing addictive behaviours. The sessions annotated based on the global rating of empathy for the therapist in each session.	Not available
	IEMOCAP database (Mariooryad and Busso, 2013; Lee et al., 2009)	Audiovisual database for expressive acted human interactions in dyadic sessions. It is labelled with specific emotions for each participant based on dialogue turns.	Available
Therapeutic Alliance	Psychoanalytic Therapy Over Time (Safran, 2009)	It is six sessions of psychoanalytical therapy with a young depressed woman. The therapist emphasis on the therapeutic relationship.	Not available
	Resolving Therapeutic Impasses (Safran and Muran, 2006)	It is a training sessions to guide the therapist to deal with therapeutic impasses by demonstrating different therapeutic tools labelled with therapist comments about the patient impasses. Each session also labelled with a specific type of rupture.	Not available
Mental Disorders	EMFACS (Cummins et al., 2011)	The database consists of 23 depressed and 24 control participants. It was initially recorded as an audiovisual database for measuring the facial activity in depressed people.	Not available
	Mundt database (Helfer et al., 2013)	The database was collected for depression severity study which are based on telephone recordings of 35 participants. The database was labelled with depression outcome measure.	Not available
	Distress Assessment Interview Corpus (DAIC) (Scherer et al., 2013)	In the corpus, the participants interacted with a virtual agent in semi-structured manner. Each participant completed a series of questionnaires such as PHQ-9.	Not available
	AVEC 2014 database (Valstar et al., 2014)	It is audiovisual depression corpus recorded in human computer interaction scenario. The participants performed two tasks: reading an excerpt and describing an emotional experience. The data is labelled with depression outcome measure.	Available

Table 4.4 databases used in the research domain

4.7 Summary

This chapter reviewed the previous research relating to the system proposed in Section 3.2 and for each module proposed in the system. These modules are based on the patients' and therapists' behaviours that have been found to correlate positively with the therapeutic alliance as discussed in Chapter 2. Furthermore, previous studies that have investigated the automatic detection or tracking of those behaviours were reviewed in this chapter. Detecting and tracking those behaviours in a full automatic system could assist therapists in comprehending a bigger picture about those behaviours and how they are interacting and relating to the therapeutic alliance in the session. Furthermore, it would help therapists in making better treatment choices for patients that could eventually minimise the drop-out rate and enhance treatment outcomes. The review in this chapter showed that most of these studies concentrated on one human behaviour rather than investigating several behaviours that could relate to each other in therapy. In addition, there were studies that focused on investigating full automatic systems for the detection of human behaviours or therapeutic attributes, especially for the purpose of studying the therapeutic alliance. For that reason, this thesis will investigate a full automatic system that can detect and track patients' and therapists' behaviours to reveal a positive therapeutic alliance in the session. The next chapter will investigate the first step towards achieving this proposed system, which is the ASR system for conversational psychotherapy sessions.

Chapter 5

Automatic Speech Recognition for Conversational Psychotherapy Sessions

The previous chapter reviewed the previous research relating to the thesis's proposed system in Chapter 3. Chapter 3 outlined the blocks for implementing a full system for the automatic analysis of the THEPS dataset. Furthermore, it described the psychotherapy sessions' environment and the reasons for collecting the sessions' recordings. To start implementing the first block of the proposed system, the development of an Automatic Speech Recognition (ASR) system for the conversational psychotherapy sessions (THEPS dataset) is described in this chapter.

This chapter describes the ASR system, including the challenges, the previous research on the related systems, the data used for training the system, the architecture of the ASR, the experiment involved in conducting the system and the obtained results.

5.1 Introduction

ASR is a fundamental part of the overall proposed system. Speech recognition is considered a particular form of pattern recognition. The optimal goal of ASR is to find the word sequence based on the input speech signal (Wang, 2020). Conventional ASR systems consist of the following main parts: feature extraction, acoustic modelling, language modelling and decoding (finding the most likely string of words). The feature extraction step transforms the input waveform into a sequence of acoustic features vectors such that each one represents the information in a specific time window of the signal. The acoustic model measures the likelihood of the observed spectral feature vectors given linguistic units such as words and phones. The language model calculates probability estimations for a word sequence using

the most common language models, the N-gram models. The N-gram model assumes that the probability of a word depends on the previous N words. The probabilities of an N-gram model is typically trained on the training set of the manually transcribed audio corpora. Then, this trained model is used to compute probabilities on the test set (Jurafsky and Martin, 2013). Recently, Deep Neural Networks (DNN) have replaced the traditional Gaussian Mixture Model (GMM) for the likelihood evaluation while keeping all the aforementioned ASR components as hybrid ASR systems. The speech community has seen increased studies investigating moving from hybrid modelling to End-to-End (E2E) modelling. In E2E models, a speech sequence is translated to a token sequence using a single network (Li, 2021). However, the performance of E2E ASR systems highly depends on the availability of training data. When it comes to data scarcity problems in low-resource scenarios, E2E ASR systems need more training data than hybrid ASR systems to achieve similar Word Error Rates (WERs) and this relates to the tendency of E2E systems to overfit the training data (Gakuto et al., 2019; Wang et al., 2021).

Several challenges in the THEPS dataset could make it difficult for ASR to achieve a high performance. As mentioned in Section 3.3, the sessions may contain environmental noise disturbance in the background of the patient and therapist conversations, such as door slams, keyboard typing and sounds from an open window. Also, an echo noise may occur in the sessions' recordings due to the use of unprofessional recording equipment in some sessions. The distance between the speakers and the microphone could also affect the recognition performance, as it may mean capturing only part of the speakers' speech, or it could pick up only the reflected sound in the room.

Additionally, challenges that could influence the recognition performance are the speaker characteristics, such as the patient's emotions and the therapist's empathy. These speakers' behaviours could influence their acoustics, yielding a variation on the extracted feature vectors compared to their acoustics recorded when the speakers are speaking with more neutral emotions. Due to the Coronavirus Disease 2019 (Covid-19) pandemic, some sessions were recorded using a mobile phone. The recognition of mobile phone sessions could differ from that of face-to-face sessions due to bandwidth limitations, a lack of control over the speaker's environment and handset noise (De Wet et al., 2006). In many cases, overlapping speech in psychotherapy sessions has been noticed, such as when the therapist tries to maintain engagement with the patient through verbal communication (back-channelling). It is commonly known that ASR systems perform poorly when recognising overlapped speech (Yoshioka et al., 2018). Finally, the most common challenge in speech recognition for the medical domain is data sparsity and untranscribed audio data, which could limit the prediction performance of the Machine Learning (ML) algorithms (Kang et al., 2020). Despite the

need for manually transcribed corpora corresponding to the audio data in ASR, transcribing audio data manually is generally expensive, time-consuming and labour-intensive (Bada et al., 2017).

This chapter outlines how a conventional ASR system has been built for the THEPS dataset considering the data sparsity and untranscribed audio data difficulties. Speaker diarisation and Voice Activity Detection (VAD) are outside the scope of this project. The rest of this chapter is organised as follows: Section 5.2 presents the related research concerning ASR and the common methods used in the instance of sparse data, Section 5.3 explains the data partition used from THEPS dataset for this chapter, Section 5.4 defines the architecture used for the ASR system, Section 5.5 presents the experiments and the results conducted in this chapter and Section 5.6 summarises the chapter's findings.

5.2 Related Work

Training the ASR system using the THEPS dataset could suffer from limited training data (data sparsity) and minimal transcribed audio sessions. Achieving a high performance ASR system requires a sufficient quantity and quality of training data (Mustafa et al., 2014). Recent advances in ASR research addressing these challenges have demonstrated that transfer learning of DNNs can be a feasible solution to enhance the performance of the trained model. The transfer learning framework was selected to transfer the already gained knowledge from the source domain or task to the target domain or task (Pan and Yang, 2009). The transfer learning mechanism could be described as follows: a specific number of trained neural network layers are duplicated from a pre-trained source model. A particular number of new layers are defined and appended to the old layers and updated according to the target model. Using a different method, the original layers (one or many) can be fine-tuned based on the target model (Meyer, 2019). Several studies have demonstrated the efficiency of the transfer learning in sparse datasets (Kunze et al., 2017; Matassoni et al., 2018; Mirheidari et al., 2021, 2020; Yi et al., 2018). Taking into consideration the sparse amounts of data used in those studies, it could be helpful to investigate transfer learning as a method for enhancing the ASR system's performance on the THEPS dataset.

A number of corpora have become standard when training ASR systems. One of those is LibriSpeech which consists of 1000 hours of audiobooks speech available to download. Previous research has established that the acoustic models trained on LibriSpeech as a source model achieved lower WER in comparison to other corpora (**?**). Ghahremani et al. (2017) conducted an experiment relating to transfer learning from the LibriSpeech corpus to the Wall Street Journal (WSJ) that includes 80 hours. They noticed improvements in the WER using

the transfer learning method compared to the baseline trained directly on WSJ. Furthermore, the inclusion of more speakers involved in the training improved the results.

The medical field usually suffers from the scarcity and poor quality of recorded data due to difficulty collecting recordings and also occasionally the effects of a health disorder on a patient's language. A study by Gale et al. (2019) explored improving ASR systems for children with autism and language impairment. Due to the scarcity of data in their research, the authors investigated how to implement transfer learning to enhance ASR system performance. They pre-trained a Time-Delay Neural Network (TDNN) on the LibriSpeech corpora. Then, they used transfer learning to adapt the LibriSpeech model to the target data. The best WER was obtained when combining the training with recordings of first and second-grade children. The age match in the training data profoundly impacted the training data for children's ASR. They found out that data scarcity and domain specificity influenced the overall model performance.

Transfer learning was also investigated by Huang et al. (2020) for several ASR tasks, such as when dealing with multiple English accents, languages and application-specific domains. In the different accents experiment, they trained the model on several language corpora, such as LibriSpeech and WSJ. They achieved higher accuracy from the transfer learning models than the model trained from scratch. Furthermore, the large source data sped up the convergence and produced better performance results. Mirheidari et al. (2020, 2021) developed an automatic system for detecting dementia by analysing patients' speech and language while they spoke to an Intelligent Virtual Agent (IVA). They demonstrated improvements in their ASR system by applying transfer learning on a base TDNN acoustic model trained on the LibriSpeech corpus using a 10 fold cross-validation approach. They used a four-gram language model with the Turing smoothing interpolation technique along with a language model from the LibriSpeech corpus.

The probability smoothing is another language modelling technique that can address the problem of zero probabilities assigned to events that were unseen in the training data. This enhance the probability mass over more events, yielding more smooth probabilities distribution by adjusting low probabilities upward and high probabilities downward (Chen and Goodman, 1999). Adopting a technique such as probability smoothing in the THEPS ASR pipeline could improve the transfer learning results when using a source model trained on LibriSpeech, especially since the LibriSpeech consists of 1000 hours of recordings.

The other challenge posed by the THEPS dataset is minimal transcribed audio sessions. Due to the effort and time required to transcribe the databases manually, several studies have explored techniques for using untranscribed data. One such technique is semi-supervised training (or self-training). Semi-supervised training has been shown to improve ASR per-

formance in low resource settings and several acoustic conditions (Lamel et al., 2002; Ma et al., 2006; Novotney and Schwartz, 2009; Zavaliagkos et al., 1998). The common practice in semi-supervised training is to utilise a small amount of real transcribed training speech utterances to build a basic acoustic model. Then, the acoustic model is employed to generate automatic transcripts for the remaining un-transcribed training speech utterances. The real transcriptions and the automatically generated ones are used together to estimate a better acoustic model. To improve the accuracy of the training speech, the literature proposed confidence-based filters to automatically discriminate the low confident speech utterances for semi-supervised acoustic models (Lo and Chen, 2019). A study by Thomas et al. (2013) used semi-supervised training on multilingual data for training neural networks and acoustic models. They observed a 16% improvement in the word recognition accuracy using a low-resource setting, such that one hour is used for transcribed training data. Another study by Biswas et al. (2019) investigated the South African language pairs. They considered the effectiveness of semi-supervised training to increase the size of the minimal acoustic training sets. Using around 11 hours of un-transcribed speech, the system reduced the WER from the baseline systems. However, Semi-supervised training is out of scope of this study.

5.3 Data

The dataset used to establish the ASR system is described in Section 3.3. Table 5.1 shows the patient demographics and therapy session information of the transcribed part of the dataset using rounded values. In-person sessions denote face-to-face sessions that were not conducted over mobile phones. For clarification, the speaker turns are referred to as speech *segments*.

The dataset has been partitioned into training and test sets based on the segments (utterances). The training set represented 80% of the data, with 7417 utterances and the test set represented 20% of the data, with 1854 utterances. The number of sessions used for training is 43 sessions with around 1307 minutes, while 11 sessions with around 327 minutes have been used for testing. According to the literature, it is challenging to automatically recognise overlapping speech (Renals and Swietojanski, 2017). The data included in Table 5.1 exclude overlapping speech. For this reason, the overlapped speech has been eliminated from the dataset for the ASR experiment. Furthermore, the time alignments for each segment's start and end times have been reviewed and adjusted manually by the thesis's author to produce more accurate timestamps for each segment in the dataset and thus achieve more accurate results for the ASR system.

Table 5.1 Patient demographics and therapy session information for the transcribed dataset

Patient demographics	**Total (all sessions)**	**Average**	**Min**	**Max**
Number of patients	54	-	-	-
Female	39%	-	-	-
Age	-	37	16	74
In-person sessions	34%	-	-	-
Session information				
Length (min)	1634	30	20	52
Number of words	264069	4890	354	5462
Number of segments	9271	172	35	194
Time talking per session (min)				
Patient	644	12	3	22
Therapist	927	18	8	28
Words spoken per session (N)				
Patient	104981	1944	354	3997
Therapist	159088	2946	1351	5462
Number of segments per session (N)				
Patient	4526	84	35	193
Therapist	4745	88	36	194

5.4 ASR Architecture

Kaldi is an open-source speech recognition toolkit developed in C++. Kaldi provides several recipes for creating an ASR system by utilising famous databases such as LibriSpeech and WSJ (Povey et al., 2011). ASR consists of several components and Kaldi treats each component as an independent module. Kaldi is a pipeline system that joins all the ASR modules together to perform speech recognition. These modules consist of feature extraction, acoustic modelling, language modelling and Kaldi decoding. The complete structure of the Kaldi ASR pipeline is described in Figure 5.1. The feature extractor module extracts the speech features from the speech signal. The main extracted features are Mel filterbanks and Mel Frequency Cepstral Coefficients (MFCCs). They are extracted from window frames of 25 ms with a shift of 10 ms. The feature extractor also extracts 100-dimensional i-vectors capturing speaker and environment information (noise). As described in Section 5.1, the language model investigates the word sequences to match them with phrases that have logical meanings. They measure the probability of certain words in a larger context of many former or latter words. Kaldi's training and decoding algorithm depends on Weighted Finite State Transducers (WFSTs). WFSTs are used to prepare a well-studied graph operation that can be

effective in acoustic modelling. The Kaldi decoding phase is implemented using a decoding graph constructed from the WFSTs graph.

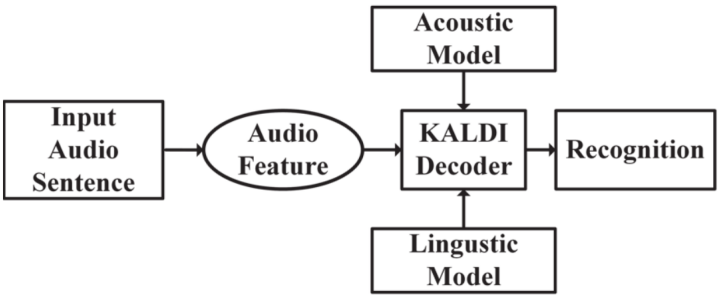


Fig. 5.1 Kaldi ASR structure pipeline (Upadhyaya et al., 2017)

The acoustic modelling estimates each signal sequence's probability described by the extracted speech features for the corresponding sequence of words. It provides frameworks for acoustic modelling based on a Hidden Markov Model (HMM). HMM is a statistical model representing the evolution of observable events depending on internal factors (Yoon, 2009). HMM is used to model sequential speech characteristics where each state in the HMM corresponds to a sub-phonetic unit. The HMM output probability density is modelled using a GMM or a DNN. The GMM and DNN models offer posterior probabilities for all the possible sub-phonetic units related to an acoustic frame. The acoustic model is also a pipeline process starting with a basic HMM-GMM acoustic model where each HMM state model a context-independent phone. The resulting model has been used to force align the training set to roughly estimate the next model's phone boundaries. This model is then used to train the context-independent phone (triphone) model. The triphone model is used to attain a better forced alignment for the training set used for training the next complex model. The next HMM-GMM acoustic model has been created by applying Linear Discriminate Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT). LDA and MLLT are feature transformation techniques applied to the input features as a preparation procedure for the next model for adaptive speaker training. The next acoustic model performs speaker adaptive training using the previously created model. The forced alignment process is performed after each acoustic modelling process to gain high-quality alignments for the next GMM or DNN training process (Georgescu et al., 2019; Upadhyaya et al., 2017). Each triphone step in the HMM-GMM training includes the number of Gaussian and number of leaves. The number of Gaussian is the number of mixture models that the training is aiming to achieve. The number of leaves is the number of leaf nodes aimed for in the process of state tying (Kullmann, 2016).

The DNN based model used in Kaldi is TDNN which is a specific type of artificial neural network which mainly deals with sequential or speech data. It was first introduced for phoneme recognition (Shaharin et al., 2014; Waibel et al., 1989). The TDNN used two types of features: 40-dimensional MFCCs extracted from 25 ms frames with 10 ms shift and 100-dimensional i-vectors calculated from 150 consecutive frames. The i-vectors have shown their usefulness for discriminative adaptation for neural networks due to their ability to capture both speaker and environment-related information (Peddinti et al., 2015). The architecture of the TDNN is described in Peddinti et al. (2015).

5.5 Experiments and Results

The LibriSpeech recipe in the Kaldi tool has been adopted for the training of the ASR system. The use of commercially available systems such as cloud-based ASR services (Google Cloud Speech-to-Text) is not applicable due to the confidentiality of the THEPS dataset. The recipe enhances the acoustic model training by introducing further acoustic models. The subsequent acoustic model selects the most similar state in the previous pipeline model by judging the similarities based on the overlap of counts in tree stats to minimise the time of training. The language model has been built using the SRI Language Modeling (SRILM) toolkit as advised in Kaldi (Povey et al., 2011). Kaldi requires specific data preparation for the sessions' recordings and transcriptions. For that reason, several data processing stages were implemented to create the files required by the Kaldi toolkit. This includes reviewing the alignment start and end times for each segment, checking for any faults in the transcriptions' syntax, introducing Kaldi reserved words, transforming the text to upper case and removing all the punctuation marks from the text. The reserved words used in Kaldi for this experiment are <UNK> for unknown words; <NOISE> for the background noise, such as door slam; and <SPOKEN _NOISE> for non-verbal vocalisation, such as laugh and cry.

5.5.1 Transfer Learning

There are three major transfer learning methods described in the literature: weight transfer, multi-task training and domain assimilation. The weight transfer method configures the parameters (weights) of the target DNN model by the learned parameters of the source model, based on the idea that the low-level features of the source and the target data are similar. It has been demonstrated that weight transfer can empirically improve the performance given a small amount of the target data and this can also speed up the training process due to the small number of parameters that require fine tuning (Ghahremani et al., 2017). In multi-task

training, the shared hidden layers of the DNN model are learned by several tasks at the same time to benefit from the correlated information shared through tasks (Heigold et al., 2013; Sahraeian and Van Compernolle, 2016). Unlike the other two methods, domain assimilation selectively employs a subset of the target data closer to the target domain. Although this method is popular in the image processing field, it has been rarely explored in the ASR field (Gale et al., 2019; Ge and Yu, 2017).

The main idea of the weight transfer method in Kaldi is that the internal layers of DNN learn intermediate level representations of input, which can be pre-trained on one dataset (or task) and re-used on the other tasks. The method is implemented in Kaldi by first training the model on a large dataset, retaining only n layers and adding new task-specific adaptation layers over those. The typical way is to do a two-stage training by freezing the transferred layers and training task-specific layers in the 1^{st} training stage and then fine-tune the whole network in the 2^{nd} stage of training using a smaller learning rate. However, a more efficient way could be implemented in Kaldi is a single-stage training by training the transferred layers with a lower learning rate while training the task-specific layers with a larger learning rate (Ghahremani et al., 2017).

The THEPS ASR system has been implemented using the 54 manually transcribed sessions described in the data section using the LibriSpeech recipe, including HMM-GMM, DNN and transfer learning. Using the weight transfer method described earlier, a transfer learning approach has been applied to adapt the LibriSpeech model to the THEPS dataset. Following the LibriSpeech recipe, the LibriSpeech corpus has been used to train a base TDNN acoustic model. The neural network has 4 hidden layers with p-norm input dimension of 3000 and group size of 10. High-dimensional MFCCs features have been extracted from the THEPS dataset and 100-dimensional i-vectors calculated based on the same model trained for the LibriSpeech model. The LibriSpeech model has been applied to generate frame-level acoustic alignments for training and the weight have been copied from the LibriSpeech model to initialise the target model. The acoustic model has been adapted to the acoustics of the THEPS dataset such that both structure and weights have been transferred from the training of the LibriSpeech corpus. Afterwards, the training stage was conducted using two epochs on the training set. The language model used for the transfer learning approach has been trained as 4-gram with Turing smoothing interpolated with the 4-gram language model from the LibriSpeech corpus following the technique used by Mirheidari et al. (2020, 2021), who reported that the best performance was gained with 60% weight for the training set and 40% weight for the LibriSpeech language model. Further system validation has been implemented using K-fold cross-validation such that the THEPS dataset has been organised into six equal folds, each containing nine sessions, on account of the total dataset size. The micro-average

WER has been estimated for the cross-validation considering the number of word imbalances between the different folds. Table 5.2 presents the results of the different ASR systems using 54 sessions for the HMM-GMM, DNN, transfer learning and cross-validation.

Table 5.2 ASR system results using transfer learning and cross-validation techniques

System	Number of sessions	Train	Test	Cross-Validation	Adapt/Transfer	%WER
HMM-GMM	54	THEPS	THEPS			67.18
DNN	54	LibriSpeech	THEPS			70.30
DNN	54	THEPS	THEPS			47.18
DNN	54	THEPS	THEPS		LibriSpeech model + THEPS dataset	35.86
HMM-GMM	54	THEPS	THEPS	6 folds		65.49
DNN	54	THEPS	THEPS	6 folds		45.29
DNN	54	THEPS	THEPS	6 folds	LibriSpeech model + THEPS dataset	33.88

The results gained after applying the adaptation method using an acoustic model trained on the LibriSpeech corpus and adapted to the THEPS dataset with a total number of 54 sessions is a WER of 35.86%. The cross-validation technique gained an average absolute improvement of 1.85% with WERs of 65.49%, 45.29% and 33.88% on the HMM-GMM, TDNN and transfer learning systems, respectively. The final results gained after performing the transfer learning is comparable with what has been achieved in the literature. Xiao et al. (2015b) reported a mean WER of 43.1% after implementing an ASR for Motivational Interviewing (MI) sessions for drug and alcohol counselling using the same tool used in this chapter (Kaldi). Chen et al. (2021) developed an automatic pipeline that transcribed Cognitive Behaviour Therapy (CBT) sessions' recordings using Kaldi and gained a WER of 44.01%. Further discussion on those findings will be introduced in Chapter 9.

5.6 Summary

This chapter investigated the development of an ASR system for the THEPS dataset, which has shown promising results toward achieving the study goals. However, there are several concerns about the THEPS dataset from its sparsity to the limited transcriptions available for the data. Using the Kaldi LibriSpeech recipe, several ASR systems have been investigated including the use of transfer learning and cross-validation techniques to improve the results gained in the initial training phase. The results gained from transfer learning using a model trained on the LibriSpeech corpus and adapted to the THEPS dataset showed better results compared to the initial system's results. The final automatic transcriptions results will be

used in Chapter 8 to investigate the therapist's competence using language-based features. The next chapter will investigate the automatic methods for detecting depression and anxiety using mood outcome measures, which is one of the modules in this thesis's proposed system.

Chapter 6

Automatic Detection of Depression and Anxiety

The previous chapter investigated the development of an automatic speech recognition system (ASR) for the THEPS dataset as a first step toward achieving the system proposed in Chapter 3. The proposed system aims to explore how to automatically analyse psychotherapy sessions by detecting patients' and therapists' behaviours. One of the patient's behaviours of interest is the patient's mood, where the related outcome measures, Patient Health Questionnaire (PHQ-9) and Generalised Anxiety Disorder (GAD-7) are available within the THEPS dataset. Automating the procedure of predicting these outcome measures could assist therapists in identifying a more suitable treatment plan and producing more efficient therapy for the patient. The aim of this chapter is to establish a system for predicting mood outcome measures and classifying the depression and anxiety levels based on their corresponding outcome measures. The first step toward this aim is to set up and evaluate a system using recordings of consultations for people with Dementia or Dementia-related symptoms (referred to as the Dementia Database; it will be further described below). The patients in the Dementia Database were asked to fill out the questionnaires related to the mood outcome measures, which makes this database applicable for evaluating this chapter's main system. Furthermore, this could ensure a more efficient system and explore different diagnostic phenomenons. The THEPS database, as described in Section 3.3, consists of recordings for conversational psychotherapy sessions for anxiety disorders treatments. The patients in the THEPS dataset are also required to complete mood outcome questionnaires, such as those for depression and anxiety outcome measures. This chapter starts by implementing and validating a reliable baseline for automatic classification and prediction of depression and anxiety outcome measures. Then, the baseline system is explored further and evaluated on the databases

mentioned earlier. This chapter will concentrate on the acoustic features and the automatic transcription outputs from Chapter 5 will not be used in this chapter.

6.1 Introduction

As described in Section 2.1, the most commonly used mood outcome measures for depression and anxiety are PHQ-9 and GAD-7. Those scores are calculated from descriptive diagnosis manual questionnaires filled out by the patients before each session. Both THEPS and the Dementia Database have accompanying PHQ-9 and GAD-7 scores. This chapter explores to what degree it is possible to automatically predict these outcome measures in a person's speech. In particular, a system has been developed using an in-house dataset (Dementia Database) and this system is then further evaluated on the THEPS dataset.

To start building a system for the automatic detection of depression and anxiety, the Audio/Visual Emotion Challenge (AVEC) 2014 challenge baseline explored in Section 4.3.5 has been re-implemented for further evaluation using a publicly available database provided by the challenge (the AVEC 2014 Database). This included the extraction of the AVEC 2014 feature set provided in the challenge. Then, the baseline system performance has been explored using a database of recordings of people with Dementia or Dementia-related symptoms. The Dementia database is useful for detecting depression and anxiety because it is labelled with depression and anxiety outcome measures. Due to the size differences between the AVEC 2014 challenge database (240 hours) and the Dementia database (around 32 hours), some techniques have been investigated for cross-validation. The use of an enormous number of features in the AVEC 2014 (2268 feature) led to the use of the feature selection method. One of the challenges that emerged when applying feature selection and cross-validation methods was that there were no common selected features between the folds, which led to the need for advanced methods in Machine Learning (ML) to resolve this matter. The results gained from the Dementia database has been explored based on the diagnosis categories related to Dementia as a classification and regression problems. Due to the sparsity of the data and the potential use of maximised features in the THEPS dataset, the enhanced system has been used for classifying and predicting depression and anxiety outcome measures. Parameter optimisation has been introduced to the system to improve the results. The obtained results have been predicted and classified based on the depression and anxiety scores and score levels.

This chapter proceeds as follows, Section 6.2 describes the related work on the detection of depression and anxiety. Section 6.3 describes the standard baseline system for depression and anxiety score prediction and classification. Section 6.4 describes the validation process

of the baseline system on the AVEC 2014 database using the provided feature set. Section 6.5 illustrates the evaluation process on the Dementia database and a walk through the steps of the system implementation. Section 6.6 provides a further implementation of the enhanced pipeline on the THEPS dataset. Section 6.7 outlines the chapter's findings.

6.2 Related Work

Based on the related research review presented in Section 4.3.5, previous studies investigating the ability to classify speech automatically as a measure for depression based on mood outcome measures can be divided into three groups of problems: depression presence, severity or score level prediction. This chapter focuses on score level prediction and severity of depression. Various studies have assessed the efficacy of using acoustic features for depression and anxiety classification and score level prediction. A study by Scherer et al. (2013) investigated the voice quality characteristics, especially the breathy to tense dimension, to predict depression and Post-Traumatic Stress Disorder (PTSD). The corpus used in the study involved interacting with a virtual human in a controlled scenario. Participants completed a series of questionnaires, including PHQ-9 and the PTSD Checklist-Civilian version (PCL-C), after a short explanation of the study. They used a Support Vector Machine (SVM) with leave-one-speaker-out to classify depression and PTSD based on the polarities of specific affective questions, such as positive, negative and neutral. The results showed that participants with depression and PTSD conditions showed more tense voice features compared with participants without PTSD. Furthermore, SVM could distinguish depression from no-depression with an accuracy of up to 75% and PTSD from no-PTSD with an accuracy of 72.09%. Based on the relationship between depression and the neurophysiological changes in the brain that could disturb articulatory precision in speech production, a study by Helfer et al. (2013) investigated the vocal tract formant frequencies and their dynamic features from sustained vowels and conversational speech. They used a dataset of 35 speakers labelled with the 17-item Hamilton depression rating scale. Systems based on SVMs and Gaussian Mixture Models (GMMs) were evaluated using a cross-validation approach through the analysis of vowels and free responses. The SVMs performed better than the GMMs with an Area Under the Curve (AUC) of 0.76 for the fusion of vowels and free-response features.

In 2013, the first AVEC challenge was released to predict the score of a single depression indicator for each recording in the dataset. It provided the participants with the depression database and depression related feature set and as a result, speech-based depression score prediction gained greater attention (Cohn et al., 2018; Valstar et al., 2013). The AVEC multi-model depression severity prediction challenge baseline used a combination of multivariate

acoustic features and a Support Vector Regressor (SVR) at the back-end (Valstar et al., 2013, 2014). Various studies have assessed the efficacy of the AVEC audio baseline feature set in a range of depression-related approaches. Lee et al. (2021) proposed an approach for how the AVEC 2013 baseline feature set could be employed in a voice-based screening test for Major Depressive Disorder (MDD) for male and female elderly Koreans when reading a series of mood-inducing sentences. The dataset was labelled with the Korean version of the Geriatric Depression Scale scores. They employed the AdaBoost classifier as a decision tool for male and female participants, separately. The results showed that the acoustic features needed to discriminate between participants with MDD and healthy participants differed between males and females. They found that the discriminative features for males were spectral and energy-related acoustic features, while for females were prosody related features. The current chapter has not investigate gender-related features due to the sparse dataset.

In a study investigating the prediction of Parkinson's Disease (PD) severity, Tracy et al. (2020) reported that the employment of voice signals in conventional ML models could be used to distinguish participants with PD who exhibit little to no symptoms (mild PD) from healthy controls. They extracted the AVEC 2013 baseline feature set along with other acoustic feature sets devoted to clinical speech analysis. Several ML models were selected for classification, such as L2-regularised logistic regression, random forest and gradient boosted decision trees. Cross-validation was implemented along with hyper-parameter tuning. The gradient boosted model performed better than the other model with a recall of 0.646, f1-score of 0.688 and precision of 0.849. According to the authors, these results showed that voice features and ML approaches can be used as an early detection indicator for PD.

In a more related study to the topic, Tasnim and Stroulia (2019) used the AVEC 2013 dataset to explore the automatic detection of the prevalence and severity of depression from acoustic features using different classification and regression algorithms. The study experimented on the AVEC 2013 baseline feature set along with Principal Component Analysis (PCA) to reduce the dimensionality of the features, resulting in 791 features. They selected random forest, SVM, a gradient boosting tree and a Deep Neural Network (DNN) as the training models in the study. The results showed that the random forest algorithm performed best on the AVEC 2013 dataset, outperforming the other models with a Mean Absolute Error (MAE) of 8.21. Another relevant study by Özkanca et al. (2018) merged the German AVEC 2013 database with a Turkish database to select better features and improve depression assessment accuracy. They applied the minimum redundancy maximum relevance feature selection method on the AVEC 2013 baseline feature set before using SVR to model the relationship between the depression scores and the features using the leave-one-out feature method. They found that larger multilingual databases improved the

model performance and enhanced the model's ability to detect the similarities between two entirely distinct languages.

It is worth mentioning that the AVEC 2013 and AVEC 2014 baseline feature sets are identical, but the AVEC 2014 database is a subset of the AVEC 2013 database. The AVEC 2013 database contains 340 recordings, while the AVEC 2014 database contains 150 recordings (Valstar et al., 2013, 2014). From the studies reviewed here, it is evident that applying the AVEC 2013 feature set in depression score prediction systems could be suitable for predicting the depression severity and scores in combination with various feature reduction methods. Several systems have adopted the techniques of parameter optimisation (tuning) and cross-validation to regulate and evaluate the model training. Several classifiers and regressors, such as SVM, SVR, gradient boosting tree and random forest, have achieved noticeable performance in predicting and classifying depression in speech.

Depression and anxiety can be comorbid with several other illnesses, such as bipolar disorder, pseudo-dementia, parenting stress and dysthymia (Cairney et al., 2008; Devanand et al., 1996; Leigh and Milgrom, 2008; Moscati et al., 2016; Perlis, 2005). Dementia is a collection of related symptoms concerning the loss of cognitive ability, such as logic, memory and language. Due to the temporary cognitive decline caused by depression, confusion may accrue in classifying patients with Dementia. *Pseudo-dementia* is a condition that resembles Dementia in cognitive impairment but appears due to depression. A study by Sumali et al. (2020) investigated the ability to differentiate between dementia patients and depression patients. Furthermore, they examined the possibility of automated pseudo-dementia screening using acoustic features. They extracted three acoustic features: Mel Frequency Spectrum Coefficients (MFCCs), Harmonics-to-Noise Ratio (HNR) and Zero-Crossing Rate (ZCR). An SVR with several kernels, decision trees and boosted trees was explored as the ML models in the research. The results showed that the MFCC features were positively correlated with Dementia scores and showed an inverse correlation with depression scores, highlighting that their system could be effective as a tool for screening pseudo-dementia.

Bipolar disorder can be recognised by the presence of recurring manic or hypomanic episodes alternating with depressive episodes (Carvalho et al., 2020). A study by Yang et al. (2017) investigated approaches to distinguish between depression and bipolar disorder due to differences in the intensity and duration of mood between the two disorders. An SVM classifier was built via extracted facial and acoustic features. They extracted the acoustic features, energy, ZCR, Fundamental frequency (F0), HNR and MFCCs using the openSMILE tool. A Coupled Hidden Markov Model (HMM) multimodal fusion approach was proposed to model contextual information based on temporal changes in speech and facial responses. The results showed that the final proposed model was feasible for mood disorder detection. A

study by Parlato-Oliveira et al. (2021) explored *prenatal motherese*, which is when pregnant women speak to their unborn baby using *motherese* (a specific form of speech used by parents or caregivers to address their infants). They also investigated the relationship between prenatal motherese and maternal depression status. They used acoustic features related to the F0, energy and duration of the speech along with the SVM classifier. Prenatal motherese was found to be correlated with depression scores which means the more future mothers were depressed, the less they spoke to their fetuses during pregnancy.

Dysthymia is the insidious onset of less severe depression. The symptoms of Dysthymia that overlap with depression are excessive guilt, loss of interest and problems in concentrating (Hirschfeld, 1994). Additionally, depression can be comorbid with anxiety, characterised by a more chronic course, greater likelihood of relapse and stronger suicidal tendencies (Rohde et al., 1991). Wang et al. (2017) examined the effects of acoustic features on distinguishing the comorbidities of anxiety disorder and Dysthymia from pure depression. The authors extracted acoustic features from the patients' speech to identify if the patient had a particular comorbidity or not. They extracted 26 acoustic features, such as intensity, loudness, ZCR, voicing probability, F0 and 12 MFCCs, using the openSMILE tool. For classification, they implemented SVM with 5-fold cross-validation to train and test the model. The results indicated that anxiety disorder and Dysthymia can be effectively distinguished from depression using acoustic features. Together, these studies provide important insights into the usefulness of applying acoustic biomarkers in classification tasks to detect depression and anxiety that can be comorbid with other disorders.

6.3 Depression/Anxiety Score Prediction and Classification

Figure 6.1 illustrates the block diagram of the system for detecting depression and anxiety using outcome measures such as PHQ-9 and GAD-7. The system starts by preprocessing the data samples depending on each database and detecting the voiced segments from the samples using Voice Activity Detection (VAD) or manual annotations has been applied if two speakers are found in the database recordings. For the THEPS and Dementia databases, manual annotations have been used to separate the patient's segments from the therapist's or neurologist's segments. After that, feature extraction is applied to the patient speech segments because the mood outcome measures only need to be detected for the patient. In some cases, there is a need for feature selection approaches to select the most correlated features. Finally, a classifier or regressor is built to classify or predict the depression/anxiety score.

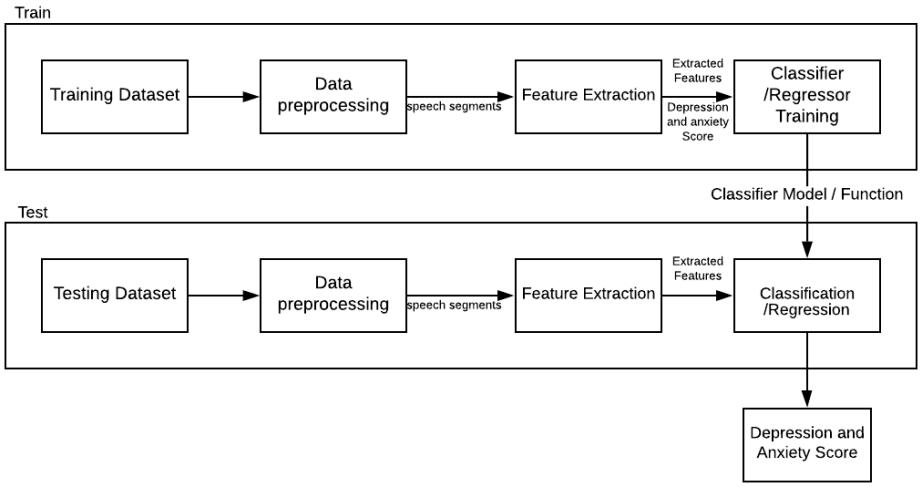


Fig. 6.1 The block diagram of the system

6.4 Validating the System on the AVEC 2014 Database

In order to conduct the experiment of predicting the mood outcome measures on the Dementia database, the AVEC 2014 challenge baseline system has been re-implemented first to validate the feature extraction and classifier pipeline setup. The data and the feature set used in this validation both came from the AVEC 2014 challenge baseline. The AVEC 2014 database used in the challenge is described in Section 4.5.1.

Challenge Feature Set

In this chapter, the AVEC 2014 feature set has been chosen for the system implementation. The non-verbal activity of people with depression is generally considered by clinicians as an indicator of depression (Hall et al., 1995). Clinicians found that there is a high correlation between the prosodic, articulation and acoustic features of depressive patients' speech (Flint et al., 1993; Nilsonne, 1988). As mentioned in Section 4.3.5, early studies that investigated the speech of depressed patients found that the listener could observe a change in the pitch, loudness, speaking rate and articulation of the speech (Darby and Hollien, 1977). Several acoustic features in patients' speech, such as prosodic, source, formants and spectral features, have been found to be positively correlated with depression. The AVEC 2014 challenge baseline used a specific acoustic feature set related to depression (2268 features total), which is publicly available for the participants to use. The OpenSMILE toolkit (Eyben et al., 2010;

Valstar et al., 2014) is amended to enable the extraction of the challenge feature set; the feature set is outlined in Table 6.1.

Table 6.1 The AVEC 2014 feature set (Valstar et al., 2014)

Energy & spectral (32)
loudness (auditory model based); zero crossing rate; energy in bands from 250-650 Hz, 1 kHz-4kHz; 25%, 50%, 75% and 90% spectral roll-off points; spectral flux; entropy; variance; skewness; kurtosis; psychoacousitc sharpness; harmonicity; flatness; MFCC 1-16
Voicing related (6)
F_0 (sub-harmonic summation, followed by Viterbi smoothing), probability of voicing; jitter, shimmer (local), jitter (delta:"jitter of jitter"); logarithmic Harmonics-to-Noise Ratio (logHNR)

Challenge Baseline and Results

The challenge consists of two sub-challenges: detection of emotions according to three affective dimensions (Arousal, Dominance and Valance) and depression recognition according to a depression score for each recording. In the depression sub-challenge, SVR was implemented with no feature selection and/or parameter optimisation applied. The results has been reported using MAE and Root Mean Squared Error (RMSE). The experiment presented in this chapter re-implemented the challenge baseline for further evaluation to build the depression and anxiety prediction system. Table 6.2 shows the AVEC 2014 challenge's reported results and the replicated results using the same depression sub-challenge database and conditions.

Table 6.2 Comparison between the AVEC 2014 challenge's reported results and the replicated results

Partition	**MAE**	**RMSE**
Development (AVEC results)	8.934	11.521
Development (Replicated results)	8.881	11.539
Test (AVEC results)	10.036	12.567
Test (Replicated results)	10.173	13.953

It is clear that there is not much difference between the replicated results and the reported ones and this difference may have been caused by configurations of the implemented systems.

Several classifiers have been investigated to classify the level of depression in the AVEC 2014 database. As mentioned earlier, the used depression labels in the AVEC 2014 database are the BDI such that the scores range from 0 to 63. Ranges can be organised as follows:

0–13: indicates no or minimal depression, 14–19: indicates mild depression, 20–28: indicates moderate depression, and 29–63: shows severe depression. Table 6.3 presents the classification results for the depression score ranges of the AVEC 2014 database using the following classifiers: SVM, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier and Gradient Boosting Classifier. It is clear that SVM achieved better performance results than thew other classifiers.

Table 6.3 Depression score ranges (BDI) classification results for the AVEC 2014 database using several classifier models

Classifier	**Accuracy**	**Precision**	**Recall**
SVM	59.33%	0.42	0.40
Decision Tree	58.00%	0.41	0.39
Random Forest	53.77%	0.39	0.37
AdaBoost	49.00%	0.34	0.33
Gradient Boosting	47.86%	0.31	0.28

6.5 Exploring the System on the Dementia Database

A study by Diniz et al. (2013) found that depression is highly correlated with an increased risk of all causes of Dementia. Dementia is a collection of symptoms related to deterioration in memory function, such as memory loss, thinking speed and the capability to handle problems. Several cognitive signs are associated with Dementia, such as depression. Depression can accrue in 40% of patients diagnosed with Dementia (Salary and Moghadam, 2013). A database concerned with Dementia people is used in this chapter for model training and system exploration. This database can be eligible for detecting depression and anxiety, especially since it provides depression and anxiety outcome measures aligned with the baseline system.

The audio recordings collected from a study by Elsey et al. (2015) has been used in this chapter for exploration purposes. In this chapter, these audio recordings are referred to as (Dementia Database). The study was mainly conduced due to the challenge of differentiating between Dementia and other memory disorders. The study aimed to distinguish between patients with a neurodegenerative disorder causing Dementia (ND) and patients with Functional Memory Disorder (FMD) through the analysis of clinical conversations. FMD can be defined as memory complaints relating to dysfunction that affect the patient's level of memory and functioning in their professional and/or social life (Schmidtke et al., 2008). ND and FMD share the same memory complaint symptoms, while ND and Depressive Pseudo-Dementia

(DPD) share the same depressive symptoms. Although Dementia assessment usually conflicts with DPD due to the matching symptoms, DPD is considered one of the depressive disorders. The study has been motivated by the high rates of ND and FMD diagnostic errors and the existing evidence that early signs of Dementia can be found in a patient's language. The patients attended the memory clinic at the Department of Neurology of the Royal Hallamshire Hospital, Sheffield, United Kingdom. The patients who participated in the experiment signed an ethical document for participation approval. They were encouraged to be accompanied by other people, such as a relative or friend. In the clinic, the patients were requested to fill out standard outcome measure questionnaires, such as PHQ-9 and GAD-7.

The Dementia database was used by Mirheidari et al. (2016) to investigate approaches to automatically detect Dementia based on conversation analysis. The study used ASR to extract features related to conversation analysis style. They mentioned that the database was not initially recorded for speech recognition purposes in that there was noise in the background and acoustic interference that made the speech challenging to process, especially the overlapping speech situations. As a result, the overlapping speech segments were discarded from the analysis. The nature of the conversations in the Dementia database is based on systemic questions introduced by the neurologists. These questions were designed to help reveal the standard signs of impairments of FMD and ND in conversation. The questions included closed questions to evoke long term memory, compound questions where patients diagnosed with Dementia would have difficulty answering those questions, questions that referred to the patient memory concerns and open questions related to the patient's life.

In this experiment of automatically detecting depression and anxiety, 39 recordings (around 32 hours total) were investigated. The average length of the recordings is 50 minutes and when extracting the patient-only turns, the duration of the available audio ranges from 2 to 30 minutes. In addition, the participants were requested to complete standard depression (PHQ-9) and anxiety (GAD-7) questionnaires. The structure of the PHQ-9 and GAD-7 questionnaires were described in Section 2.1. No definite diagnosis decision could be reached for two speakers in the Dementia database, but they were still included in the database because they had PHQ-9 and GAD-7 scores. Table 6.4 summarises the number of speakers and the duration of the speech samples (patient-only turns) per diagnostic category, including the two un-diagnosed speakers. Figure 6.2 shows the line graph distribution based on histogram counts of the PHQ-9 score contained in the database per diagnostic category. In the database, the highest PHQ-9 score is 22. Likewise, the highest GAD-7 score is 20 in the database. Figure 6.3 presents the line graph distribution of the GAD-7 scores histogram counts per diagnostic category.

Table 6.4 The Dementia database information

Diagnostic Category	Spk. Num.	Duration
DPD	15	2 h 44 min 7 sec
FMD	16	2 h 41 min 38 sec
ND	6	0 h 30 min 59 sec
Un-diagnosed	2	0 h 12 min 3 sec

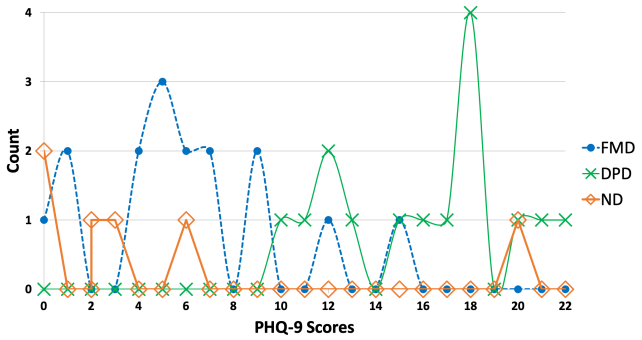


Fig. 6.2 Distribution of PHQ-9 score counts in the Dementia database based on the diagnostic classes

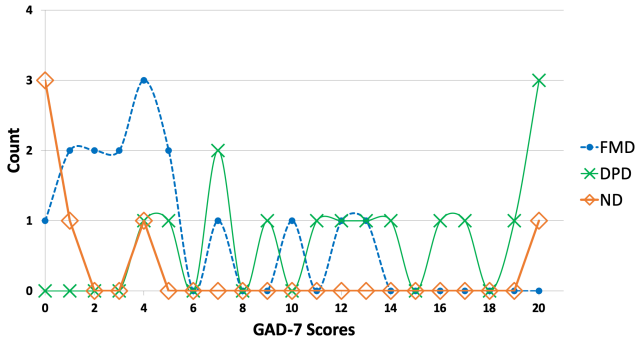


Fig. 6.3 Distribution of GAD-7 score counts in the Dementia database based on the diagnostic classes

It is clear that the FMD group have a minimal range of PHQ-9 and GAD-7 scores, while the people with DPD almost all have higher scores. In contrast, people with ND could have either have low or high PHQ-9 and GAD-7 scores. Furthermore, Figures 6.2 and 6.3 show that there is an increased number of scores that occur only a few times in the database, which is a likely challenge for prediction models.

6.5.1 System Implementation

The proposed system pipeline comprises pre-processing, feature extraction/selection and classification/regression. The classification or regression step has been implemented using

the cross-validation method due to the limited recordings in the database, which is a common practice when assessing the performance of ML models (Kerkeni et al., 2019). This experiment aims to explore the performance of the score prediction and score level classification concerning the diagnostic classes and the effect of using feature dimensionality reduction on the model performance using the Dementia database.

Pre-Processing

To predict the PHQ-9 and GAD-7 scores for the patient only, the parts of the neurologist-patient conversations that contained only the patient talking has been extracted. After that, the extracted segments have been concatenated to form a single audio file for each patient. The resulting grouped audio files ranged from 2 to 30 minutes.

Feature Extraction and Selection

The AVEC 2014 challenge baseline feature set mentioned previously has been used in this exploration (Valstar et al., 2014). A total of 2268 features have been extracted, consisting of 32 energy and spectral-related features and six voice-related features. The feature set has been extracted using the OpenSMILE toolkit (Eyben et al., 2010; Valstar et al., 2014).

Due to the high dimensionality of the feature set, the Recursive Feature Elimination (RFE) has been used as the feature selection method. It has been proven to be successful for similar sparse data domains (Mirheidari et al., 2019). The aim of RFE is to compute the coefficient of each feature to eliminate the features with the minimum coefficients in a recursive manner. The RFE strategy depends on first building an estimator that is trained on the training set features vector. Then, the individual feature's importance is obtained based on the retrieved coefficient of each feature. After that, the features that gained the smallest coefficients are eliminated from the current feature set. This procedure is repeated recursively until the desired number of features is reached (Guyon et al., 2002; Pedregosa et al., 2011).

One of the difficulties of implementing RFE with cross-validation is that on each fold the selected features would likely be different from some arrived at when running RFE on another fold. It is generally not possible to arrive at one common, overall reduced feature set. For this reason, the Recursive Feature Elimination Cross-validation (RFECV) has been implemented using the Scikit-learn library in python (Guyon et al., 2002; Pedregosa et al., 2011). RFECV specifies the number of feature sets by fitting over the training folds and selects the features that produce the least averaged error across all folds.

Classification/Regression

In the following section, the process of score prediction and score levels classification of the depression score (PHQ-9) and anxiety score (GAD-7) has been investigated. The predicted score is a continuous value ranging from either 0 to 27 or 0 to 21. For that reason, the process of prediction is considered a regression problem rather than a classification problem. However, to detect the severity of a specific banded score, a classification model has been applied to classify the score according to one of the nominated score bands.

Several common classifiers have been evaluated from the python scikit-learn toolbox to determine the most efficient model for classification, such as SVM, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier and Gradient Boosting Classifier. In addition, several common regressors have been compared from the python scikit-learn toolbox: SVR, Decision Tree Regressor, Gradient Boosting Regressor and AdaBoost Regressor. SVM and SVR gained the highest correlation coefficient results and the lowest error rates in predicting and classifying depression and anxiety scores and score levels, in agreement with the results gained from the challenge baseline in that SVR delivers the best results in predicting depression diagnostics (Cummins et al., 2013c; Grimm et al., 2007). Furthermore, SVM has been previously used to classify depression score levels, indicating its ability to classify mental state (Cummins et al., 2011, 2013a,b). Therefore, SVR and SVM models have been applied for regression/classification and RFE has been chosen as a feature reduction method. Due to the small size of the database, validating the SVR and SVM models has been implemented using parameter optimisation and cross-validation methods. Deep learning methods have not been investigated due to the sparse amount of data.

SVR is a methodology based on SVM that seeks to define the margin that the training set can align to (Drucker et al., 1997). The SVR model has been used with a linear kernel to predict depression and anxiety scores. Based on the parameter optimisation results, the parameters are set at a C of 1.00 and epsilon of 1.00. For the SVM, the parameters have been fine-tuned to a C of 1.00 and gamma of 1.00 and a linear kernel is used. One of the difficulties with determining the score distribution of the data, as presented in Figure 6.2 and Figure 6.3, is that the number of score occurrences in the database is limited and some score values may exist in a given test, but the model may not have seen sufficient examples of those values in the training set. For that reason, the stratified method of K-Folds cross-validation has been implemented. This cross-validation method can distribute the scores' occurrences in the training and test splits fairly, which improves the results in a way appropriate for such small databases that are often prevalent when working in the healthcare domain. The number of splits that helped in gaining better results is three folds.

To classify the severity levels based on the scores, the PHQ-9 scores have been rearranged according to the following bands (Kroenke et al., 2001): Minimal (0-4), Mild (5-9), Moderate (10-14), Moderately Severe (15-19) and Severe (20-27). The GAD-7 has been rearranged according to the following bands (Spitzer et al., 2006): Minimal (0-4), Mild (5-9), Moderate (10-14) and Severe (15-21). Each score range is called a 'band score' starting from band score 1 for the Minimal scores range, up to band score 5 for PHQ-9 and band score 4 for GAD-7, resulting in 5 and 4 bands for the PHQ-9 and GAD-7, respectively.

Regression Results

The use of stratified cross-validation method using three folds gained good results according to the Pearson Correlation Coefficient (R), MAE and RMSE. The results of predicting PHQ-9 and GAD-7 using RFECV and SVR are shown in Table 6.5. It is clear that depression score prediction requires more features for prediction. It gained a slightly lower R and higher error rates in comparison to the anxiety score prediction.

Table 6.5 Depression and anxiety score prediction results for the Dementia database (MAE and RMSE percentages of the total dataset)

Mood Outcome Measure	Num. Feat.	MAE	RMSE	R
GAD-7	31/2268	1.64 (3.22%)	1.94 (3.80%)	0.97
PHQ-9	148/2268	2.25 (4.41%)	2.78 (5.45%)	0.96

The results of the system after applying RFECV are shown in Figure 6.4 (depicting both actual and predicted scores) and likewise for GAD-7 in Figure 6.5. It seems that the model reasonably predicts the FMD,DPD, and ND people scores, while it finds it more difficult to predict the higher scores for DPD people. That might relate mainly to the minimal occurrences of these high scores in the database regardless of their diagnostic class.

Classification Results

Table 6.6 shows the depression and anxiety scores' bands classification results using SVM, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier and Gradient Boosting Classifier. It is clear that SVM achieved higher accuracy results in classifying the score bands of depression and anxiety.

To investigate the effect of depression and anxiety on the Dementia diagnostic category, in Figures 6.6 and 6.7, a plot has been shown of the confusion matrix of the depression and anxiety severity classification based on the mentioned banded scores for each of the available diagnoses in the database (FMD, ND and DPD). It is clear that the classifier could correctly

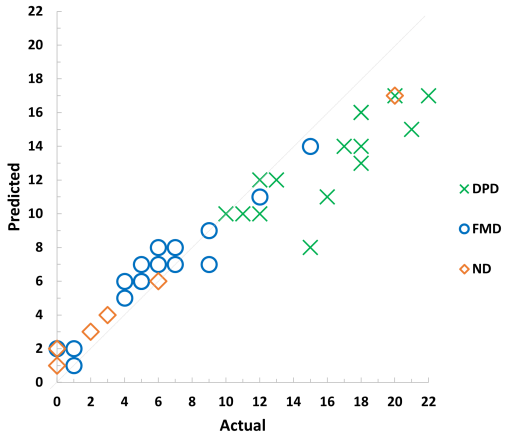


Fig. 6.4 The results of the actual versus predicted PHQ-9 scores for Dementia database

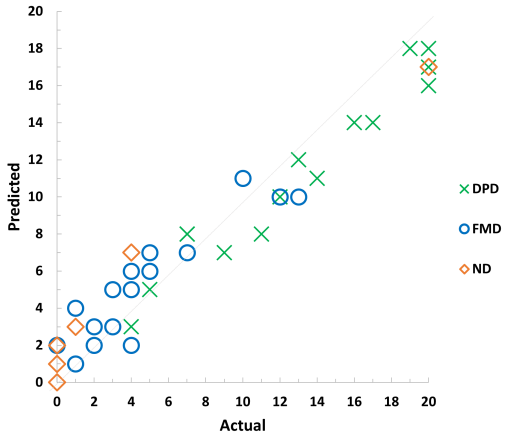


Fig. 6.5 The results of the actual versus predicted GAD-7 scores for Dementia database

Table 6.6 Depression and anxiety score bands classification results for the Dementia database using several classifier models

Mood Outcome Measure	Classifier	Accuracy	Precision	Recall
PHQ-9	SVM	80.73%	0.78	0.76
PHQ-9	Decision Tree	52.72%	0.47	0.47
PHQ-9	Random Forest	46.42%	0.39	0.41
PHQ-9	AdaBoost	44.94%	0.35	0.38
PHQ-9	Gradient Boosting	42.63%	0.31	0.29
GAD-7	SVM	85.00%	0.89	0.79
GAD-7	Decision Tree	56.94%	0.57	0.51
GAD-7	Random Forest	49.81%	0.42	0.42
GAD-7	AdaBoost	48.33%	0.39	0.40
GAD-7	Gradient Boosting	47.05%	0.35	0.31

classify the scores in bands 1 and 2 for people with FMD and ND, which are the bands for which the most PHQ-9 and GAD-7 scores occur for the FMD and ND groups. Likewise, the classifier struggled to classify the scores in the higher bands due to few occurrences of these scores in the range from 11 to higher scores. The higher PHQ-9 and GAD-7 scores mostly occur in the DPD group.

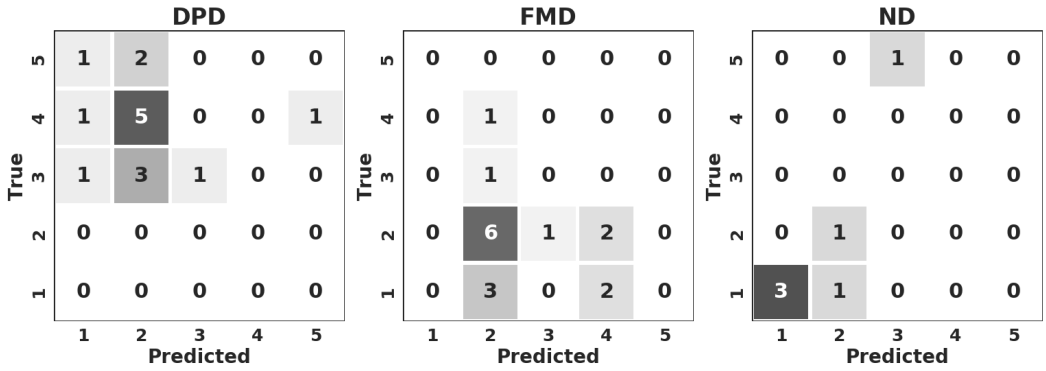


Fig. 6.6 The confusion matrix for classifying PHQ-9 band scores per each diagnosis (1 = Minimal, 2 = Mild, 3 = Moderate, 4 = Moderately Severe, 5 = Severe)

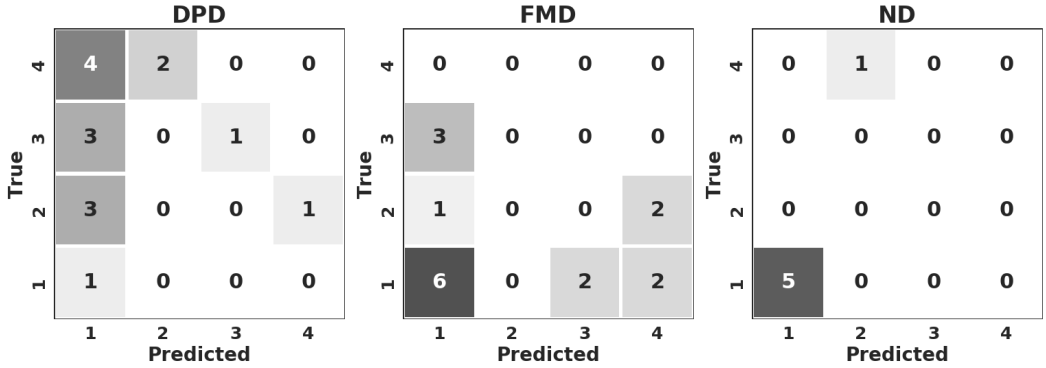


Fig. 6.7 The confusion matrix for classifying GAD-7 band scores per each diagnosis (1 = Minimal, 2 = Mild, 3 = Moderate, 4 = Severe)

6.6 Evaluating the System on the THEPS Dataset

The system pipeline implemented to evaluate the Dementia database has been investigated using the THEPS dataset. The subset of the THEPS dataset used in this experiment is shown in Figure 6.8.

The THEPS dataset has been labelled with PHQ-9 and GAD-7 as shown in Figures 6.9 and 6.10 and organised based on the levels of each score. It is clear from the figures that the mild and moderate score levels for PHQ-9 are the most prevalent levels in the dataset. For GAD-7, the mild and severe score levels are the most prevalent levels in the dataset.

The AVEC 2014 feature set has been used to extract the features from the patient's segments that are manually annotated for the ASR experiment described in Chapter 5. Each patient's segments have been merged to construct a single patient recording for each session. Then, the AVEC 2014 feature set was extracted from those recordings, resulting in 2268 features for each session. The regressors and classifiers (SVR and SVM) that showed the most promising results when evaluated using the Dementia dataset were applied

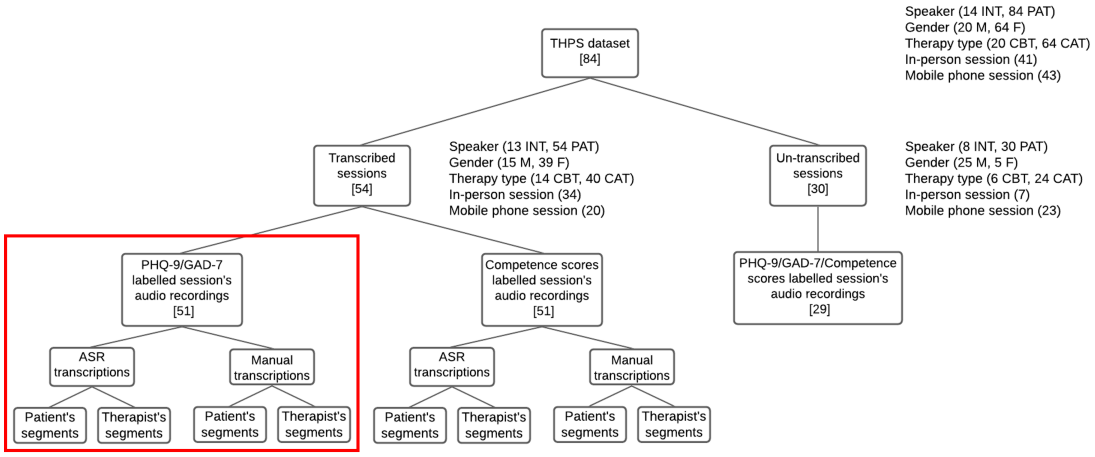


Fig. 6.8 THEPS dataset tree diagram highlighting the part used in this chapter

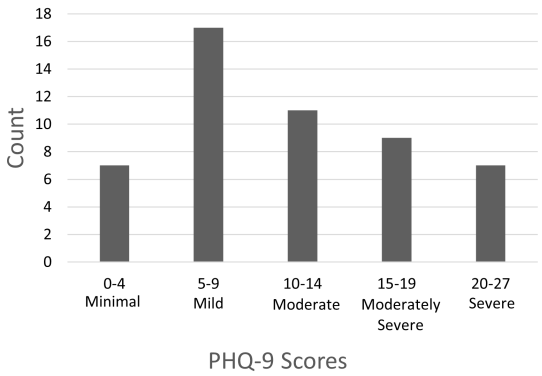


Fig. 6.9 Distribution of PHQ-9 score counts in the THEPS dataset based on score levels

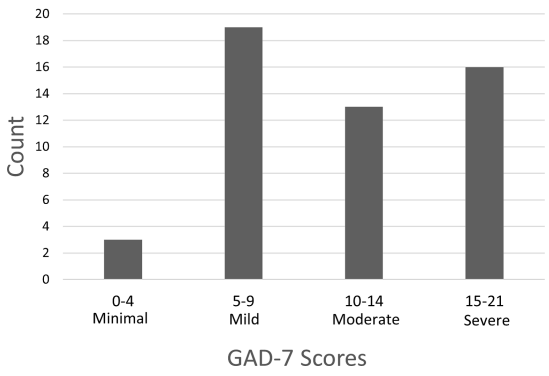


Fig. 6.10 Distribution of GAD-7 score counts in the THEPS dataset based on score levels

again for the THEPS-based system. Afterwards, parameter optimisation and RFECV have been implemented using SVR and SVM with three representative folds. The parameter optimisation resulted in a C of 1.00, epsilon of 1.00, and a linear kernel for SVR; and a C of 1.00, gamma of 1.00, and a linear kernel for SVM. The regression results for the depression and anxiety prediction based on the PHQ-9 and GAD-7 score for the THEPS dataset using SVR and RFECV are shown in Table 6.7. The results confirm the prediction results gained using the Dementia database (Table 6.5) in that depression requires more features for prediction and give results with lower correlation coefficients and higher error rates in comparison to the anxiety prediction results. Furthermore, as mentioned in Section 3.3, the THEPS dataset was initially created for a study focusing on delivering treatment sessions for anxiety disorders and that might contribute to the higher prediction rates for the anxiety scores compared to the depression scores (Kellett et al., 2020).

Table 6.7 Depression and anxiety scores prediction results for THEPS dataset (the MAE and RMSE percentages of the total dataset)

Mood Outcome Measure	Num. Feat.	MAE	RMSE	R
GAD-7	40/2268	1.39 (2.73%)	1.79 (3.51%)	0.96
PHQ-9	334/2268	2.70 (5.29%)	3.53 (6.92%)	0.88

The THEPS dataset is labelled with the therapist's competency measure as described in Section 3.3, but PHQ-9 and GAD-7 could not be clustered based on the level of the therapist's competence due to the lack of correlation between the mood outcome measures and the overall competency ratings as resulted from a study investigated by Power (2021) on the THEPS dataset. For this reason, the results of the prediction and classification have been categorised based on the depression and anxiety score levels. Figures 6.11 and 6.12 show the results of predicting depression and anxiety using the THEPS dataset based on the scores of PHQ-9 and GAD-7 scores, respectively. From Figure 6.11, it is clear that the mild and moderate levels of depression are almost accurately predicted scores for depression due to the high occurrences of those score levels in the dataset as shown in Figure 6.9. In Figure 6.12, the mild and severe score levels are mostly predicted accurately, thanks also to their high occurrences in the dataset as shown in Figure 6.10.

The classification results for the depression and anxiety score bands for the THEPS dataset are reported in Table 6.8. Also, the confusion matrix results based on the results gained from the SVM and RFECV techniques are presented in Figures 6.13 and 6.14. The score bands that occurred the most in the dataset affect the classification of the other score bands, especially for the mild score level in PHQ-9 and mild and severe score bands in GAD-7.

Table 6.8 Depression and anxiety score bands classification results for the THEPS database using several classifier models

Mood Outcome Measure	Classifier	Accuracy	Precision	Recall
PHQ-9	SVM	74.50%	0.80	0.77
PHQ-9	Decision Tree	53.92%	0.57	0.54
PHQ-9	Random Forest	46.40%	0.43	0.43
PHQ-9	AdaBoost	43.13%	0.37	0.38
PHQ-9	Gradient Boosting	42.33%	0.35	0.30
GAD-7	SVM	76.47%	0.67	0.69
GAD-7	Decision Tree	50.00%	0.41	0.43
GAD-7	Random Forest	45.09%	0.38	0.38
GAD-7	AdaBoost	43.52%	0.34	0.31
GAD-7	Gradient Boosting	41.08%	0.29	0.27

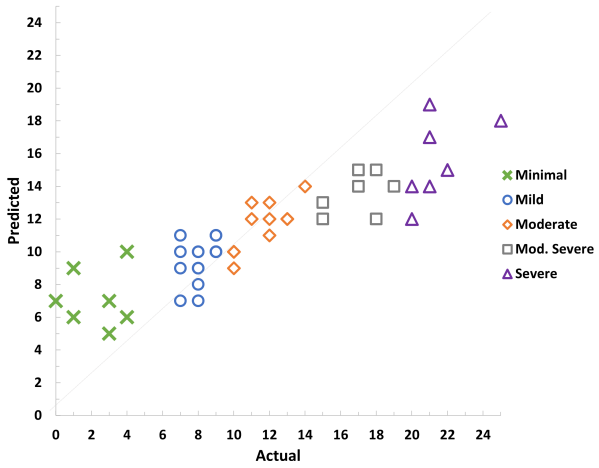


Fig. 6.11 Results of the actual versus predicted PHQ-9 scores for the THEPS dataset

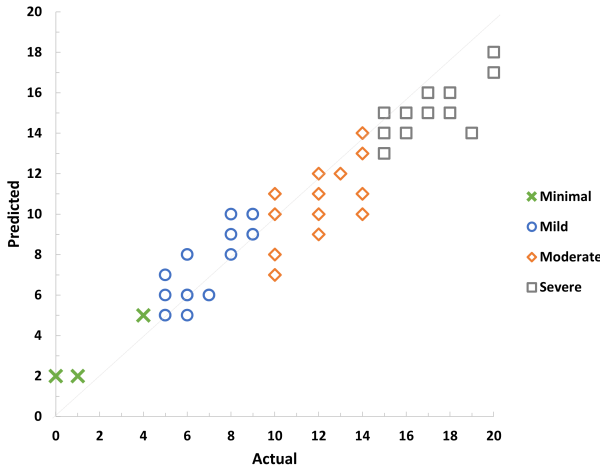


Fig. 6.12 Results of the actual versus predicted GAD-7 scores for the THEPS dataset

6.7 Summary

To conclude, this chapter described the implementation and evaluation of the automatic detection of depression and anxiety using mood outcome measures. The validation process of the baseline system has been successfully implemented for predicting depression outcome measures using the AVEC 2014 database and their provided baseline feature set. The prediction and classification of the depression and anxiety systems have been validated using mood outcome measures. The exploration and evaluation processes have been implemented using two datasets: the Dementia database and the THEPS dataset and it showed good results after using feature reduction and cross-validation. Furthermore, the results showed that classifying depression and anxiety using severity score levels can assist in classifying FMD and ND diagnoses, especially if the occurrence of these scores is fairly distributed. It is

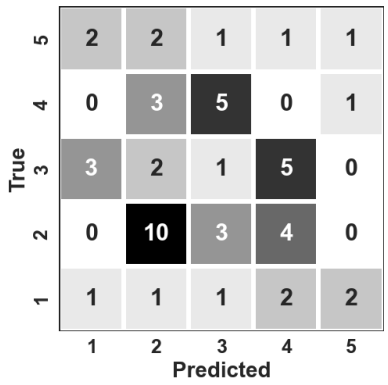


Fig. 6.13 The confusion matrix for classifying PHQ-9 band scores per each score level (1 = minimal, 2 = mild, 3 = moderate, 4 = moderately severe, 5 = severe)

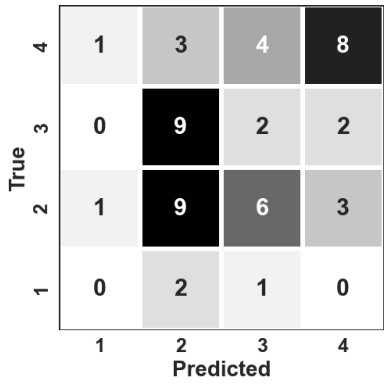


Fig. 6.14 The confusion matrix for classifying GAD-7 band scores per each score level (1 = minimal, 2 = mild, 3 = moderate, 4 = severe)

clear that detecting depression and anxiety using outcome scores as a regression problem achieved higher performance results in comparison to the classification problem, which is more related to the ability of the ML models to predict scores rather than classifying more than one score under a single class. Another shown result is that depression requires more features for prediction and is more challenging to predict than anxiety. The next chapter will investigate another module in the proposed system that is the automatic time-continuous recognition of emotional dimensions.

Chapter 7

Automatic Time-Continuous Speech-Based Recognition of Emotional Dimensions

Chapters 6 explored the feasibility to detect the mood outcome measures automatically. The questionnaires related to those measures address several patients' behaviour patterns that could correlate with the their behaviours expressed in the therapy sessions. The patient's and therapist's emotions could be one of those powerful behaviours that could influence the treatment outcomes. This chapter explores the efficacy of detecting the patient's and therapist's emotions and their relationship to the mood outcome measures and the therapist's competency measures using acoustic features. This chapter will concentrate on the acoustic features and the automatic transcription outputs from Chapter 5 will not be used in this chapter.

7.1 Introduction

As have been discussed in Section 2.3, the patient's emotions are considered a dynamic factor that can help establish a positive therapeutic alliance in psychotherapy sessions. The therapist should acknowledge and motivate the patient's positive emotions through talk therapy. The small variations in emotions could help the therapist determine the appropriate treatment plan and guide the patient toward a better therapeutic experience. Eventually, the therapist should obtain a clear view of the patient's distressful situations and help them to distinguish their thoughts from their emotions. Furthermore, the therapist should be empathetic to the patient's emotions and evaluate any flawed thinking that has impacted the patient's mood

(Beck, 2011). The therapist could accomplish that by fostering interpersonal synchrony with the patient which could increase the patient's willingness to collaborate in therapy (Koole and Tschacher, 2016). The therapist's ability to assess the patient efficiently would indicate a highly competent therapist. Assessing the therapist's competence includes evaluating the therapist's use, knowledge and implementation of the treatment (Barber et al., 2007; Fairburn and Cooper, 2011). Predicting the patient's and therapist's emotions using a dimensional approach could help capture the small variations in the emotions expressed in the session.

One part of the proposed system mentioned in Section 3.2 is to detect patient's emotions. Therefore, due to the importance of detecting emotions in therapy, this chapter investigates automatic methods for continuous emotion recognition in therapy sessions using the emotional dimensions: *arousal* and *valence* as discussed in Section 4.3.3. Moreover, a qualitative study is presented to understand how the power of the therapist's empathy and the interpersonal synchrony relates to the therapist's competency measure. It is important to highlight that, based on the literature, the systems aimed to extract the detailed representations of emotion in a continuous domain along with the temporal dynamics of emotion, usually referred to as emotion tracking or continuous-time representation of the affective content (arousal and valence) (Malandrakis et al., 2011). The time-continuous concept here refers to continuous predictions in time rather than a single prediction for the whole session. Each prediction does not depend on the previous prediction in time or value.

The rest of this chapter is organised as follows: Section 7.2 describes related work concerning this chapter's aims, Section 7.3 presents the implemented automatic dimensional emotion recognition system, Section 7.4 explains the experiment of validating the baseline system on the REmote COLlaborative and Affective interaction (RECOLA) database, Section 7.5 presents the results of evaluating the baseline system on the THEPS data as well as both qualitative and quantitative analyses. Finally, Section 7.6 summarises the chapter's findings.

7.2 Related Work

Speech Emotion Recognition (SER) is the process of predicting human emotions from speech signals. SER is a significant area of research in several fields of human-machine interactions such as healthcare, call centres, automotive environments, robots, mobile services and psychology (France et al., 2000; Grimm et al., 2008; Schuller et al., 2004; Huahu et al., 2010; Yoon et al., 2007; Boateng et al., 2020). Even though SER is applied in several fields, it is considered a challenging task due to the subjectivity in accurately measuring human emotions. Perceived emotions are subjective to human understanding of emotions and it would be challenging to apply a suitable measure for rating emotions. Furthermore, there is

no standard agreement on how to label emotions. They are usually labelled using the human perception of emotion continuously over time, even though the labelled emotions can vary between raters (Tversky, 1969). Furthermore, recordings with naturally expressed emotions are challenging to obtain due to ethical and legal concerns. It has been commonly agreed, through the literature, that the valence emotional dimension is more challenging to predict using acoustic features alone than the arousal dimension (Wöllmer et al., 2008).

The databases for SER can be explored according to the method used to express emotions, such as *acted* speech emotion databases, *elicited* speech emotion databases and *natural* speech emotion databases. Acted speech databases are recorded by actors in a soundproof environment. These are considered easier to record, but acted speech cannot convey real-life emotions and might be exaggerated. As a result, the acted emotions might not accurately match the speaker's true feelings, which could minimise natural emotions' recognition rates. Elicited speech databases are established by putting the speakers in a simulated emotional situation that can mimic several emotions. This type of database could lead the speaker to use emotions that are not fully elicited but closer to real ones. The natural speech emotion databases are collected from real-world situations such as talk shows, radio talks, call centre recordings and similar sources. These databases could also contain spontaneous speech. It can be challenging to obtain those types of data due to the ethical and legal concerns relating to the data processing and distribution (Akçay and Oğuz, 2020).

It is important to focus on the natural speech emotion databases for their eligibility to match the emotions expressed in the THEPS dataset. A list of the common natural speech emotion databases is outlined in Table 7.1. The databases aforementioned in the table are labelled with discrete and/or dimensional emotions, which some are described in Section 4.5. The most relevant databases are the RECOLA, SEMAINE, Sensitive Artificial Listener (SAL) and Vera Am Mittag (VAM) databases because they are labelled with dimensional emotions relevant to this chapter's aims. The SEMAINE is an audio-visual database created to build automatic agents to induce the speakers into an emotional coloured conversation. It was built based on the interactions between users and simulated agents for 150 speakers with 959 conversations (McKeown et al., 2011). Since the SEMAINE database does not include a human to human interaction, which is the nature of the THEPS dataset rather a human to a machine, it would be a less appropriate choice of database which to establish the thesis SER system presented in this chapter. The VAM database consists of 12 hours of audio-visual talk show recordings capturing spontaneous emotional speech. The database recordings were authentic discussions between the talk show guests, who were 47 speakers in 959 conversations (Grimm et al., 2008). The RECOLA database is a multimodal database devoted for spontaneous collaborative and affective interactions. It consists of 46 speakers

recorded in dyads during a video conference while completing a required task. The speakers were required to complete self-reports related to their emotional and social behaviours. The organisers of the data collection experiments decided which speaker could receive a mood induction that simulates either positive or negative emotions based on the completed self-reports (Ringeval et al., 2013). The self-reports approach is the same as the approach used in the mood outcome measures in the THEPS dataset. The mood induction strategy is similar to the therapist's behaviour, trying to induce some of the patient's emotions in some moments in the therapy. The SAL database consists of audio-visual recordings of human-computer conversations elicited through an interface designed for the users to work through a range of emotional states. The data were collected for four users with 20 minutes each (Douglas-Cowie et al., 2007). The Sentiment Analysis in the Wild (SEWA) database includes audio-visual recordings of human interactions by German participants' revealing spontaneous and natural behaviours. The participants were required to discuss a commercial they had viewed in pairs (Kossaifi et al., 2019). Several studies that employed the RECOLA, SEMAINE, SAL, SEWA and VAM databases in their systems are mentioned later in this section. In order to fully evaluate a SER system on the THEPS data, it would be necessary to have access to the continuous emotional labels. THEPS dataset is not labelled with emotional labels. Therefore, a system is trained on a benchmark database (RECOLA) and then this model is used to predict emotions in the THEPS dataset.

Table 7.1 Natural speech emotion databases

Database	Size	Emotion labels
Chinese Annotated Spontaneous Speech corpus (CASS) (Li et al., 2000)	7 speakers (2 male, 5 female),6 h of speech	Anger, fear, happiness, sadness, surprise, neutral
RECOLA Speech Database (Ringeval et al., 2013)	46 speakers (19 males, 27 females), 7 h of speech	Five social behaviors (agreement, Natural, dominance, engagement, performance, rapport); arousal and valence
SEMAINE Database (McKeown et al., 2011)	150 speakers, 959 conversation	Valence, activation, power, expectation, overall emotional intensity
Vera Am Mittag Database (VAM) (Grimm et al., 2008)	47 speakers from talkshow, 947 utterances	Valence, activation,and dominance

FAU Aibo Emotion Corpus (Batliner et al., 2008)	51 children talking to robot dog Aibo, 9 h of speech	Anger, bored, emphatic, helpless, joyful, motherese,neutral, reprimanding, rest, surprised, touchy
TUM AVIC Database (Schuller et al., 2009)	21 speakers (11 male, 10 female), 3901 utterances	Five level of interest; 5 non-linguistic vocalizations (breathing, consent, garbage, hesitation, laughter)
AFEW Database (Kossaifi et al., 2017)	330 speakers, 1426 utterances from movies, TV-shows	Anger, disgust, surprise, fear, happiness, neutral, sadness
Sensitive Artificial Listener (SAL) Database citepdouglas2007humaine	4 speakers, 20 minutes each from human-computer conversations	Activation and valence
Sentiment Analysis in the Wild (SEWA) database (Kossaifi et al., 2019)	66 speakers, 44 hours in total human-computer conversations	Valence, arousal, liking and social signals

The literature on SER has highlighted several approaches for the Machine Learning (ML) process. The most general framework for the ML process can be summarised in the following steps:

- Feature extraction: The acoustic features are extracted from speech segments to create a master feature vector for model training. The acoustic features that are typically adopted in the SER system have been described in Section 4.3.2.
- Classifiers: A ML classifier/regressor is trained on the feature vectors to implement emotion classification/regression. It can adopt several ML models such as Hidden Markov Model (HMM) (Nwe et al., 2003), Support Vector Regression (SVR) (Ringeval et al., 2015b) and Deep Neural Network (DNN) (Partila et al., 2015).

Researchers have focused on evaluating the effect of various features and classifiers in SER systems. Wu et al. (2010) focused on estimating the emotions in three dimensions space (valence/activation/dominance) using the VAM database. Several acoustic features were extracted from the database, showing promising results in earlier studies. The extracted features were Fundamental frequency (F0), duration, energy and Mel Frequency Spectrum

Coefficients (MFCCs) features. They studied several approaches in estimating emotions using Robust Regression (RR), SVR and Locally Linear Reconstruction (LLR). They found that the used estimators were suitable for dimensional emotion estimations. Furthermore, the estimations of the valence dimension showed lower correlation coefficients than the activation and dominance dimensions. Han et al. (2017) focused on the realistic continuous emotion recognition from audiovisual signals in the valence and arousal dimensions. Two databases were investigated to study the effectiveness of the system: RECOLA and SEMAINE. Based on earlier studies in the literature, MFCCs features were chosen due to their effectiveness in relation to RECOLA and SEMAINE. SVR and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) were concatenated in a hierarchical framework as the ML models. The experimental results showed that the suggested model could deliver a significant performance improvement for audio and audiovisual systems. Wöllmer et al. (2008) implemented an LSTM-RNNs architecture for continuous emotion recognition in a 3D space using SAL database. The three dimensions included in the study were activation, valence and time. The selected features were energy, pitch, voice quality and spectral features. The results showed that activation gained the best results with a Mean Squared Error (MSE) of 0.08 using LSTM-RNNs. The valence dimension achieved an MSE of 0.18 using LSTM-RNNs and SVR equally. They noticed that the classification performance of the valence was quite low as the detection of valence has been known to be challenging. The systems, as mentioned earlier, provided a successful architecture for the SER systems, including the use of features, databases and classifiers.

The Audiovisual Emotion recognition Challenge (AVEC) is well-known in the field of continuous emotion recognition. The main goal of the challenge is to join the researchers from the visual and audio analysis communities around mainly the topic of emotion recognition. It started in 2011 by defining the goal of the challenge to recognise four emotional dimensions (arousal, expectation, power and valence) continuously at the word level (Schuller et al., 2011b).

The AVEC 2012 changed the challenge goal to recognise the same four emotional dimensions continuously at every moment in the recordings and per word uttered by the speaker (Schuller et al., 2012). The challenge participants were provided with the feature sets required for each modality. Furthermore, the challenge baseline used the SEMAINE database for recognising the labels. The proposed audio feature set consists of energy, spectral and voicing related features defined as Low-Level Descriptors (LLD) features. Additionally, the feature set includes the functionals required to be extracted from the LLDs, such as root quadratic mean and standard deviation. The AVEC 2012 audio baseline feature set contains

1841 features. The feature sets are provided using the openSMILE toolkit (Eyben et al., 2010). The challenge baseline used SVR as a classifier for recognising emotions.

The AVEC 2013 changed the challenge goal to predict the continuous labels of emotions for arousal and valence each moment in time. The used feature set was upgraded to include the spectral flatness and the MFCCs 11-16, leading to a total of 2268 features (Valstar et al., 2013). In 2014, the challenge baseline further expanded the dimensions that needed to be recognised to include the dominance dimension. The AVEC 2014 specified the same audio feature set as AVEC 2013 (Valstar et al., 2014). The 2015 version of the challenge baseline introduced the RECOLA as the benchmark database for the participants to use. The database included the physiological modality with the audio and the video modalities to extend the challenge for combining audio, visual and physiological emotion recognition (Ringeval et al., 2015a).

The AVEC 2016 challenge baseline used the same database as AVEC 2015 despite a change in the audio feature set to be the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2015). The eGeMAPS feature set was selected to be used in the challenge baseline due to its compact use of features based on expert knowledge. Furthermore, the features showed their high robustness in modelling emotional speech. The eGeMAPS acoustic LLDs includes spectral, cepstral, prosodic and voice quality information as a total of 88 features and they are extracted using the openSMILE toolkit (Valstar et al., 2016).

In 2017, the challenge baseline introduced another database for the continuous emotion recognition sub-challenge, namely a subset of the SEWA database. Furthermore, they expanded the dimensions required to be recognised in the challenge baseline to be: arousal, valence and liking (Ringeval et al., 2017). The 2018 version of the challenge baseline introduced two sub-challenges related to emotion recognition (Ringeval et al., 2018). The first goal is the Cross-cultural Emotion Sub-challenge (CES) to predict the level of the three emotional dimensions (arousal, valence, liking) continuously under several cultures environment. They introduced an extended version of the SEWA database recorded in the same conditions with Hungarian participants as a blind test set. The second goal was the Gold-standard Emotion Sub-challenge (GES) that focused on generating dimensional emotion labels by fusing continuous annotations of dimensional emotions rated by several annotators. Then, the fused annotation is used to train and evaluate a baseline emotion recognition system using the RECOLA database.

The need for the gold-standard emotion labels relates to the differences found in the annotations that could highly affect the system's ability to learn the proper mapping between the input data. As mentioned before, the emotion labels are subjective and depend on human judgments of human behaviours. Those judgments are by nature highly variable

and subjective (Tversky, 1969). For that reason, the GES challenge baseline followed the dominant approach in the literature referred to as *gold standard* that is to combine the annotations for each recording across time. Nonetheless, several challenges arose in the fusion process, such as the inconsistencies that appeared between the annotators' values and a delay presented between the emotional events and the corresponding annotation value that relate partially to the reaction time (Tversky, 1969). The challenge baseline used an Evaluator Weighted Estimator (EWE) based approach to combine the annotations into a single gold standard per recording per dimension (Pandit et al., 2018; Grimm and Kroschel, 2005). The winners of the GES sub-challenge evaluated the baseline on the RECOLA database against a proposed scheme for annotation fusion. The results showed that their method achieved similar or better correlations for the valence dimension (Booth et al., 2018).

Through the latter large body of literature related to the history of the AVEC challenge, it has been clear that several baselines were introduced, including various label dimensions, databases with different recording environments and several acoustic features. Furthermore, several studies in the medical field incorporated the challenge baseline in their system architectures, especially the use of the eGeMAPS feature set (Ringeval et al., 2016; Mencattini et al., 2018; Gideon et al., 2019; Zhang et al., 2020). This represents a highly efficient and thoroughly investigated baseline which has been adopted in similar fields of study. It can therefore be considered as a reliable and successful choice of pipeline on which to base this chapter's experiments.

7.3 Automatic Dimensional Emotion Recognition System

The AVEC 2018 challenge baseline system has been used for the dimensional emotion recognition system presented in this chapter. In particular, the GES has been selected due to its compatibility with the experiment requirements. Having the RECOLA database labelled with the gold-standard labels for the emotions could ensure the efficiency of the emotional labels for training the proposed SER system. Furthermore, the database used for the challenge baseline (RECOLA) is a natural emotions database that resembles the THEPS dataset used in this chapter. Initially, the AVEC 2018 baseline system has been re-implemented using the same baseline training and development sets to set up the proposed SER system for further experiments. Then, the proposed SER system has been analysed on the THEPS dataset using the RECOLA training and development sets as the training set.

7.4 Validating the System on the RECOLA Database

7.4.1 AVEC 2018 Challenge Baseline

This section will describe the pipeline used in the AVEC 2018 challenge baseline including the use of features and ML models. The RECOLA database is described in Section 4.5.2. The challenge baseline used three methods that involve different levels of supervision in the feature extraction phase, described bellow:

Supervised: Expert knowledge

At the supervised level, features depend directly on the emotion rating expert's knowledge-based representations. The eGeMAPS was used in this challenge as the supervised feature set. It comprises 88 features covering the formerly mentioned acoustic features: spectral, cepstral, prosodic and voice quality information. Table 7.2 presents the 42 LLDs computed as part of the eGeMAPS feature set (Valstar et al., 2016). The functionals of all the LLDs computed consists of several mathematical representations such as the arithmetic mean and the coefficient of variation. The details of those functionals and the feature set can be found in (Eyben et al., 2015). The features were extracted using the openSMILE toolkit (Eyben et al., 2010; Valstar et al., 2016).

Table 7.2 eGeMAPS LLDs (Valstar et al., 2016)

1 energy related LLD	**Group**
Sum of auditory spectrum (loudness)	Prosodic
25 spectral LLD	**Group**
α ratio (50-1000 Hz / 1-5 kHz)	Spectral
Energy slope (0–500 Hz, 0.5–1.5 kHz)	Spectral
Hammarberg index	Spectral
MFCC 1–4	Cepstral
Spectral flux	Spectral
16 voicing related LLD	**Group**
F0 (linear & semi-tone)	Prosodic
Formants 1, 2, 3 (frequency, bandwidth, amplitude)	Voice quality
Harmonic difference H1–H2, H1–A3	Voice quality
Log. HNR, jitter (local), shimmer (local)	Voice quality

Semi-supervised: Bags of Audio Words

Another type of feature that the AVEC 2018 challenge baseline implement is the Bag of Audio Words (BoAWs). The BoAWs method involves generating words with a clustering algorithm and quantising the original features to generate the bag-of-words in the form of a histogram. The MFCCs are used in BoAWs as a front end to compute the acoustic features. The process starts by quantisation of the LLD vectors from single frames according to a codebook after the process of extracting the MFCC LLDs from the audio signal. This codebook is a result of a random sampling of the LLDs for all the training partition. Finally, a histogram is generated based on the distribution of the codebook vectors over the whole audio segment for each recording in the test set as a bag-of-words (Schmitt and Schuller, 2017). The technique of BoAWs was described by the authors of the AVEC 2018 challenge baseline as semi-supervised representation learning because it represents the distribution of LLDs according to a dictionary or codebook learned from them (Schmitt et al., 2016; Cummins et al., 2018). The BoAWs were extracted using the open-source toolkit openXBOW (Schmitt and Schuller, 2017). Further explanations on BoAWs can be found in (Schmitt et al., 2016; Schmitt and Schuller, 2017).

Unsupervised: Deep Spectrum

The unsupervised features are directly learned from the raw signal (Trigeorgis et al., 2016) or produced with the use of out-of-domain data (Cummins et al., 2017). The challenge used the Deep Spectrum features as the unsupervised features that were introduced earlier for snore classification (Amiriparian et al., 2017). They have been extracted using deep learning representations inspired by image processing. Several studies found that Deep Spectrum features have been effective in multimodal emotion recognition systems (Cummins et al., 2017).

AVEC 2018 Challenge Baseline used several dimensional regressors as the ML models such as SVR from the liblinear toolkit and Generalised Linear Models (GLMs) such as Ridge regression, Elastic Net and Lasso from the scikit-learn toolbox (Fan et al., 2008; Pedregosa et al., 2011). Furthermore, multi-task formulation of the Elastic Net and Lasso algorithms have been implemented to make use of the correlations between the dimensions. The AVEC 2018 challenge baseline reported in the challenge research paper that the best results was gained using the BoAWs in recognising the dimensional emotions under the audio modality using the RECOLA database. The valence dimension results are quite challenging, comparable with the arousal, which has been commonly agreed in the literature (Ringeval et al., 2018).

7.4.2 Challenge Baseline Results

The AVEC 2018 challenge baseline has been re-implemented on the RECOLA dataset using the same feature methods described earlier. Table 7.3 presents the gained replicated results and baseline results reported in the paper for the development partition as the test partition labels are hidden (Ringeval et al., 2018). The table shows the results based on the Concordance Correlation Coefficient (CCC) and the best regression model for the arousal and valence dimensions. The difference between the replicated results and the reported one is considered minimal and that difference might be caused by configurations of the implemented systems. The best results were gained using the BoAW features with the SVM for the arousal dimension and Lasso for the valence dimension. It is important to highlight that the AVEC 2018 challenge baseline authors referred to the dimensional regressors as exploring various linear models using "SVMs" rather than SVRs, which is used in this thesis.

Table 7.3 Comparison between the challenge reported results (Ringeval et al., 2018) and the replicated results, best regression model (S:SVMs, L:Lasso, E:ElasticNet) is given in superscript.

Data Partition	eGeMAPS	BoAWs	Deep Spectrum
Arousal			
Development (AVEC Results)	0.749^{S}	0.760^{S}	0.621^{L}
Development (Replicated results)	0.749^{S}	0.751^{S}	0.621^{L}
Valence			
Development (AVEC Results)	0.319^{S}	0.364^{L}	0.220^{E}
Development (Replicated results)	0.325^{S}	0.353^{L}	0.220^{E}

7.5 Analysing the System on THEPS Dataset

The validated system in the previous section has been re-trained using the full obtained RECOLA database as a training set (train and development sets) to analyse the THEPS dataset (test set). The THEPS dataset described in Section 3.3 consists of 54 transcribed sessions. The transcribed sessions have been used due to the availability of the time alignments for each speaker separately to detect each speaker's emotions individually. Considering the unavailability of the emotion's labels in the THEPS dataset, the Patient Health Questionnaire (PHQ-9), Generalised Anxiety Disorder(GAD-7) and the competency measures were treated as indicative, correlated measures when analysing the system results. Figure 7.1 presents the subsets from the THEPS dataset that have been used in this analysis. The manual

transcriptions have been used for the speaker's utterance segmentation to obtain the time alignment information.

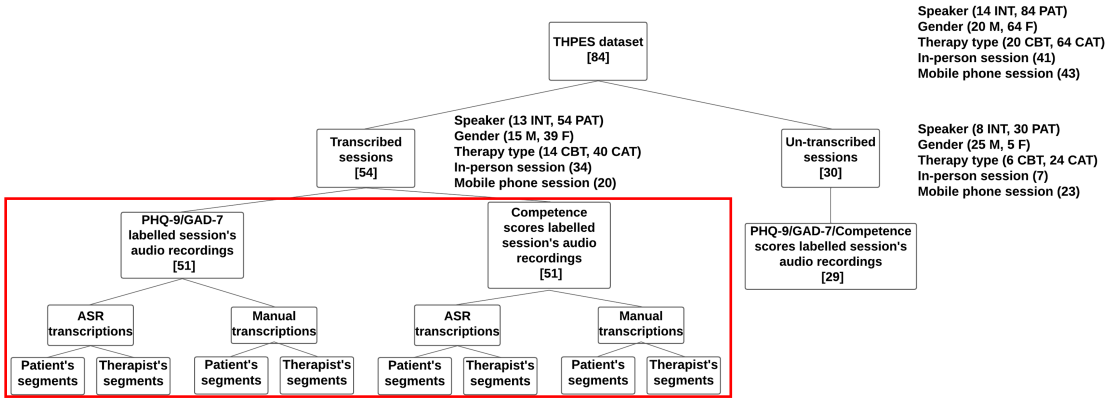


Fig. 7.1 THEPS dataset tree diagram highlighting the part used in this Chapter

The mood outcome measures related to the patient's mental disorder are PHQ-9 for depression disorder and GAD-7 for anxiety disorder. Using the extracted patient's segments has been vital to investigate the relationship between the mood outcome measures and the recognised dimensional emotions. The patient's segments ranged from 0.14 seconds to 1069.85 seconds in each session. The optimal length of the included segments in the experiment in order for the patient's emotions are clearly defined by human perception of emotion has been investigated. After investigation, segments smaller than 0.5 seconds were excluded due to their unsuitability with the chapter aims.

As described in Section 2.3, the competency measure is a rating scale used by the psychotherapy supervisors to measure the therapist's performance in treatment sessions. As mentioned in the referred section, the scale depends on the COM-B model (Michie et al., 2014) that is a model of behaviour and behaviour change to conceptualise the patient's problem behaviour as resulting from the interaction of three factors. After reviewing those factors, it is clear that the competency measure is a rating that relies on both the patient and the therapist to be assessed in the session. For that reason, the full session recordings were used to investigate the relationship between the recognised dimensional emotions and the competency measure.

The features selected for analysing the THEPS dataset are eGeMAPS and BoAWs because BoAWs features gained the best correlation coefficients in the AVEC 2018 baseline for the audio modality (Ringeval et al., 2018). The eGeMAPS has been used at the beginning to build and train the SER system using the RECOLA dataset. The eGeMAPS features have been extracted using the openSMILE toolkit, while the BoAWs have been extracted using the

openXBOW toolkit. The AVEC 2018 challenge baseline reported in their configuration file that the eGeMAPS and BoAWs were computed with a sliding window from 3 to 9 seconds with a step size of 100 ms. An optimisation process was then performed on the development set by implementing a grid search over the window size. Due to the unavailability of the emotional labels for the THEPS dataset, the performance evaluation results were missing, leading to difficulty conducting the grid search as implemented in the baseline system. For this reason, a full investigation of the recognition results for each window size has been reviewed separately, leading to selecting the window size to be 3 seconds with a step size of 100 ms because those configurations achieved minimal empty results.

As mentioned earlier, the BoAW feature set starts by extracting 12 MFCCs, delta and delta delta for training and test datasets. The BoAWs are audio representations formed by bagging acoustic LLDs such that each frame-level LLD vector has been allocated to an audio word from a codebook retained from the training data. A fixed-length histogram representation of an audio recording has been generated by counting the number of assignments for each audio word. As implemented in the baseline, the codebook size consists of 100 words (Cummins et al., 2017). The extracted features from the training set are then concatenated and normalised for the classification phase. The regressors used in the recognition system are the SVR and the Generalised Linear Models (ridge regression, lasso, multi-task lasso, elastic net and multi-task elastic net). The results reported in the following section are based on the models that achieved the best recognition results in the baseline system.

7.5.1 Analysis and Results

The section explores two main objectives. The first objective is to investigate the effect of different types of feature sets in the SER system on the prediction of the arousal and valence dimensions from a quantitative perspective, including changes in the patient's segments or the whole sessions (therapist's and patient's segments). Furthermore, the effect size (Cohen's d) has been calculated between several trends in the study, including the selected features, the predicted dimensional emotions and the mood outcome measures as suggested in (Baird et al., 2020). The other objective is to conduct a qualitative study to understand the subtle changes in the predicted patient's emotions and emotions within the interaction between the patient and the therapist during several periods in the therapy treatment. Due to the unavailability of any labels related to the therapist's empathy or the interpersonal synchrony in the THEPS dataset, several sessions were explored qualitatively to study the relationship between them and the competency measures. Furthermore, due to the unavailability of the emotion labels in the THEPS dataset, it is impossible to produce any correlation coefficient results.

Quantitative Analysis

To explore the differences between the results gained using the eGeMAPS and BoAWs features, the averaged predictions of the dimensions arousal and valence were plotted for the patient's segments as shown in Figure 7.2 and 7.3. Figure 7.4 and 7.5 presents the same measures for the full sessions. The figures show that most of the patient's emotions are presented by a negative valence and low arousal. That is, a more negative depressed emotion, as shown in Chapter 4 (Figure 4.1). For that reason, the predicted patient's emotions is align with the most common patient's mental disorder represented in the dataset as discussed in Section 3.3.

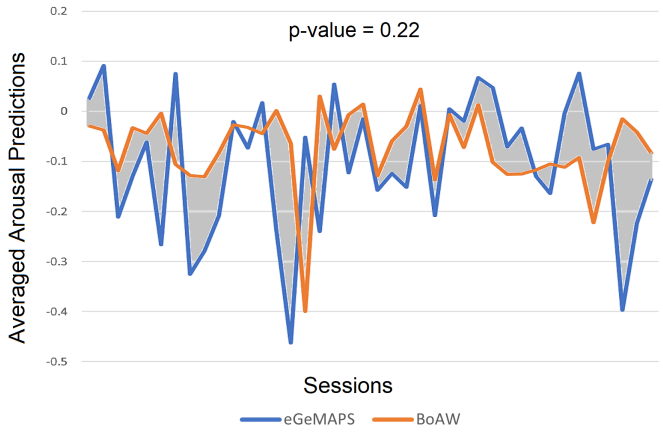


Fig. 7.2 Arousal averaged predictions for patient's segments using eGeMAPS and BoAWs features

The figures show that the predictions of both features in the arousal dimension are close to each other, indicated by the p-value, which resulted in values higher than 0.05, which means the two features' results are not significantly different. While the BoAWs gained more stable results than the eGeMAPS in the valence dimension, especially in Figure 7.3. The results gained confirms the baseline results in that the valence is somewhat challenging to predict, in comparison to with the arousal. Furthermore, the BoAWs enhance the valence predictions, while both features' predictions are very similar in the arousal dimension. Some prediction results were missing after the model was tested which might relate to the inability of the model to predict some of the dimensional emotions.

The effect size has been explored between the predicted dimensional emotion results from eGeMAPS and BoAWs based on each type of recording (patient only or full session) as presented in Figure 7.6. It is clear that the valance indicated a higher effect size on both types of recordings that could relate to the efficiency of the BoAWs in predicting the valance dimension. Also, the effect size has been calculated between the patient's predicted

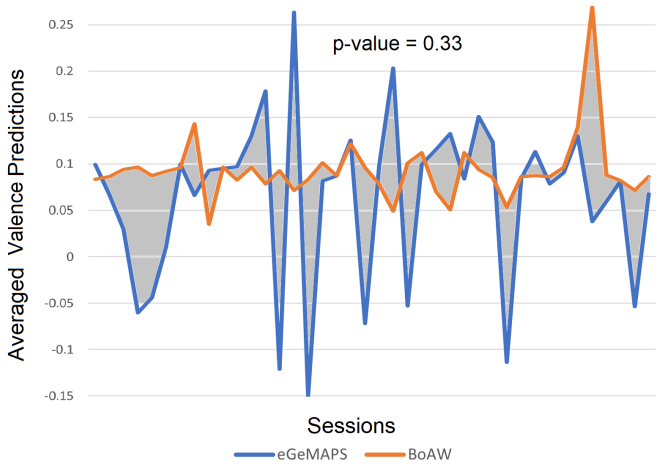


Fig. 7.3 Valence averaged predictions for patient's segments using eGeMAPS and BoAWs features

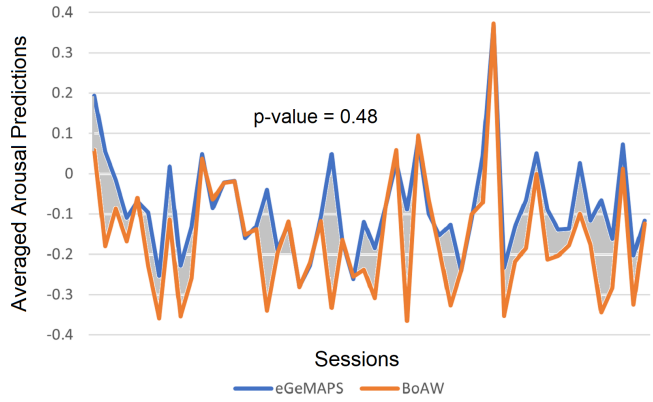


Fig. 7.4 Arousal averaged predictions for full sessions using eGeMAPS and BoAWs features

dimensional emotion and the mood outcome measure levels as presented in Figure 7.7 and 7.8. To achieve that, it is important to select a cut-off score for each mood outcome measure to define the low and high end of the scales. For GAD-7, the selected cut-off score is ten as described in (Spitzer et al., 2006). For PHQ-9, a cut-off score of 11 has been selected as described in (Manea et al., 2012). The figures illustrate that the BoAWs appear to have almost a moderate effect size, more prominent than eGeMAPS, particularly for the valence dimension. This result is matching with the baseline results in that the BoAWs improved the prediction results comparable with the eGeMAPS features in the valence dimension.

The relationship between the full session (patient and therapist) and the competency measure has been investigated using the effect size as shown in Figure 7.9. Due to the

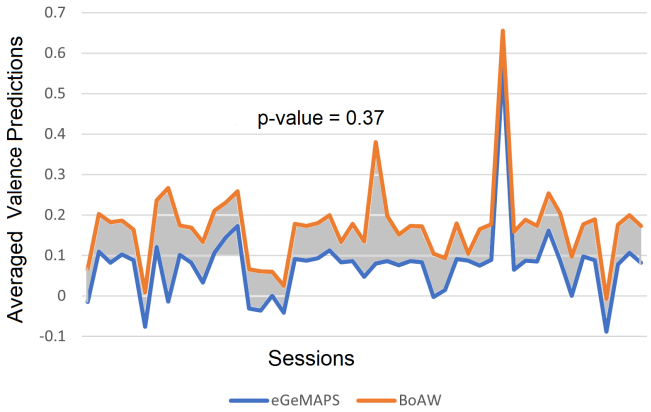


Fig. 7.5 Valence averaged predictions for full sessions using eGeMAPS and BoAWs features

unavailability of a cut-point score for the competency measure, the averaged score has been used, which equals 20. The figure shows that the eGeMAPS appear to have a larger effect size than the BoAWs features, particularly for the valence, which might be due to the longer segments that included both patient's and therapist's speech.

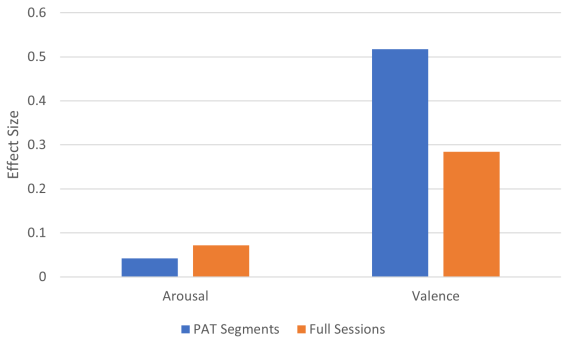


Fig. 7.6 The effect size between the eGeMAPS and BoAWs

Qualitative Analysis

A qualitative study has been conducted to understand the subtle changes in patient's emotions and emotions within the interaction between the patient and therapist. The predicted arousal and valence values and the interactional behaviours were qualitatively analysed across the sessions and several observations have been made in relation to the therapist's empathy and interpersonal synchrony. A series of case studies have been conducted inspecting individual sessions. Some individual patient's session have been further explored with a special emphasis on the patient's segments (case studies 1 and 2), while other analyses

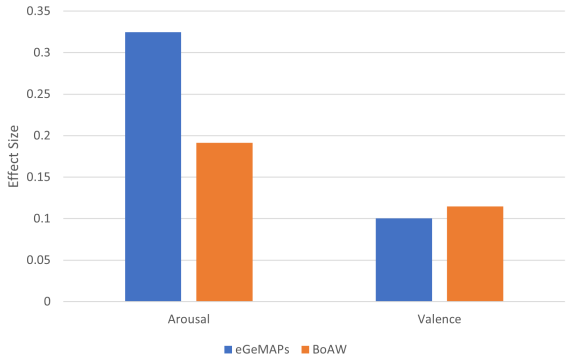


Fig. 7.7 The effect size between the predicted features for Low-GAD7 and High-GAD7

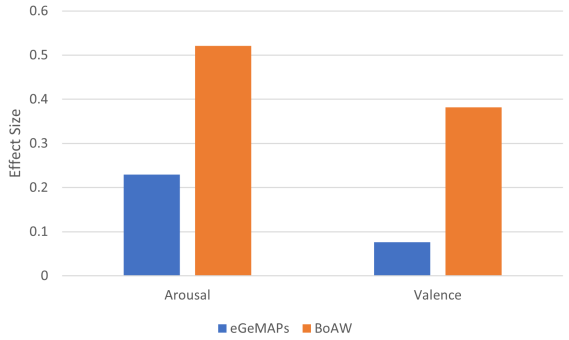


Fig. 7.8 The effect size between the predicted features for Low-PHQ9 and High-PHQ9

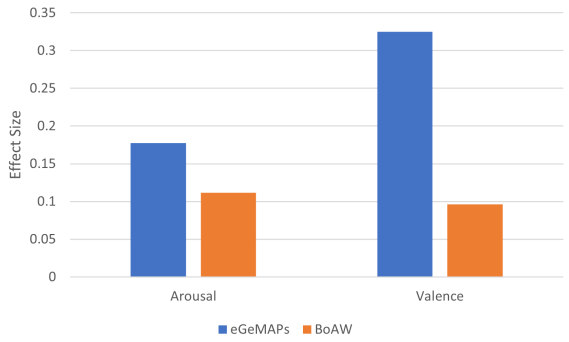


Fig. 7.9 The effect size between the predicted features for Low-Competency and High-Competency

explored the full session, including both the patient's and the therapist's segments (case studies 3 and 4). Each scene in the case studies related to the predicted values of arousal and valence is highlighted with an alphabetic letter to correspond to the resulting figure.

Case study 1: The patient S044CAT is male and aged 44. The session recorded is the 5th session and it is a face to face session. The total length of the patient's segments is 5

minutes and 37 seconds (337 seconds). The GAD-7 and PHQ-9 values for the recorded session are 5 (none) and 8 (mild), respectively. The competency measure for the therapist is 19.5 (medium). The predicted arousal and valence are presented in Figure 7.10. The vertical slashed lines in the figure indicate the borders of the patient's segments. In the session, the patient and the therapist discussed the patient's strength and resilience. The patient started by describing his strength, (A) then he engaged in describing an unpleasant memory which might decrease the valence. Afterwards, the therapist tried to motivate the patient gain a more positive attitude. The therapist asked about the patient's resilience, problem-solving, positive relationships and how the patient feels about it. It can be seen that there is a corresponding (B) increase in the predicted valence caused by the patient positively expressing his feelings toward this in a positive manner. The patient thinks a bit longer before answering questions, which might lead to (C) fluctuations in the valence. The patient spoke in a naturally aroused voice, which might be the cause of the stable arousal levels.

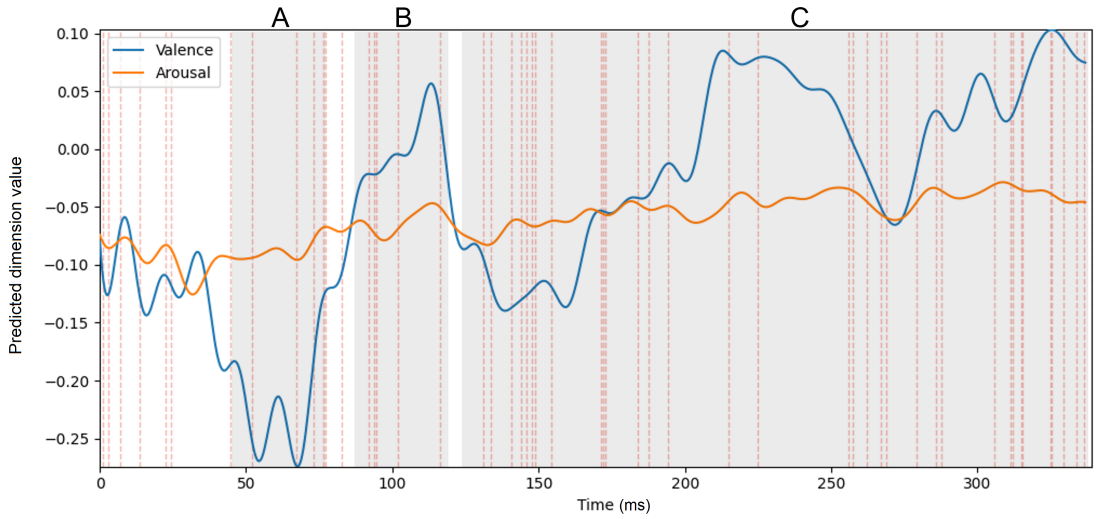


Fig. 7.10 The predicted dimensional emotion for patient S044CAT over time

Case study 2: The patient S214CAT is male and aged 22. The recorded session is the forth and it is a face to face session. The GAD-7 and PHQ-9 values for the recorded session are 6 (mild) and 10 (moderate), respectively. The competency measure for the therapist is 14 (Low). The total length of the patient's segments is 16 minutes and 4 seconds (963 seconds). The predicted arousal and valence are presented in Figure 7.11. The session started with the therapist asking the patient about his last week. The patient started by describing his health problem in the hospital that was worrying him, but he tried to calm himself by mentioning that he could be discharged and get well. This might be the cause of (A) an increase in the

arousal and valance to positive values. Another (B) sharp rise in arousal and valence was when the therapist asked the patient what might make him feel better. The patient answered by describing his work and how he enjoyed doing that. The patient moved towards more negative emotions when describing the worrying part in his career and how he might fail. During this part, it is seen that (C) the arousal and valence values move toward more negative values. He explained how this worry could be seen as a positive in the case of his work and the therapist agreed. This is seen to lead to (D) a rise in the emotional dimensions. The therapist then asked the patient for more worrying issues during the last week. The patient described his relationship with a family member and how this is affecting his life, leading to (E) low and fluctuating arousal and valence values for around 200 seconds. Another (F) sharp emotional rise is seen when the patient described what motivates him. Toward the end of the session, the therapist tries to guide the patient in the session to improve his stress by being less critical and more encouraging, leading to (G) stability in the arousal and valance values. At the last minute of the session, the therapist asked the patient about his genuine emotions in facing those stress points. The patient answered "nervous and uncertain", which led to (H) a drop in the arousal and the valence values. The final (I) rise in the emotions was due to the patient reflecting on the therapist's remarks on the positive consequence of facing worry and stress.

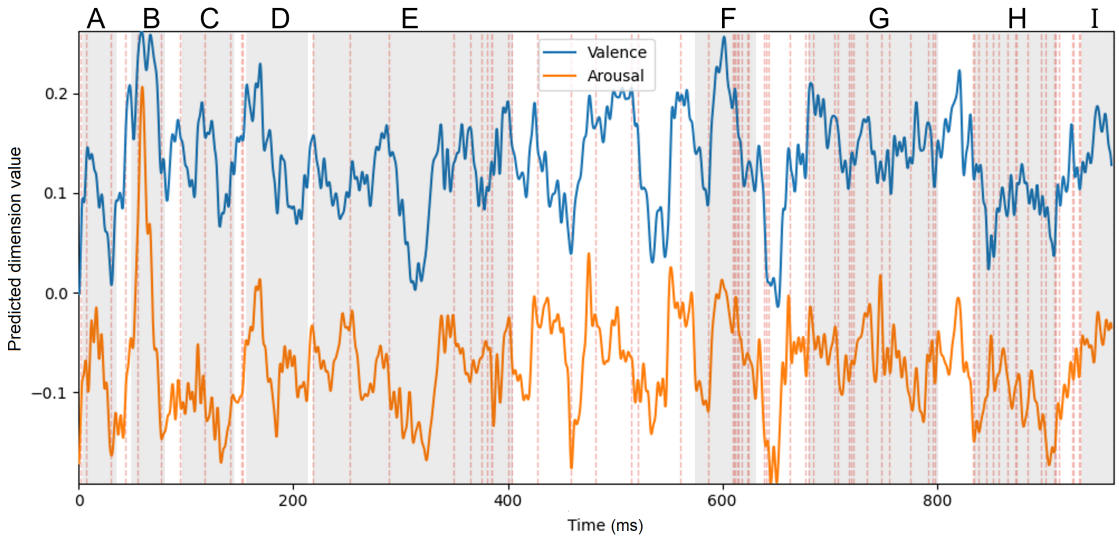


Fig. 7.11 The predicted dimensional emotion for patient S214CAT over time

Case study 3: The patient S067CAT is female and aged 59. The number of the session recorded is session six and it is face to face session. The GAD-7 and PHQ-9 values for the

recorded session are 7 (mild) and 9 (mild), respectively. The competency measure for the therapist is 19 (medium). The session recording length is 26 minutes and 43 seconds (1603 seconds). The predicted arousal and valence over the whole session are presented in Figure 7.12. The red highlighted regions are the patient's segments, while the green highlighted segments are therapist ones. The patient started the session by describing the difficulties she had faced since the last session, which led to (A) a drop in arousal and valence values. The therapist is trying to revise the last session with (B) positive arousal and valence moving forward to a more empathetic role toward the patient. Then, the patient starters to describe her role toward therapy and what struggles she faced. As a result, there were (C) fluctuations in the emotional dimensions. Then, the therapist tried to highlight the advantages of the therapy compared with before therapy. The patient approves of those advantages and (D) an increase in emotions was noticed in that part of the session. However, the sixth session is the final session in therapy for this particular patient and the patient is still struggling with stress. That is clear in the (E) fluctuation in the dimensional emotions. Furthermore, it is clear that the therapist is playing (F) an empathetic role and trying to sync the patient's emotions. Toward the end of the session, the therapist tries to guide the patient to resolve the stress and worry. The therapist scored with medium competency and that is clear in the session that the therapist tried to assist the patient in a treatment session. Still, there is more room for improvements and guides toward therapy.

Case study 4: The patient S140CAT is a female and aged 46. The recorded session is the first and it is face to face session. The GAD-7 and PHQ-9 values for the recorded session is 18 (severe) and 25 (severe), respectively. The competency measure for the therapist is 22.5 (high). The session recording length is 29 minutes and 37 seconds (1777 seconds). The predicted arousal and valence over the whole session are presented in Figure 7.13. Considering this is the first session, the therapist tries to give an overall introduction to the therapy ahead and talks about what feedback is required from the patient. This is clearly reflected in the (A) predicted emotions that trend toward naturalness. The patient described her struggles with medicines earlier in the session, which was evident in the (B) low arousal and valence values. The therapist noticed the severeness in depression and anxiety measures reflected in the questionnaires and asked about the thoughts of harm. In this part of the session, the patient starts to describe very negative thoughts in therapy. It is evident in the predicted emotions that (C) the arousal reaches -0.3, which is the lowest value for arousal in the whole session. At that moment, the therapist asked the patient, "What things of life that helped to make you safe?". This evidences the high ratings of the therapist's competence. Also, the predicted emotions lends evidence to this by the (D) sharp increase in the arousal

and valance for the patient. The therapist has tried to catch up with the patient since that last meeting in the assessment phase. The patient attempts to describe her problem and what she has been through in her life. The therapist feeds back with an empathetic voice in a sensible way. Toward the end of the therapy, the therapist guides the patient toward treatment and give her a road map for the therapy workbook. Furthermore, the therapist built a rapport relationship with the patient to improve the working alliance shown in the (E) positive arousal and valence values toward the end of the therapy.

Some common themes were noticed in the predicted arousal and valence values during the full-session qualitative analysis.

- In sessions recorded earlier in the course of the therapy treatment, there were more patient speaking turns. This has been especially true in sessions with high therapist's competence ratings. This is likely because the patient in those earlier sessions has been invited to spend time describing their issues and how these affect their normal life.
- In the later sessions of the course of treatment, the therapists have more speaking turns, which could relate to the therapist trying to resolve the patient's problems and assist them in dealing with those mental difficulties.
- In the sessions rated with high competency, the therapist tried to sync the patient's emotions and shows various emotions as a sign of approval and empathy toward the patient. In contrast, sessions with a corresponding low competency rating show that the therapist does not reveal a variation in emotions or show natural emotions along with the session.
- The sessions rated with medium competency showed a combination results of high and low rating sessions results.

7.6 Summary

In this chapter, the efficacy of predicting continuous emotional labels on the THEPS dataset has been investigated. The AVEC 2018 challenge has been used as a baseline for building and training the SER system. Due to the unavailability of dimensional emotion labels in the THEPS dataset, the RECOLA database has been used as a benchmark database for training the system. The trained system has been analysed on the THEPS dataset. The feature set used for training the system were eGeMAPS and BoAWs. After calculating the affect size between the eGeMAPS and BoAWs for the valence dimension, the BoAWs features achieved a higher affect size (0.39) comparable to the eGeMAPS features (0.09) which outlined with the AVEC

2018 baseline results, especially for the case of depression. Furthermore, several remarks were reported in the qualitative study relating to the number of the speaker's speaking turns and the therapist's competence in several stages in the therapy treatment. The next chapter will investigate one of the modules suggested in the proposed system that is the prediction of the competency measure.

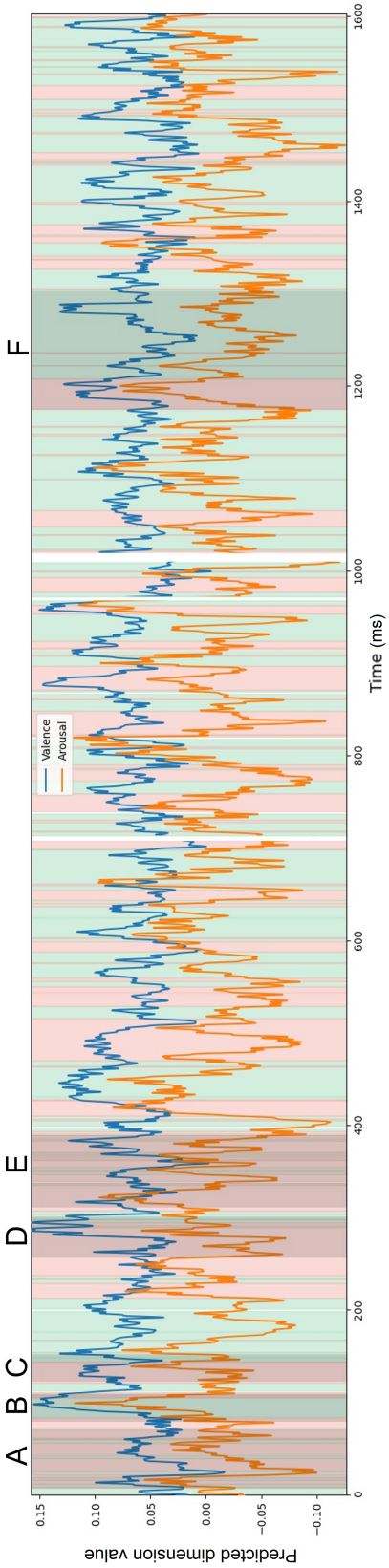


Fig. 7.12 The predicted dimensional emotion for the full session for patient S067CAT over time

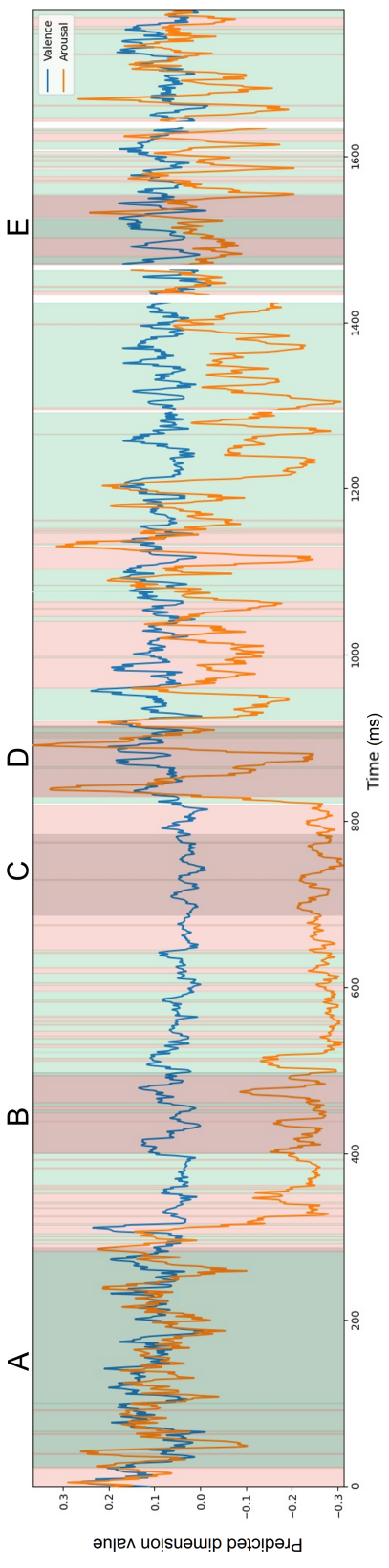


Fig. 7.13 The predicted dimensional emotion for the full session for patient S140CAT over time

Chapter 8

Automatic Prediction of Competency Measures

The prediction of the emotional dimensions for the therapist and the patient, presented in Chapter 7, has shown the ability of the acoustic features, Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) and Bag of Audio Words (BoAWs), to predict the continuous emotional labels in the THEPS dataset. Furthermore, the extracted acoustic features have shown a moderately sized effect on predicting the emotional labels for low and high groupings of therapist's competency measures. This chapter investigates the feasibility of automatically predicting a therapist's competency measure from patient's and therapist's speech using the THEPS dataset. Both acoustic and linguistic features will be explored in this chapter and the automatic transcription outputs from Chapter 5 will be used in this chapter.

8.1 Introduction

A therapist's competence is considered one of the criteria that could influence therapy quality. Therapy quality estimates the standard of delivering a psychological treatment in a qualified way to achieve its expected effects and eventually enhance the patient's outcome (Kohrt et al., 2015). As mentioned in Section 2.3, there are several methods for assessing a therapist's competence, such as the evaluation of treatment sessions using raters who evaluate the therapist's performance in delivering a treatment. The ratings are applied based on the presence and quality of certain therapist-determined features using a scaled standardised procedure, such as the Low-Intensity CBT (LI-CBT) treatment competency scale mentioned in Section 2.3. These raters could be persons who provide supervision for the therapists for training purposes. Supervisors prepare and train the therapists to be competent, committed

and effective psychotherapy practitioners. There are some concerns relating to the rating process, such as time-consuming, the variability of those ratings and the bias of ratings applied by supervisors who might know the therapists (Fairburn and Cooper, 2011; Watkins Jr, 2012). Therefore, there is a need for an automatic tool to predict a therapist's competency ratings. To the best of our knowledge, this is the first work to predict competency ratings automatically using the LI-CBT treatment competency scale.

The therapist's skills, such as competence, have been found to contribute positively to the therapeutic alliance. Furthermore, the therapist's personal qualities and techniques have a positive influence on the understanding or repair of ruptures in the therapeutic alliance (Ackerman and Hilsenroth, 2003). As shown in Figure 3.1, the detection of the therapeutic alliance is considered one of the blocks in the full system diagram. Predicting a therapist's competence automatically could assist in detecting the therapeutic alliance due to the positive connection between the two.

The LI-CBT treatment competency scale is a scaled measure used for rating therapists during treatment sessions. As mentioned in Section 2.3, the scale consists of six items that qualify the raters to inspect a range of competencies, including focusing the session, continued engagement competencies, interpersonal competencies, information gathering: specific to change, within session self-help change method and planning and shared decision making competencies (Kellett et al., 2021b). According to the manual used by the raters as a guide for the rating process, several scale items have been found to be directly or indirectly associated with the emotional and empathetic behaviours expressed by the therapist and the patient. Furthermore, the results reported in Chapter 7 demonstrated the feasibility of predicting the continuous emotional labels from the THEPS dataset using acoustic features (eGeMAPS and BoAWs) with respect to the therapist's competency measure. Therefore, those acoustic features could be a suitable indicator for predicting the therapist's competency measure. On the other hand, the speakers' language could reveal the emotions raised in the session, along with the degree of that emotion. The use of language-based features to capture speakers' emotions could improve the accuracy of the prediction of the competency measures. As mentioned in Section 4.3.1, the Referential Activity (RA) is a dictionary-based model that estimates the degree to which a person's language is associated with emotions. Due to the strong association between a speaker's emotional behaviours and the competency measures, the RA model scores have been implemented as the language-based (linguistic) features in this chapter. The linguistic features could provide complementary information to the acoustic features. Consequently, the fusion of acoustic and linguistic features mentioned earlier has been examined to predict the therapist's competency measure.

The rest of this chapter is organised as follows: Section 8.2 describes the related work for predicting therapist's competency measure, Section 8.3 presents the data description and analysis from a linguistic perspective , Section 8.4 explains the experiment of the automatic prediction of a therapist's competency measure, Section 8.5 presents the results of the experiment and Section 8.6 summarises the chapter's findings.

8.2 Related Work

A growing body of literature recognises the importance of acoustic features in the field of therapy. A study by Nasir et al. (2017) examined the significance of several acoustic features extracted from recordings of couples therapy interactions to predict the success or failure of the couples' marriages. In addition, the researchers explored behavioural codes rated by human experts as a feature of marital outcome prediction. Two stages of preprocessing were implemented: Voice Activity Detection (VAD) and speaker diarisation. The acoustic features were extracted for each speaker individually across the sessions. They extracted several acoustic features, such as speech prosody (pitch and energy), voice quality (jitter, shimmer) and spectral envelope characteristics which are Mel Frequency Spectrum Coefficients (MFCCs). Then, several functionals were computed for all the former acoustic features, such as mean, minimum and maximum. The results showed that the acoustic features outperformed the behavioural codes in predicting the state of the marriage. The fusion of both acoustic features and behavioural codes showed the best results in outcome prediction. Amir et al. (2010) analysed a corpus of speech recorded during psychotherapy focused on tackling unresolved anger towards an attachment figure. The recordings were from therapy sessions of 22 females; 283 stimuli were extracted and evaluated for emotional content by 14 judges. The emotions were rated dimensionally according to three scales: activation, valence and dominance. The features used for classification were acoustic features representing several prosody components: Fundamental frequency (F0), intensity, duration and voice quality. The automatic classification results showed that the acoustic features were better at predicting activation comparable with valence and dominance such that the features based on F0 were the dominant features.

In a recent study related to competency by Sümer et al. (2021), automatic nonverbal analysis of presentation competency estimations was applied to presenters' behaviours during a speech from audiovisual recordings of a real-world setting. The presentation competency rating consisted of six items: addressing the audience, structure, language use, body language and voice, visual aids and content credibility. The research used several modalities to extract features: speech (eGeMAPS), facial (head pose, gaze direction and facial action units)

and body poses (the estimated locations of body joints). They found that acoustic features transcended face and body pose features in both classification and regression tasks. The study highlighted that speech features were founded to be the most dominant nonverbal cues used to estimate presentation competence.

People use language to express their behaviours and actions. From a conversational therapeutic perspective, the linguistic features could be organised into four groups of features: sentence features (sentence position and length), context features (labels and N-grams), dialogue features (speaker change and turn index) and sentiment features (counts of words that can be determined from two common psychological dictionaries, Linguistic Inquiry and Word Count [LIWC] and Discourse Attributes Analysis Program [DAAP]) (Pennebaker et al., 2015; Bucci and Maskit, 2005; Lee et al., 2019). A study by Flemotomos et al. (2018) predicted the quality of Cognitive Behaviour Therapy (CBT) using a gold standard measure for CBT quality. The used features are a set of occurrences of N-grams in the sessions that is documented by weighting each N-gram by the Term Frequency-Inverse Document Frequency (TF-IDF), which is a measure for how relevant a word is to a document, along with the LIWC features. The results indicated that the therapist-related features have more predictive power than the patient-related features.

Several studies have explored the use of sentiment features in the field of therapeutic alliance. Atta et al. (2019) examined how DAAP, as a linguistic measure, could indicate the correlation between emotional elaboration and therapeutic alliance within a single session. In the study, they allocated 40 patients with varying diagnoses to be videotaped, transcribed and analysed using linguistic measures of the referential process (DAAP), followed by a human-centered scoring with the Working Alliance Inventory (WAI) for every 5-minute interval. The results showed that if the ratings of the patients indicated more emotional engagement with their experience and was followed by an experience reflection by mid-session, those patients would have higher scores in the therapeutic alliance by the final part of that same session. Recent work by Christian et al. (2021) has investigated the connection between non-verbal emotional experiences and how verbal language is affected by ruptures during treatment. They employed linguistic measures of the referential process in association with measures of the therapeutic alliance. They employed a scored measure to identify ruptured from non-ruptured segments in 27 psychotherapy sessions. The classified segments were scored based on the key linguistic dimensions of the referential process. The results showed that, during ruptured segments, the patients showed a decrease in emotional engagement, an increase in negation as compared to non-ruptured segments and an increase in a measure of distancing. The therapists showed similar patterns to the patients during rupture, in addition

to self-disclosure, increased attempts at emotional control, increased references to bodily experiences and natural affect words.

It is well known that the use of linguistic features could complement the use of acoustic features in classifying or predicting common observations in the medical field. To further investigate the role of combining acoustic and linguistic features, Wiegersma et al. (2020) carried out a study to examine the possibility of recognising *hotspots* for patients experiencing a trauma-focused treatment for Post-Traumatic Stress Disorder (PTSD). Hotspots can be defined as moments of traumatic experience that include the highest emotional impact. To discriminate between hotspot and non-hotspot phases, they extracted a combination of text and speech features from recordings and transcriptions of patients' CBT sessions to develop an automatic supervised classification model. They mapped nine properties that could differ between hotspots and non-hotspots to acoustic and linguistic features. The linguistic features used were N-grams, grammatical tags, lexicon-based tags and LIWC features. The acoustic features used were pitch, loudness, duration, spectral features and voice quality features. The results gained from the study showed that combining the best linguistic and speech features could train a sufficient model to discriminate between hotspots and non-hotspots. Furthermore, the model trained separately on acoustic features showed less training performance.

A study conducted by Tavabi et al. (2020) involved analysing behavioural cues in a patient's language and speech that indicate behaviour towards change across a therapy session. They used a dyadic Motivational Interviewing (MI) dataset that involved patients with alcohol-related problems. The dataset was labelled with the Motivational Interviewing Skill Code (MISC) for therapist and patient language during the MI sessions to permit the analysis of the link between the patient's language and the subsequent behavioural outcome. The patient utterances were sorted based on the language suggesting willingness or resistance to change such that the labels were categorised into three scales: Change Talk, Sustain Talk and Follow/Neutral. The textual features extracted were: the Bidirectional Encoder Representations from Transformers (BERT), a pre-trained language model extracted per utterance for both therapist and patient and the LIWC features. They used two different speech representations and feature sets, eGeMAPS and a pre-trained deep convolutional neural network pre-trained on audio spectograms extracted from a large database of audio event categories. The study results showed that the multimodal that fused the acoustic and linguistic features slightly underperformed the text unimodal, which they believed was mainly due to the low-quality speech files.

Overall, these studies highlighted the beneficial effects of the acoustic and linguistic features and their fusion in the counselling domains. In this chapter, to predict the therapist's

competency measure, the fusion of the acoustic features (eGeMAPS and BoAWs) and the linguistic features (DAAP) has been investigated as a complete feature set for the prediction model.

8.3 Data Analysis

The data used in this chapter is the transcribed part of the THEPS dataset described in Section 3.3. Figure 8.1 shows the part of the dataset used in this chapter in relation to the total THEPS dataset.

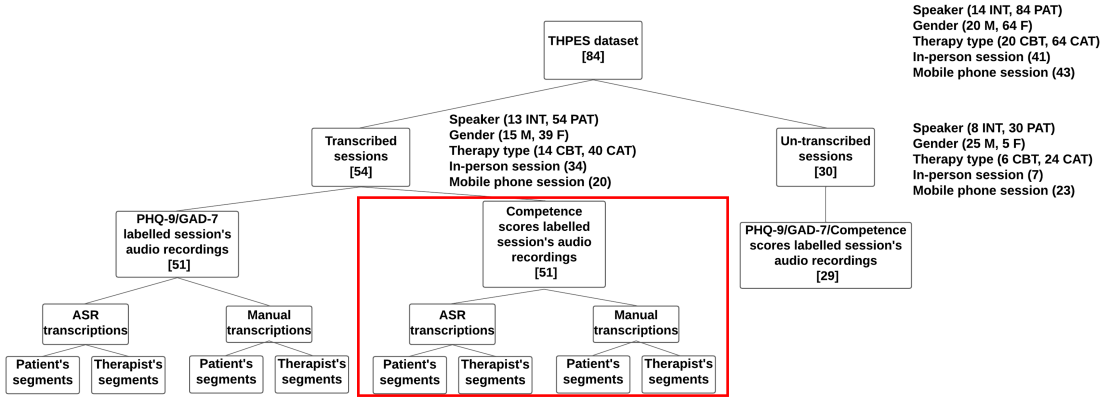
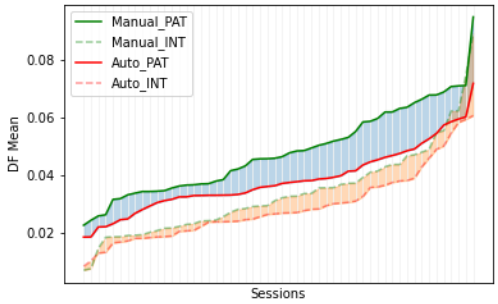


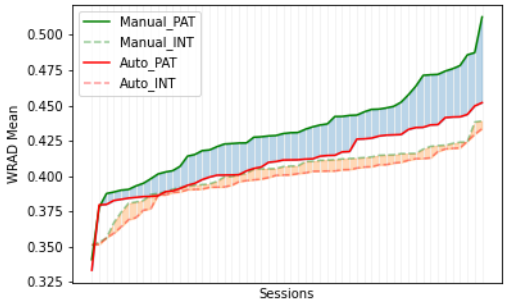
Fig. 8.1 THEPS dataset tree diagram highlighting the part used in this chapter.

The reason behind selecting this part of the dataset is the need for the transcriptions to extract the linguistic features and the need for the annotation information to extract the acoustic and linguistic features for each speaker independently. Both the manual transcriptions and the automatic ones gained from the Automatic Speech Recognition (ASR) in Chapter 5 have been used as inputs for the therapist's competence measure prediction system. The therapist's and patient's segments have been joined in a single speech recording per speaker based on each session. Afterwards, the acoustic and linguistic features have been extracted for each speaker per session.

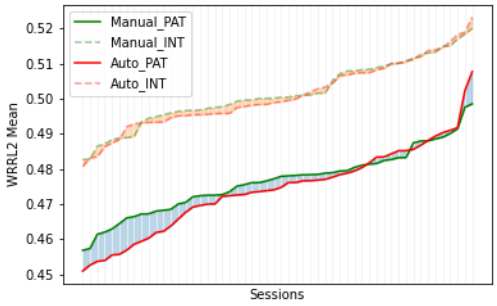
To validate the arguments in the literature relating to the DAAP linguistic dictionaries scores for patients and therapists (described in Section 4.3.1), several dictionary scores have been presented in Figure 8.2 based on manual and automatic transcriptions for patient's and therapist's speech, individually. Figure 8.2a shows the mean scores of the Disfluency Dictionary (DF) for each speaker. It is clear from the figure that the patient's speech achieved higher disfluency mean rates as compared to the therapist's speech in both the automatic and manual transcriptions and that is relate to the emotional state the patient is experiencing



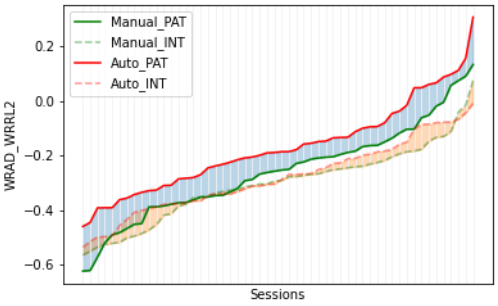
(a) Disfluency dictionary mean score



(b) WRAD dictionary mean score



(c) WRRL dictionary mean score



(d) WRAD/WRRL covariation score

Fig. 8.2 Dictionary scores based on manual and automatic transcriptions (The straight green lines represent the results for the patient's speech using the manual transcripts, the dashed green lines represent the results for the therapist's speech using the manual transcripts, the straight red lines represent the results for the patient's speech using the automatic transcripts and the dashed red lines represent the results for the therapist's speech using the automated transcripts.)

in the session (Atta et al., 2019). Furthermore, the manual transcriptions resulted in higher disfluency means than the automatic ones, since the manual transcriptions represent the gold standard among the transcriptions.

As presented in Section 4.3.1, the DAAP consists of several dictionaries, each with a specific psychological function. The Weighted Referential Activity Dictionary (WRAD) dictionary measures the RA ratings that captures the full speaker engagement in a narrative through language. High scores for WRAD indicate the symbolising phase in the referential process. In the symbolising phase, the patient verbally expresses a memory, event, or dream narratively as if they were there in time and place. The Weighted Reflection/Recognising List (WRRL) measures the reflection/recognising phase in the referential process, which is the

degree to which a speaker is trying to understand and recognise the emotional significance of an event or set of events in their own life or the life of someone else being capable of reflecting on it as if from outside (Atta et al., 2019; Maskit, 2021; Maskit et al., 2005). Figure 8.2b shows the mean score of the WRAD dictionary for each speaker. The scores shown in the figure reflect those of Maskit (2021), who also found that patient's sessions mean WRAD scores are higher than the therapist's such that the patient has the privilege of telling a story comparable to the therapist. The mean scores of the WRRL are shown in Figure 8.2c. As can be seen from the figure, the WRRL mean scores for the therapists are higher than those for the patients, because the therapists are reflecting on the patients' emotional behaviours. This finding has also been reported by Maskit (2021). Furthermore, there is a slight difference between the WRRL mean scores in the manual transcriptions and the automatic ones, which may relate to the therapist's language being easier to recognise by the ASR system.

The covariation between any two dictionary scores is a measure of how the two dictionaries move together or in different directions. Positive covariation reflects that the two measures move in the same direction most of the time, while negative covariation shows that the two measures move in opposite directions most of the time. The covariation between WRAD and the WRRL scores is shown in Figure 8.2d. It appears from the figure that the reported scores are mainly negative and the scores based on the manual transcriptions are lower than the automatic ones. Maskit (2021) highlighted that a person's engagement in the referential process is that the procedures of symbolising and reflecting/recognising are separated in that the patient does not describe a story and reflect on its meaning at the same moment. Taking account of that, the session level of WRAD and WRRL covariation should be mainly negative, as shown in Figure 8.2d, such that when WRAD or WRRL is low, the other is high. Reporting larger negative scores based on the manual transcriptions verifies the correct pattern described earlier in (Maskit, 2021) relating to the person's engagement in the referential process.

8.4 Automatic Prediction of Competency Measures

The pipeline of the therapist's competency measure prediction system consists of: feature extraction, feature selection and classification/regression.

8.4.1 Feature Extraction and Selection

The extracted features used in this chapter consist of acoustic and linguistic features for each speaker individually. The acoustic features extracted are eGeMAPS and BoAWs using

the same settings described in Section 7.4.1. Since those features are extracted based on a variable number of frames for each session, majority voting has been applied to extract a single feature vector for each session. The resulting acoustic feature vector consists of 88 eGeMAPS and 100 BoAWs for each speaker, for a total of 376 acoustic features.

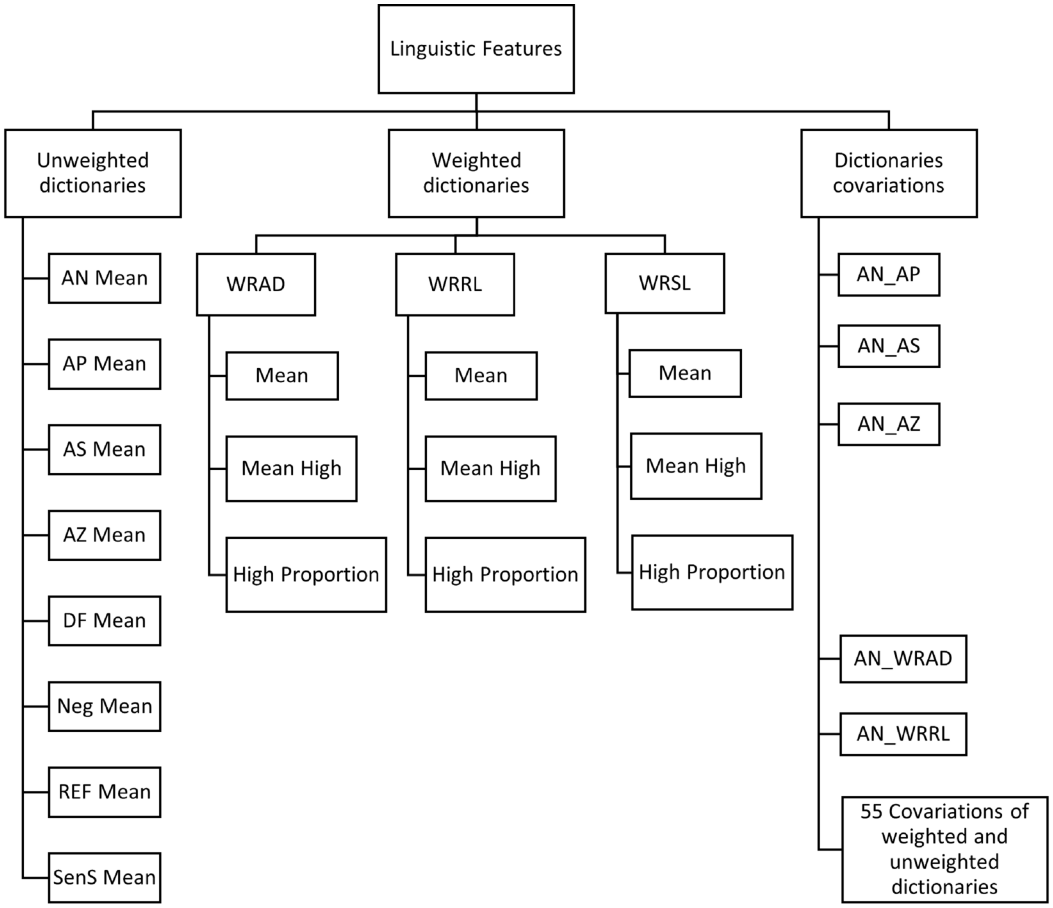


Fig. 8.3 Extracted DAAP linguistic features tree diagram

The linguistic features are extracted based on the DAAP dictionaries scores, as described in Section 4.3.1. As shown in Figure 8.3, the linguistic features consist of three main categories: unweighted dictionaries, weighted dictionaries and dictionary covariations. The DAAP calculates several measures for the dictionaries. The mean score of each unweighted dictionary value has been computed for each speaker to be included in the linguistic feature vector (Murphy et al., 2015). Each weighted dictionary has been scored based on the mean, mean high and high proportion. The mean score is the average amount of the speaker's weighted dictionary value. The Mean High (MHigh) score is the average amount by which the speaker's weighted dictionary value exceeded the neutral value of 0.5. The High Proportion (HighP) score is the proportion of words in a specified segment for which the weighted

dictionary value is greater than its neutral value of 0.5. The dictionary covariations are the covariation scores between each pair of the unweighted and/or weighted dictionary mean values. The covariations comprised 55 covariations, such as Affect Negative (AN) and Affect Positive (AP) covariation scores, AN and WRAD covariation scores, and WRAD and WRRL covariation scores. The filler words existed in the THEPS dataset manual transcriptions described in Section 3.3 has been added to the filler pauses words existed in the DF dictionary. The total linguistic features extracted based on the DAAP measures are 72 features for each speaker, for a total of 144 linguistic features.

After combining the acoustic and linguistic feature vectors, the total features used in the prediction system is composed of 520 features for both speakers. The Recursive Feature Elimination (RFE) approach (described in Section 6.5.1) has been applied for the feature selection phase in the experiment.

8.4.2 Therapist's Competency Ratings Prediction and Classification

Several regressors have been compared for predicting the competency ratings, such as SVR, Lasso, Linear Regression, Elastic Net, Decision Tree Regressor, Ada-Boost Regressor and Gradient Boosting Regressor. Lasso and SVR gained the higher correlation coefficient results and the lower error rates in predicting the competency ratings as shown in Table 8.1, which aligned with the results gained from the Audio/Visual Emotion Challenge (AVEC) 2018 challenge baseline for dimensional emotions recognition described in Section 7.4.2 (Ringeval et al., 2018). The features selection step involved using a cross validation technique adopting the approach of Recursive Feature Elimination Cross-Validation (RFECV) as described in Section 6.5.1. The number of folds selected in the cross validation process has been six folds, taking into consideration that the dataset is a relatively small and imbalanced dataset. A grid search has been implemented on the best classifier or regressor to find the optimal parameters. Using the best classifier and regressor with the optimal parameters on the same number of folds, the selected features has been used for predicting the competency rating in total and per each item on the competency rating scale. Furthermore, the therapist's level of competence has been classified based on various levels of patient's depression and anxiety. The best acoustic and linguistic features have been reported based on the total ratings and per each item in the competency rating scale for patients and therapists.

8.5 Results

This section presents the prediction results using the Pearson Correlation Coefficient (R), Mean Absolute error (MAE) and Root Mean Squared Error (RMSE) for the total therapist's competency rating and each rating item and the classification results based on the confusion matrix. Furthermore, the best acoustic and linguistic features for predicting the total competency rating and each rating item has been presented in this section. Lasso and SVR performed better than Linear Regression, Elastic Net, Decision Tree Regressor, Ada-Boost Regressor and Gradient Boosting Regressor. A grid search resulted in the following parameters being set: SVR with a C of 0.1 and epsilon of 1.00 with linear kernel, and Lasso with alpha of 0.03. The results have been reported based on R, MAE and RMSE. Table 8.1 presents the total therapist's competency rating prediction results using several regressors for manual and automatic transcriptions. The results using the Lasso regressor for the manual transcriptions are higher than the SVR results, with a higher correlation coefficient results and lower error rates. Furthermore, the manual transcriptions gained a higher correlation coefficient than the automatic ones. Based on these results, the latter investigation in this chapter concentrates on the use of Lasso and the manual transcripts. Because this is the first study that investigated the therapist's competence, it would be interesting to explore the results based on the manual transcripts, especially the features that could accurately describe the competence from an acoustic and linguistic perspective, taking into consideration that the WER of the ASR is around 30, as mentioned in Section 5.5, which might lead to a high number of misleading selected features.

Table 8.1 The total therapist's competency rating prediction results using several regressors for manual and automatic transcriptions (the MAE and RMSE percentages of the total dataset).

Regressor	Automatic Transcripts			Manual Transcripts		
	MAE	RMSE	R	MAE	RMSE	R
Lasso	1.91 (3.76%)	2.64 (5.72%)	0.88	**1.66 (3.25%)**	**2.25 (4.42%)**	**0.92**
SVR	2.91 (5.72%)	3.67 (7.19%)	0.84	2.65 (5.20%)	3.12 (6.11%)	0.91
Elastic Net	3.51 (6.88%)	4.30 (8.43%)	0.76	3.39 (6.64%)	4.24 (8.31%)	0.77
Linear Regression	3.98 (7.80%)	4.91 (9.62%)	0.60	3.80 (7.45%)	4.59 (9.00%)	0.69
Decision Tree	3.86 (7.56%)	6.24 (12.23%)	0.45	3.61 (7.07%)	5.39 (10.56%)	0.51
Ada-Boost	3.54 (6.94%)	5.53 (10.84%)	0.45	3.50 (6.86%)	5.25 (10.29%)	0.47
Gradient Boosting	4.31 (8.45%)	5.77 (11.31%)	0.44	3.51 (6.88%)	4.80 (9.41%)	0.50

The selected acoustic and linguistic features for predicting the total therapist's competency measure are presented in Figure 8.4 for patient's and therapist's speech. As shown in the figure, the best eGeMAPS features for the patient is the standard diversion of the third

formant frequency and this feature could reflect on the vowel pronunciation in the patient's speech. For the therapist, the mean of the third MFCC coefficient has been selected as the most optimal feature for predicting the competency that could reflect on the therapist's voice timbre (De Boer et al., 2021). The best linguistic feature for the patient is the MHigh of WRAD that means the patient is in the RA symbolising phase, verbally immersed in a memory or a dream which is aligned with the patterns reported earlier in Section 8.3. For the therapist, the best linguistic features are the WRRL MHigh, the AP and Affect Sum (AS) covariation, the Neutral Affect (AZ) and Weighted Arousal List (WRSL) covariation and the WRAD and WRSL covariation. The former best features can be divided into two groups based on the therapist's levels of competence: the high competency therapists mainly reflect on the patient's mental experience, which is captured by WRRL MHigh, with a positive effect showing a range of emotions, which is captured by AP and AS covariation. Low to medium competency therapists express with a AZ located either in the arousal phase of the referential process (WRSL) or in the symbolising phase (WRAD). Measuring the covariation between those dictionaries could capture the low to medium level of therapist's competence, as found in the selected features by the prediction model. It is difficult to explain the best BoAWs features, because each selected word corresponds to a combination of MFCC acoustic features.

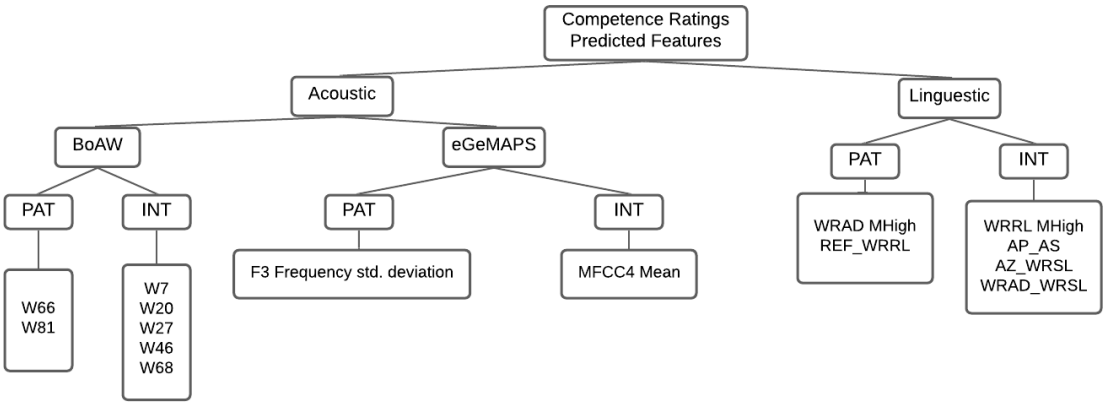


Fig. 8.4 The best acoustic and linguistic features for predicting the total competency measure (std. deviation = standard deviation, PAT = patient, INT = therapist)

The actual versus predicted total therapist's competency ratings are presented in Figure 8.5 based on each level of competence. From the figure, it is clear that the prediction results gained better results on the medium level of competence, which is highly related to the number of occurrences of this level in the dataset.

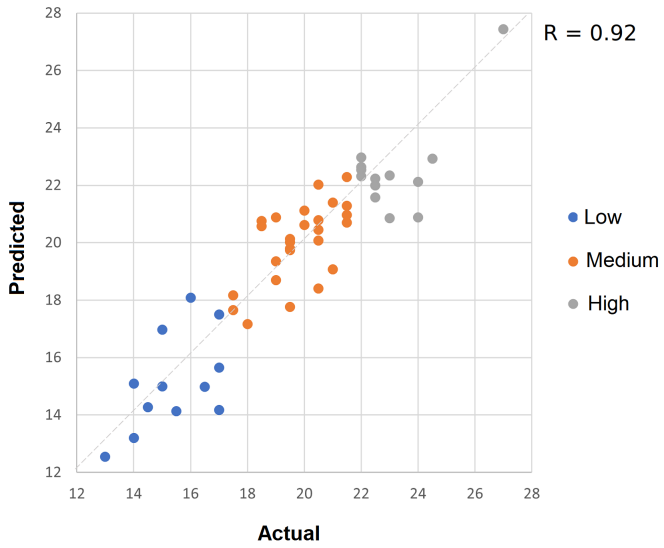


Fig. 8.5 The actual versus predicted total therapist's competence measure highlighting each level of competence.

The therapist's competency measure is calculated based on six rating items: focusing the session, continued engagement competencies, interpersonal competencies, information gathering: specific to change, within session self-help change method, planning and shared decision-making competencies. The results of the therapist's competency measures based on each item are displayed in Table 8.2. It can be shown from the results in the table that *focusing the session* and *interpersonal competencies* rating items have been easier to predict with R 0.89 and 0.87, respectively. This indicates the feasibility of detecting the therapist's empathy, which is one of the rating specifications connected to the *interpersonal competencies* rating item as described in section 2.3. On the other hand, *information gathering* and *planning and shared decision-making competencies* rating items have been harder to predict with R 0.75 and 0.73, respectively. The number of features selected to predict the *continued engagement competencies* rating item has been the lowest number of features in comparison to the other items.

Table 8.2 The competency rating prediction results based on each item using Lasso for manual transcriptions (the MAE and RMSE percentage of the total dataset).

Competency Items	Num. Feat.	MAE	RMSE	R
1-Focusing the session	27/520	0.35 (6.86%)	0.44 (8.75%)	0.89
2-Continued engagement competencies	11/250	0.30 (6.02%)	0.38 (7.55%)	0.82
3-Interpersonal competencies	23/250	0.34 (6.76%)	0.42 (8.39%)	0.87
4-Information gathering: specific to change	33/250	0.51 (10.17%)	0.62 (12.31%)	0.75
5-Within session self-help change method	27/250	0.33 (6.64%)	0.41 (8.16%)	0.81
6-Planning and shared decision-making competencies	24/250	0.34 (6.70%)	0.45 (8.98%)	0.73

The best eGeMAPS features for predicting the full ratings and each rating item has been presented in Figure 8.6 for each speaker. It is clear that the most frequently selected features are: the mean of the third and fourth MFCC coefficients for voiced regions and the standard deviation of the slope of rising signal parts of F0. These features are a combination of spectral and frequency based features (Corrales-Astorgano et al., 2019). Furthermore, the best eGeMAPS features are composed of more patient related features compared to the therapist related features. This could contribute to the ability of acoustic features to reveal the patient behaviours that could not be captured by language.

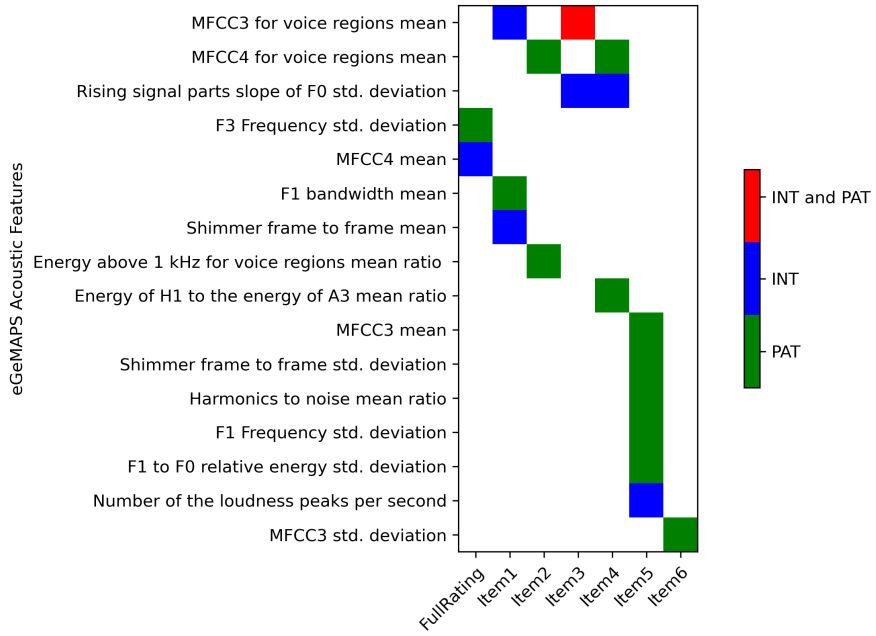


Fig. 8.6 The best eGeMAPS features represented on the horizontal axis, for predicting the full competency rating or each rating item represented on the vertical axis. The rating item numbers are based on the numbers in Table 8.2. The green blocks indicate the selected features for the patient's segments, the blue blocks indicate the selected features for the therapist's segments and the red blocks indicate the selected features for the full session, including the patient's and therapist's segments. (std. deviation = standard deviation, H1 = first F0 harmonic, A3 = highest harmonic in the third formant range).

The best linguistic features for predicting the total ratings and each rating item has been presented in Figure 8.7. The most common linguistic features are: AZ and WRSL covariation, WRAD and WRSL covariation, Sensory Somatic Dictionary (SenS) and WRSL covariation, AN and SenS covariation, WRSL mean, AZ mean and Negation Dictionary (Neg) mean. These features could indicate moments of rupture in the sessions based on the work of Christian et al. (2021), such that WRSL could indicate decrease in the emotional engagement for the patient and self-disclose for the therapist. Furthermore, the SenS score could map to

the therapist's increase in references to bodily experiences and the AZ score could capture the use of natural affect words. In addition, the number of the selected features based on the therapist's speech higher than the ones related to the patient because the most dominant approach for the therapist is to communicate and express verbally with the patient in the session. Due to the variations in each competency item characteristic, the best descriptive linguistic features have been presented in Table 8.4. The table shows that for each competency item, almost all of the linguistic features were aligned with the characteristics specified by the competency scale manual (Kellett et al., 2021b).

Each level of competence is classified using the manual and the automatic transcriptions based on the following classifiers: SVM, Decision Tree Classifier, Random Forest Classifier, Ada-Boost Classifier and Gradient Boosting Classifier. Table 8.3 presents the classification results for the manual and automatic transcriptions using the aforementioned classifiers. It is clear that the SVM achieved higher performance than the other classifiers in classifying the three levels of competence.

Table 8.3 The level of competence classification results using several classifiers for manual and automatic transcriptions (the standard deviation of the precision and recall).

Classifier	Automatic Transcripts			Manual Transcripts		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
SVM	0.61 (0.11)	0.53 (0.21)	56.86%	**0.92 (0.03)**	**0.91 (0.02)**	**90.20%**
Decision Tree	0.54 (0.09)	0.51 (0.13)	53.92%	0.63 (0.08)	0.65 (0.07)	66.66%
Random Forest	0.51 (0.08)	0.49 (0.16)	52.94%	0.62 (0.07)	0.59 (0.11)	62.09%
Ada-Boost	0.53 (0.07)	0.49 (0.16)	52.94%	0.57 (0.06)	0.55 (0.12)	58.33%
Gradient Boosting	0.52 (0.07)	0.50 (0.17)	53.33%	0.61 (0.06)	0.57 (0.10)	59.60%

The therapist's levels of competence have been classified using manual transcriptions based on the different levels of depression and anxiety. Figure 8.8 shows the confusion matrix results based on the patient's level of anxiety using the Generalised Anxiety Disorder (GAD-7) scores. In contrast, Figure 8.9 shows the confusion matrix results based on the patients' level of depression using the Patient Health Questionnaire (PHQ-9) scores. Due to a large number of occurrences of medium levels of competence in the THEPS dataset, the model predicts the medium level correctly most of the time and misclassifies the other levels as the medium level. This correlates with the results in Figure 8.5.

8.6 Summary

This chapter aims to predict the therapist's competency measures using patient's and therapist's speech in the THEPS dataset. The competency ratings used in the chapter is based

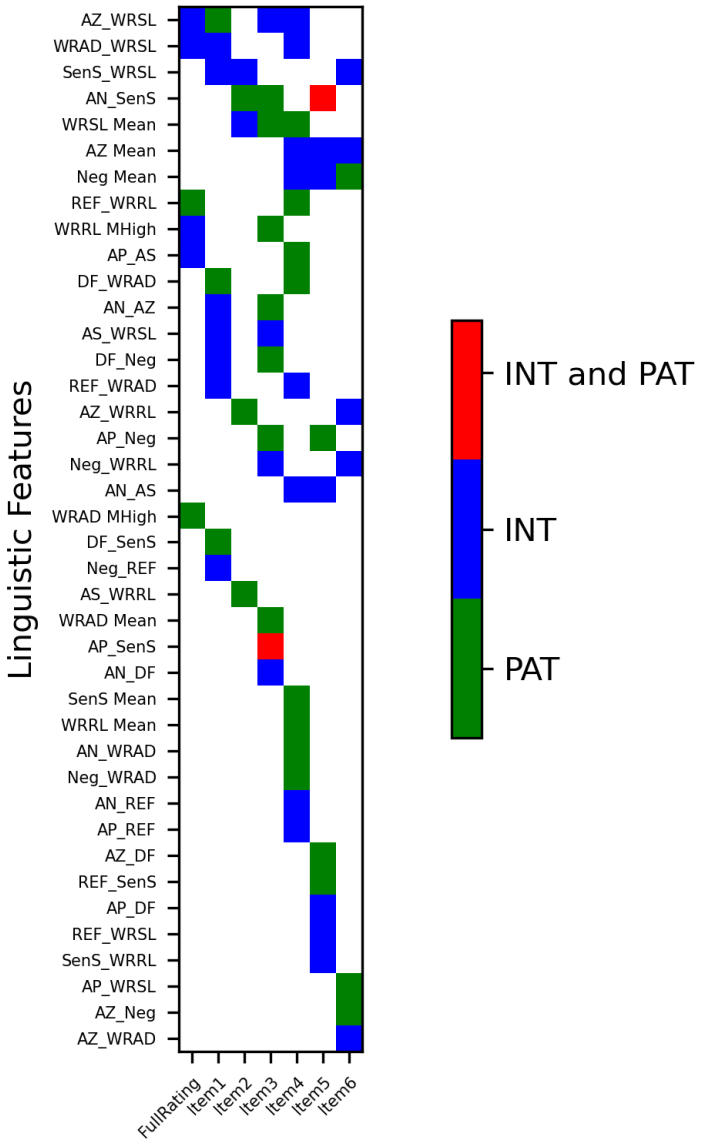


Fig. 8.7 The best linguistic features are represented on the horizontal axis, for predicting the full competency rating or each rating item represented on the vertical axis. The rating item numbers are based on the numbers in Table 8.2. The green blocks indicate the selected features for the patient's segments, the blue blocks indicate the selected features for the therapist's segments and the red blocks indicate the selected features for the full session, including the patient's and therapist's segments.

on the LI-CBT treatment competency scale. The THEPS dataset is labelled with the total ratings of the competency scale, a single score for each scale item from a six item based score and the levels of the competence measure. Since this is the first chapter that includes extracting linguistic features from the THEPS dataset, data analysis, based on the most

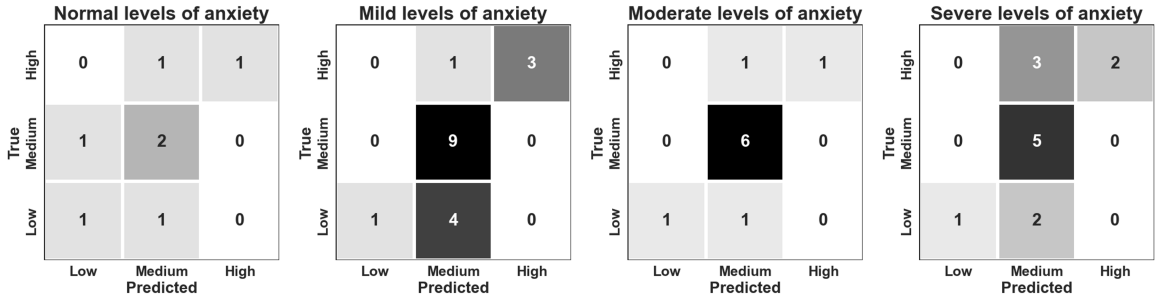


Fig. 8.8 The confusion matrix results of the therapist's level of competence (low,medium,high) based on the patient's level of anxiety (normal,mild,moderate and severe).

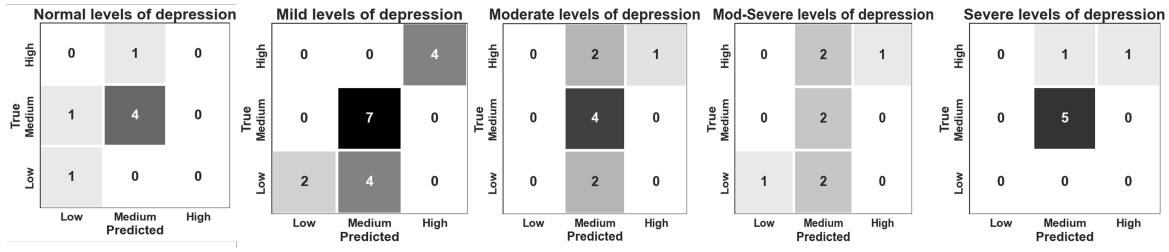


Fig. 8.9 The confusion matrix results of the therapist's levels of competence (low,medium,high) based on the patient's level of depression (normal,mild,moderate,moderately severe and severe).

dominant features in psychotherapy literature, has been used for the investigation. The features used in the prediction system are a combination of eGeMAPS and BoAWs as acoustic features and DAAP scores as a linguistic features. Feature extraction, feature selection and classification or regression have been implemented as blocks of the prediction model. The Lasso achieved the best prediction results with an R of 0.92 and lower error rates with an MAE of 1.66 and RMSE of 2.25. In addition, the best acoustic and linguistic features gained from the prediction system has been reported in the chapter. The next chapter will present the key findings for all the experiments conducted in this thesis and how they contribute to the literature.

Table 8.4 The best descriptive linguistic features based on each competency measure item and the corresponding rating characteristics.

Rating item	Rating characteristics	Best linguistic features	
		Therapist	Patient
Focusing the session	The treatment session introduction is the opportunity for the therapist to re-engage with the patient and to highlight the planned and agreed on content of the treatment.	Neg_REF	DF_SenS
Continued engagement competencies	The patient should feel positive and confident about what change they made/making or that their problems are being approached in a collaborative way. The therapist should confirm that progress is acknowledged by reflection and summaries.		AS_WRRL
Interpersonal competencies	The therapist should be able to sustain a trusting and containing therapeutic relationship with the patient. The therapist should ensure the patient has time to talk through any issues.	AN_DF,AP_SenS	WRAD mean,AP_SenS
Information gathering: specific to change	The therapist should inspire the patient to feel an active participant in the treatment process and to feel positive through the changes they made between the sessions.	AP_REF,AN_REF	SenS mean, WRRL mean, AN_WRAD,Neg_WRAD
Within session self-help change method	Cognitive restructuring is a way of influencing mood by targeting unhelpful thoughts via a process of identification and challenge.	SenS_WRRL,REF_WRSL, AP_DF	AZ_DF,REF_SenS
Planning and shared decision-making competencies	The therapist should show a patient-centered, flexible approach in delivering the self-help and show a good understanding of the delivery of the treatment methods themselves.	AZ_WRAD	AZ_Neg,AP_WRSL

Chapter 9

Toward Automatic Analysis of Psychotherapy Sessions

As stated in Chapter 1, this thesis aims to explore how to automatically analyse psychotherapy sessions by detecting and tracking positive signs of the therapeutic alliance. To achieve the aforementioned aims, the following research questions have been investigated:

1. To which degree is it possible to automatically *detect* any of the signs (behaviours) that may align with the positive therapeutic alliance, including any behaviours caused by the patient's emotional state, the therapist's competence, the therapist's empathy, the synchrony between the patient and the therapist and the patient's mood?

2. To which degree is it possible to automatically *track* any of the signs (patient's emotions, the synchrony between the patient and the therapist and the patient's mood) in a session?

This chapter presents the key findings gained from implementing the automatic modules presented in Chapter 5,6,7 and 8, that are related to the proposed full system in Section 3.2. Furthermore, this chapter connects the gained results from those chapters with the findings of the existing research literature.

9.1 Key Findings

Based on the research questions, a system model was proposed in Chapter 3 for the automatic analysis of audio recordings of psychotherapy sessions (the THEPS dataset, described in Section 3.3). Afterwards, each block in the proposed system has been investigated individually using a subset of the THEPS dataset that matched the label requirements of each experiment. Each experiment related to each block has been reported in a separate chapter:

Chapter 5, Chapter 6, Chapter 7 and Chapter 8. Those chapters' experiments yield several key findings related to the thesis research questions.

9.1.1 Chapter 5 (Automatic Speech Recognition for Conversational Psychotherapy Sessions) Findings

Chapter 5 investigated an Automatic Speech Recognition (ASR) system for the manually transcribed THEPS dataset. The results presented in that chapter showed the ability of the ASR system to automatically transcribe the psychotherapy session at an acceptable word error rate despite the common dataset challenges that occur in the medical field.

Due to the challenges concerning the nature of the recordings of therapy sessions and motivational interviewing sessions, the performance of ASR systems in those areas is still considerably lower compared to other ASR systems trained on controlled recording conditions with a large training data size. The findings reported in Chapter 5 supports the idea that the accuracy of the ASR systems implemented in the therapeutic fields is considered a reasonable result as described in the publications of other researchers, despite the use of-the-shelf ASR software in some studies. Furthermore, Chapter 5 presented the applicability of deploying an ASR system with the use of a small amount of data which is mostly the case in the medical domain, and it can act as a baseline for the other ASR systems that are based on the larger dataset. Xiao et al. (2015b) investigated an automatic system for rating the therapist's empathy in 200 Motivational Interviewing (MI) sessions for drug and alcohol counselling with human ratings of counsellor empathy. Each session was 20 minutes in total. They used ASR (Kaldi adaptation) to transcribe sessions and the resulting words were used in a text-based predictive model of empathy. For training the ASR, they used 1200 therapy transcripts to help define the typical vocabulary and language use. The ASR results gained a mean WER of 43.1%. The same ASR system was used in another study for analysing speech rate entrainment and investigating its relation to perceived empathy (Xiao et al., 2015a). A study by Chen et al. (2021) developed an automatic pipeline that transcribes 225 Cognitive Behaviour Therapy (CBT) sessions' recordings and extracts linguistic features for behavioural coding the sessions. They adopted an ASR system based on the Kaldi pipeline. The reported WER in the study was 44.01%. In a study by Miner et al. (2020) to assess the performance of ASR in psychotherapy discourse using 100 audio recordings of patient-therapist sessions, which are on average 45 minutes long, they used a commercial, cloud-based ASR service (Google Cloud Speech-to-Text) to transcribe the sessions automatically (Google, 2020). They reported that their speech recognition system's average word error rate was 25%.

9.1.2 Chapter 6 (Automatic Detection of Depression and Anxiety) Findings

Chapter 6 presents experiments aimed at investigating methods for automatically detecting depression and anxiety in the THEPS dataset using the provided mood score measures. The chapter's findings illustrated that the acoustic features extracted from the patient's speech could assist in predicting the patient's scores and score levels of depression and anxiety. The results gained presented the applicability of the automatic system to detect dementia-related diagnostics with the support of acoustic features. Furthermore, Chapter 6 showed that detecting depression and anxiety automatically in real-life psychotherapy session recordings using mood outcome measures is a promising area that mostly depends on those measures in manual diagnostics. The acoustic features investigated in the chapter's experiments included energy and voice-related features such as F0, loudness, jitter, shimmer, Harmonics-to-Noise Ratio (HNR), Mel Frequency Spectrum Coefficients (MFCCs) (as shown in Table 6.1). In the chapter's experiments, those features played an important role as the acoustic biomarkers for depression and anxiety. Various research studies have evidenced those findings by studying the influence of depression and anxiety symptoms on patients' acoustics. Some of these studies related the F0 impact to the depression and anxiety severity (Low et al., 2020; Mundt et al., 2007; Breznitz, 1992; Hönig et al., 2014; Stassen et al., 1993; Weeks et al., 2012; Hagenaars and Van Minnen, 2005; Goberman et al., 2011). Alternative voice-related features such as jitter, shimmer and HNR were found to be positively correlated with depression and anxiety (Low et al., 2020; Quatieri and Malyska, 2012; Vicsi et al., 2012; Ozdas et al., 2004; Fuller et al., 1992). Some researchers found high correlations of energy-based features such as loudness and MFCCs with depressed and anxious patients (Alpert et al., 2001; Stassen et al., 1993; Özseven et al., 2018; Taguchi et al., 2018; France et al., 2000; Cummins et al., 2013a; Ozdas et al., 2004; Cummins et al., 2011).

Furthermore, the confounding factors of cognitive decline and depressions was investigated in Chapter's 6 results. These findings make several contributions to the current literature supported in the following research studies (Sumali et al., 2020; Carvalho et al., 2020; Yang et al., 2017; Parlato-Oliveira et al., 2021; Wang et al., 2017).

9.1.3 Chapter 7 (Automatic Time-Continuous Recognition of Emotional Dimensions) Findings

Chapter 7 investigated the automatic prediction of the continuous labels of emotions using acoustic features extracted from patient's and therapist's speech. The chapter used a benchmark database for predicting and tracking the dimensional emotion labels because

the THEPS dataset is not labelled with emotional labels. The chapter's results indicated the feasibility of predicting and tracking the dimensional emotion labels (arousal and valence) in a session using a model trained on a benchmark database. Another interesting finding from Chapter 7 is that detecting and tracking the patient's and therapist's emotions in sessions using a benchmark dimensional emotion database can determine signs of the positive therapeutic alliance such as the emotion's synchrony between the therapist and the patient and the therapist's empathy. The behavioural synchrony between the therapist and the patient can be translated to emotional synchrony that eventually supports the emotional regulation for the patient (Koole and Tschacher, 2016). This emotional synchrony could be seen in the interaction between the therapist and the patient, especially in the speaker's vocal pitch, which is also known as vocal synchrony. Several studies demonstrated the vocal synchrony in conversational partners where higher levels of pitch synchrony were rated more positive emotions (Gregory, 1983, 1990; Gregory and Webster, 1996). Reich et al. (2014) examined the pitch synchrony between therapists and patients within psychotherapy sessions. The results showed that therapists and patients in these psychotherapy sessions revealed a satisfactory degree of synchrony in vocal pitch. A study by Bryan et al. (2018) explored the vocal pitch to match between patient's and therapist's mean level of emotional arousal during suicide risk assessment interviews. Results showed that patient and therapist vocal synchrony was positively correlated with emotional bond ratings as bearing on the results gained in Chapter 7.

Another concept that researchers found to be strongly correlated with the measures of the emotional bond is empathy (Horvath and Greenberg, 1989; Johnson et al., 2005). In the three-component model of the therapeutic alliance suggested by Bordin (1979), empathy is ideally most similar with the perception of the emotional bond that involves trust, mutual feelings of liking, attachment and feeling understood. In the environment of psychotherapy, the behavioural synchrony between the therapist and the patient were positively correlated with empathy, rapport and bonding (Charny, 1966; Chartrand and Bargh, 1999; Maurer and Tindall, 1983; Ramseyer and Tschacher, 2011). This behavioural synchrony could be seen in the emotional arousal reflecting empathy and emotional bonding. Several research studies that confirmed this fact was introduced in Chapter 4 (Imel et al., 2014; Xiao et al., 2014; Weiste and Peräkylä, 2014). The finding illustrates that vocal matching plays an essential role in shaping the patient's interpretation of the therapist's empathy. Those earlier studies complement the findings of Chapter 7 that indicate the positive relationship between emotion detection and tracking in session and therapist's empathy.

It is important to highlight that the tracking system implemented in Chapter 7 was based on predicting the dimensional emotions (arousal and valence) continuously for a fixed amount

of time in a single session. The THEPS dataset consisted of recordings of a single session per patient, limiting tracking to a single session rather than multiple sessions (a whole course of treatment).

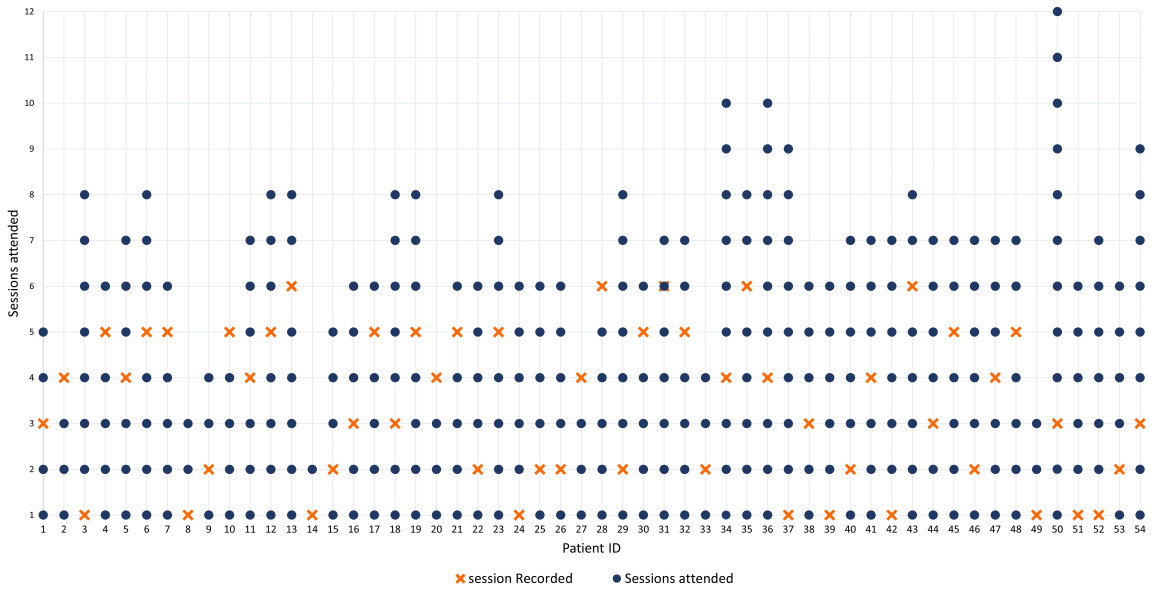


Fig. 9.1 The number of the recorded session and the attended sessions in the treatment sequence per patient in the THEPS dataset

Figure 9.1 presents the number of recorded session as well as all the sessions attended in the treatment sequence for each patient in the THEPS dataset. It is clear from the figure that the recorded session for each patient is randomly assigned within the first 6 sessions.

9.1.4 Chapter 8 (Automatic Prediction of Competency Measures) Findings

Chapter 8 presents the study of automatically predicting the competency measure using the THEPS dataset. The results showed that it is possible to automatically predict the therapist's competence using acoustic and linguistic features extracted from both the patient's and therapist's speech. In addition, the results implied the ability to detect the existence of ruptured moments in the therapeutic alliance using linguistic features. Furthermore, the findings demonstrated that selecting the RA as a language-based feature for predicting the therapist's competence is a reliable candidate for deploying the automatic system.

Surprisingly, it was found that the therapeutic alliance in Chapter 8 was related to the therapist's competence. This finding was also reported by Shaw and Dobson (1988); Whisman (1993); Trepka et al. (2004) where it was reported that the therapists are mostly

constrained by patient factors in their ability both to show competence and to assure positive therapeutic alliances. The use of both acoustic and linguistic features as shown in Chapter 8 to predict the therapist's competency measure achieved reliable prediction results on the THEPS dataset. Furthermore, the descriptive linguistic features presented in Figure 8.7 promoted rupture moments in the therapeutic alliance. Research demonstrated that the therapist's interpersonal skills could capture specific competencies, such as empathy, positive regard, dealing with criticism, or repairing of alliance ruptures (Constantino et al., 2013; Hatcher, 2015; Munder et al., 2019). Further evidence on the importance of a therapist's competence has been derived from research on rupture in the therapeutic alliance (Eubanks et al., 2018). Difficulties in those interpersonal competencies were found to be linked with adverse therapy outcomes such as patient's drop-out. The main contribution to the development of the therapists' interpersonal competencies is the effectiveness of the training received by the therapists. This training is usually received by supervision sessions adopted mindfulness training for purposes of assisting therapists to improve their observation of their own internal experience and the nature of their contributions to the interpersonal process (Muran et al., 2018). Furthermore, in a research by Daly et al. (2010), the authors found out that competent resolution of alliance rupture events mainly was dependent on the therapist's ability to recognise them and also on their adherence to the features of the rupture resolution models.

9.2 Proposed Final System Design

Based on the findings of the chapters illustrated in the previous section, the proposed model presented in Figure 3.1 can be enhanced to include the investigated features and the findings of each chapter as shown in Figure 9.2. Those findings from the thesis experiments reported in the modified proposed model illustrated in the figure broadly support the work of other studies reported earlier in the literature.

Due to the importance of integrating the several modules investigated in the aforementioned chapters, a proposed system design (layout) has been introduced in this section. The front-end of the full system introduced in this thesis can assist the therapists and the supervisors in exploring the dynamic behaviours of the therapist and the patient revealed in the session. In addition, the therapist could determine better treatment decisions based on the patient's expressed emotional and mental states. The supervisors are usually responsible for training and rating the therapist to ensure and develop the efficacy of the therapeutic alliance (Creaner, 2013). Adopting such an automated system layout could ensure a better experience for the supervisors with the therapists to maintain sufficient therapy standards. Figure 9.3

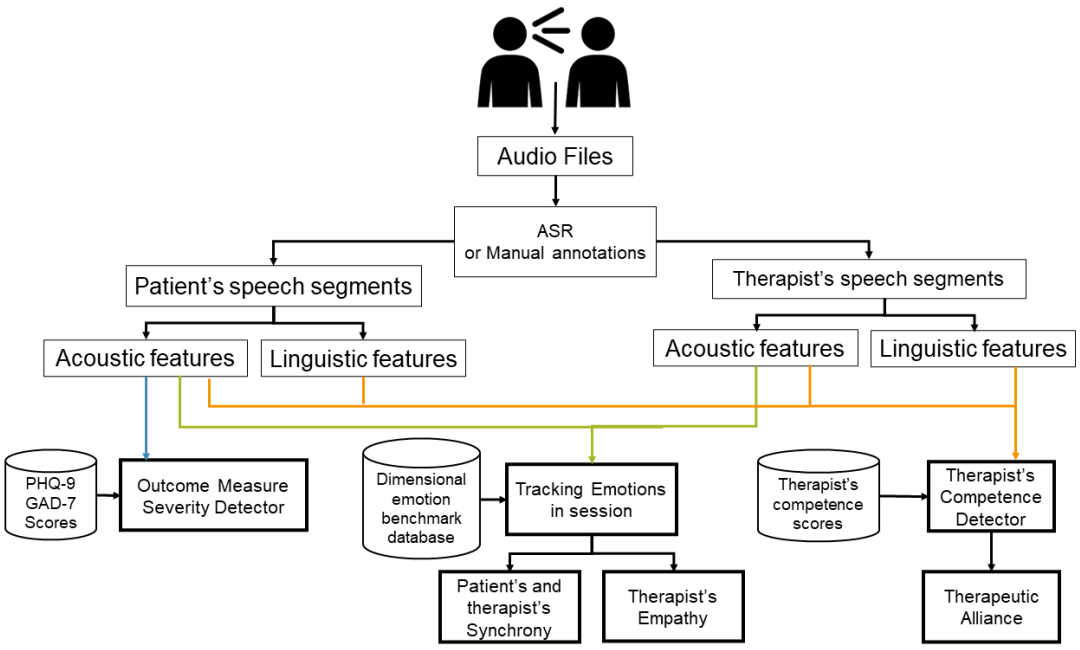


Fig. 9.2 The updated proposed model including the findings gained from Chapter 5,6,7 and 8

presents the proposed design of the full automatic system for patient S067 described in case study 3 in Section 7.5.1. Likewise, Figure 9.4 presents the proposed design of the full automatic system for patient S140 described in case study 4 in Section 7.5.1. The system design consists of general demographics information about the patient, an audio player section for playing the selected audio recording of the patient's session, a dynamic moving window presenting the arousal and valence variations through the session, the automatic transcription in a text box with a sliding bar, the predicted scores and the scores' levels of Generalised Anxiety Disorder(GAD-7), Patient Health Questionnaire (PHQ-9) and the therapist's competency ratings for the selected session's recording.

9.3 Summary

This chapter discussed the findings gained from building the blocks of the proposed system in Chapter 3 and an updated system were proposed based on the influence of those findings on the whole system. Furthermore, the findings were discussed from the perspective of other findings in the literature to support the aims and the research questions of this study. The main findings contributed in this discussion are: the ability to detect the patient's mood score

measures using acoustic features extracted from the patient's speech, the ability to detect the synchrony between the therapists and the patients based on the detection and tracking of their emotions in the sessions along with detecting the therapist's empathy and the ability to detect the therapeutic alliance from predicting the therapist's competence measures. Next chapter will conclude the thesis with a final summarisation and future work.

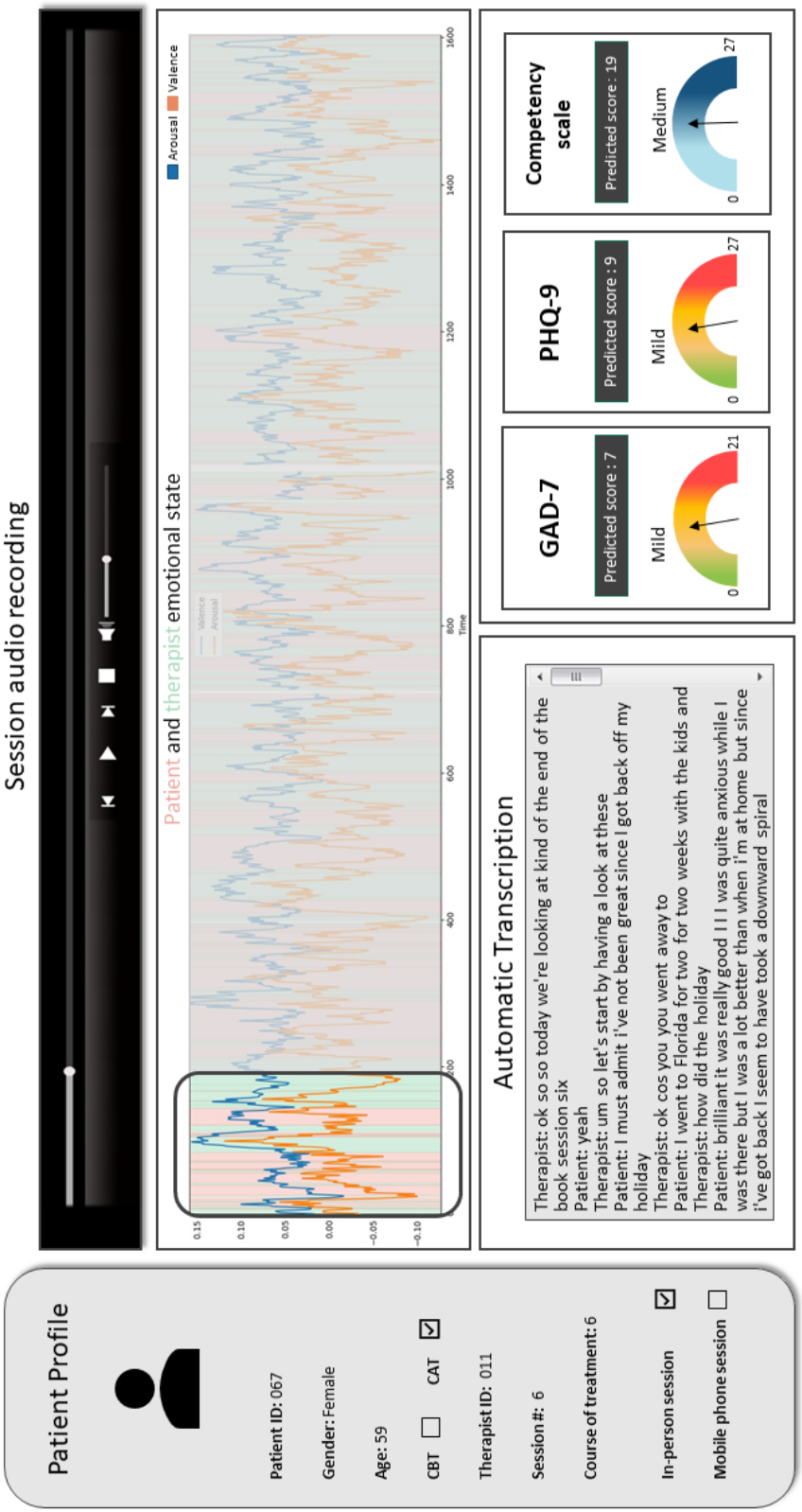


Fig. 9.3 The proposed system design for patient S067

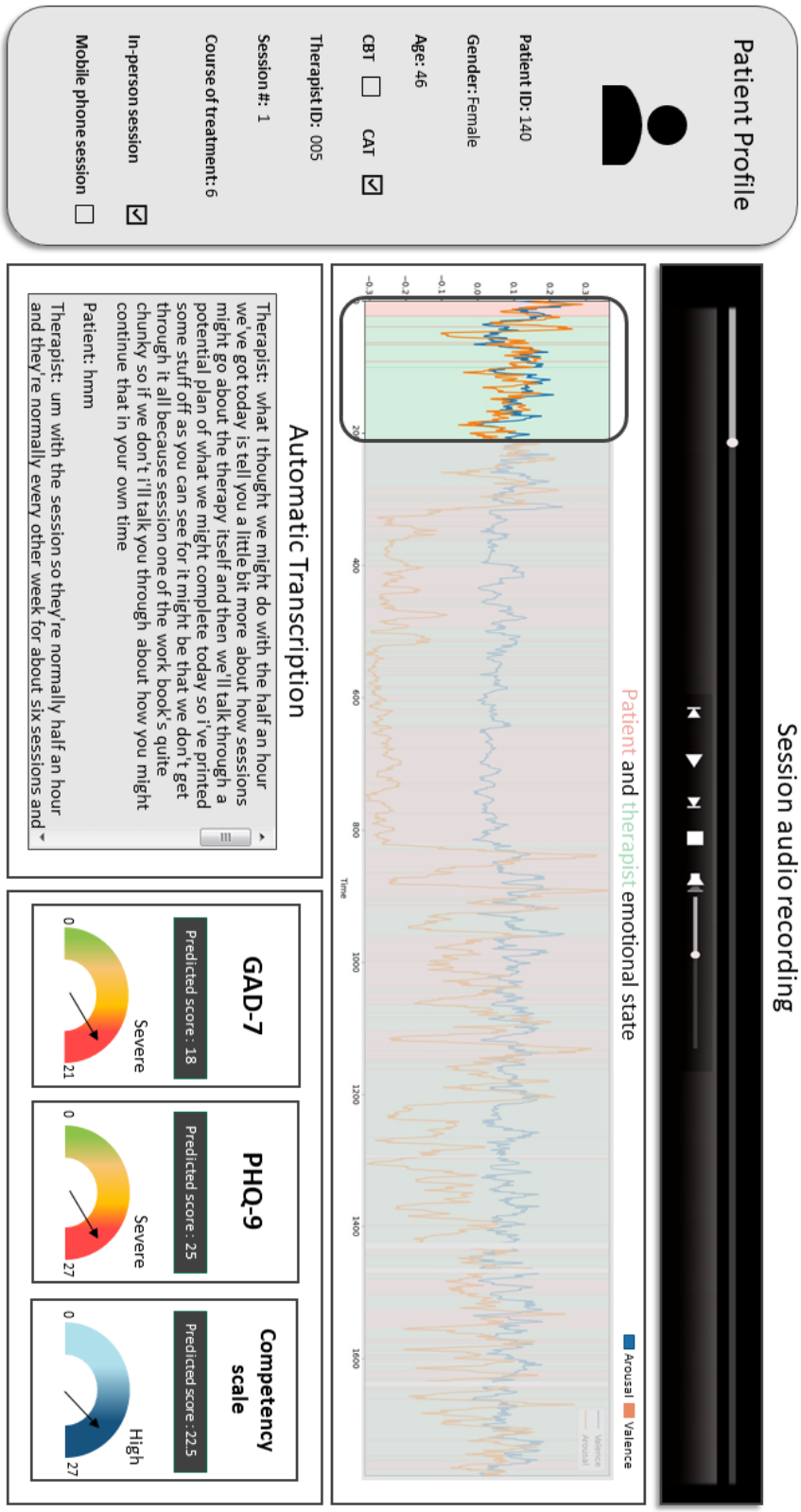


Fig. 9.4 The proposed system design for patient S140

Chapter 10

Conclusion and Future Work

In psychotherapy, therapists can review and analyse the recordings of the sessions manually to determine the relevant points for the therapy treatment. These points could be discovered by observing patients' and therapists' behaviours in the session. The main point that define the quality of the therapy is a positive therapeutic alliance. Maintaining a positive therapeutic alliance is also one of the criteria that therapists should aim to establish in a session with patients in order to be scored as competent mostly by the supervisors. One of the main disadvantages of not maintaining a positive therapeutic alliance is that the patients could drop out of therapy without finishing a complete sequence of sessions. Establishing and maintaining a positive therapeutic alliance requires hard effort from the therapist to keep an eye on several human behaviours related to the work of the therapy and the patient's behaviours as occurring during the therapy session. Furthermore, some behaviours would not be detectable only from the used language, but will also need an observation of the acoustic cues that could be accompanied with the verbal use of language. Therefore, there is a need for an automatic system that can assist therapists with maintaining a positive therapeutic alliance and eventually minimise the patient's drop-out rates.

This thesis investigated the automatic methods for analysing psychotherapy sessions by detecting and tracking several patients' and therapists' behaviours in the session found to correlate positively with the therapeutic alliance. This included establishing an Automatic Speech Recognition (ASR) system and investigating the acoustic and language-based features for the patient's and therapist's speech. Section 10.1 summarises the research reported in this thesis along with the main findings. Section 10.2 presents directions for future work. Finally, Section 10.3 concludes this chapter with remarks on the potential impact of the work in the thesis for the research and clinical communities.

10.1 Summary of thesis

The thesis began with a review of several clinical perspectives of psychotherapy relating to the positive therapeutic alliance. Patients' and therapists' behaviours were founded to contribute to establishing and maintaining a positive therapeutic alliance as highlighted in Chapter 2 including patient's emotions, the patient's mood, the therapist's empathy, the therapist's competence and the synchrony between the patient and the therapist. Afterwards, an automatic system for analysing psychotherapy sessions was proposed in Chapter 3 taking into consideration all of the human behaviours found earlier to contribute to the therapeutic alliance. In addition to this, a full review of the obtained psychotherapy sessions' recordings dataset was introduced in Chapter 3, the THEPS dataset.

As little is known about the automatic systems that investigated patients' and therapists' behaviours in psychotherapy or counselling sessions, a literature review was introduced in Chapter 4 to review the recent research studies in the therapy field that support the proposed automatic system. This review considered all the human behaviours intended to be detected or tracked in the main proposed system, either using acoustic or language-based features. Toward achieving a full automatic system, an ASR system was introduced in Chapter 5 to obtain the automatic transcriptions of the psychotherapy sessions. The results gained were in line with the ones reported in the related literature as reported in Chapter 9.

The patient's mood in the session could be discovered using the mood outcome measures reported by the patient earlier either in or before the session. Chapter 6 presented the work of automatically predicting and classifying the depression and anxiety scores and score levels as the mood outcome measures. Due to cross-sectional aspects, the experiments were evaluated on recordings of consultations related to Dementia to study the cause-and-effect relationships between Dementia and depression. A more related evaluation was established on the psychotherapy session recordings to study the feasibility of detecting depression and anxiety using the outcome measures. Those evaluations were based on the acoustic properties of the patients only in the session because the mood outcome measures are reported based on the patient's mood in the session. RFECV and SVR could be good candidates as baseline methods for future studies. Chapter 9 highlighted the results gained from the thesis along with the results gained from the literature as proof of validation.

Patient's and therapist's emotions could be detected and tracked from the dynamic emotional cues presented in the session. Chapter 7 presented work related to tracking the emotional state of the session's speakers using the dimensional models of emotion. Due to the unavailability of dimensional emotion labels in the psychotherapy sessions of the THEPS dataset, a benchmark database was used for training the system. The conducted experiments was based on the acoustic features extracted from patient's and therapist's speech. A further

analysis was conducted using the psychotherapy sessions dataset that revealed several human behaviours that occurred in the session, such as the therapist's empathy and the synchrony between the patient and the therapist. Using both the eGeMAPS and BoAWs features could be a feasible candidate as a feature set for predicting dimensional emotions for future studies. Previous research studies that aligned with the chapter findings were discussed in Chapter 9.

One of the rating scores that assess the therapists' work of therapy is the competency assessment measure, a scale used for rating therapists during treatment sessions. The work of predicting the scores and classifying the score levels of the competency measure was presented in Chapter 8. Due to the fact that the competency assessment requires observing both the patients' and the therapists' behaviour in the session, the features (both acoustic and language-based) were extracted from both the patient's and therapist's speech. Major findings were found that could relate to the positive and ruptured therapeutic alliance. Additionally, therapist's empathy was found to be one of the competency rating attributes that could be accurately detected comparable to the other competency rating attributes. RFECV and Lasso could be good candidates as baseline methods for future studies. Chapter 9 highlighted previous studies that obtained similar findings from a therapeutic perspective. Finally, a proposed system design (layout) for the fully automatic system investigated in this thesis was introduced in Chapter 9 that could inspire the clinical community.

10.2 Future work

The previous section summarised the main work established in this thesis in addition to the main findings. Still, some challenges and areas of improvement need to be addressed to leverage the overall performance of the proposed system and discover new aspects in psychotherapy. This section describes possible future directions as follows:

- State-of-art deep learning techniques are increasingly successful in various domains, requiring many data samples. It would be interesting to increase the number of session recordings in the dataset to be able to investigate more generalised models and enhance the results of the ASR system.

- Due to the existence of several variables in the psychotherapy sessions dataset, such as the in-person versus mobile phone session recordings, therapy type (Cognitive Analytical Therapy versus Cognitive Behavioural Therapy) and gender, further analysis would be useful to gain insight into the effects of those variables on the proposed system results.

- Obtaining more labels related to the emotions, empathy and the synchrony between patients therapists would improve the applicability to investigate more experiments related to those labels.

- Semi-supervised learning is one of the Machine Learning (ML) approaches that would be interesting to investigate in the work of psychotherapy due to the challenges of labelling several human behaviours in the work of therapy, especially that some of those behaviours require time-continuous labels, such as emotions. This technique would benefit from having a small labelled dataset to obtain labels for a larger unlabelled dataset.

- Feature embedding is an emerging research area focused on transforming features from an original space to a new one to support effective learning (Golinko and Zhu, 2019). This can be applied using the available models pretrained on out-of-domain large amount of data. They could enhance the work of recognising several human behaviours by extracting several high level features from the in-domain data. For example, the Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art deep learning pretrained model devoted for natural language processing tasks. Furthermore, as introduced in the Audio/Visual Emotion Challenge (AVEC) 2018 challenge baseline (Ringeval et al., 2018), DeepSpectrum is an unsupervised feature type that is directly learned from raw signals. Extracting those type of features with pretrained Image Convolutional Neural Networks (CNNs) would be interesting such that the features extracted creates visual representations for the audio data and then plots of spectrograms to be fed to a pretrained Image CNN.

- As mentioned in Section 3.1, it was difficult to automatically detect the rupture in the session due to the unavailability of the rupture markers as labels in the THEPS dataset. For that reason, it would be interesting to investigate the methods for automatic time-continues detection and recognition of rupture and rupture types in a single session and across multiple sessions for the same patient.

- As mentioned in Section 9.1, the main study's dataset lack the existence of multiple sessions' recordings (the whole course of treatment) for each patient. Having the full set of recordings for each patient could enable tracking patients' and therapists' behaviours across the sessions. This would benefit the tracking of the patient's progress in therapy and detect if the patient could drop-out of therapy.

- Conducting a follow-up study to explore psychotherapist's attitudes towards to the use of such a tool would be interesting in order to obtain insights into the project's

suitability in the field of study. This could include engaging various stakeholder groups, such as supervisors, therapists and even patients.

10.3 Concluding remarks

In conclusion, this research could have great implications for the research and clinical communities. A full system capturing deep human behaviours, such as emotions and competency, could leverage the research community to investigate more related and dynamic human behaviours, even at small intervals within the session. Furthermore, the clinical community could benefit from such a system when training and supervising therapists. Clinical supervisors (as noticed by attending several supervisory meetings conducted with Dr Stephen Kellett) train therapists or practitioners by reviewing their work using their sessions' recordings. An automatic system would also aid their personal and professional development. Having such a tool could assist supervisors in their training work by highlighting areas of limitations and improvements for the therapists in the treatment sessions. The system could also assist therapists in revealing several patient's behaviours toward regaining a positive therapeutic alliance, minimising drop-out cases and eventually enhancing patient's treatment outcomes. Additionally, therapists would find it helpful to be aware of their own behaviours in the session to regulate and enhance their emotions and competence.

References

Abhang, P., Gawali, B., and Mehrotra, S. (2016). *Introduction to EEG and speech based emotion recognition*. Academic Press.

Ackerman, S. and Hilsenroth, M. (2003). A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical Psychology Review*, 23(1):1–33.

Akçay, M. and Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

Alpert, M., Pouget, E., and Silva, R. (2001). Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69.

Amir, N., Mixdorff, H., Amir, O., Rochman, D., Diamond, G., Pfitzinger, H., Levi-Isserlish, T., and Abramson, S. (2010). Unresolved anger: Prosodic analysis and classification of speech from a therapeutic setting. In *Speech Prosody 2010-Fifth International Conference*.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., and Schuller, B. (2017). Snore sound classification using image-based deep spectrum features.

Atta, N., Christopher, C., Mariani, R., Belotti, L., Andreoli, G., and Danskin, K. (2019). Linguistic features of the therapeutic alliance in the first session: a psychotherapy process study. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 22(1).

Attas, D., Kellett, S., Blackmore, C., and Christensen, H. (2022a). Automated detection of competency in psychotherapy sessions via speech and language processing. *Applied Sciences, Applications of Speech and Language Technologies in Healthcare (in preparation)*.

Attas, D., Kellett, S., Blackmore, C., and Christensen, H. (2022b). Automatic time-continuous prediction of emotional dimensions during guided self help for anxiety disorders. *FRIAS Junior Researcher Conference: Human Perspectives on Spoken Human-Machine Interaction (SpoHuMa21)*.

Attas, D., Mirheidari, B., Blackburn, D., Venneri, A., Walker, T., Harkness, K., Reuber, M., Blackmore, C., and Christensen, H. (2021). Predicting levels of depression and anxiety in people with neurodegenerative memory complaints presenting with confounding symptoms. In *Intelligent Computing*, pages 58–69. Springer.

Aziz-Zadeh, L., Sheng, T., and Gheytanchi, A. (2010). Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability. *PloS one*, 5(1):e8759.

Bachelor, A. (1995). Clients' perception of the therapeutic alliance: A qualitative analysis. *Journal of Counseling Psychology*, 42(3):323.

Bada, I., Karsten, J., Fohr, D., and Illina, I. (2017). Data selection in the framework of automatic speech recognition. In *ICNLSSP 2017-International conference on natural language, signal and speech processing 2017*, pages 1–5.

Baird, A., Cummins, N., Schnieder, S., and Schuller, B. (2020). An evaluation of the effect of anxiety on speech–computational prediction of anxiety from sustained vowels. *Power*, 160(130):111.

Barber, J., Sharpless, B., Klostermann, S., and McCarthy, K. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology: Research and Practice*, 38(5):493.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Barrett, M., Chua, W., Crits, P., Gibbons, M., and Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy: Theory, Research, Practice, Training*, 45(2):247.

Batliner, A., Steidl, S., and Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus.

Beck, J. (2011). *Cognitive behavior therapy [electronic resource] : basics and beyond*. Guilford, New York ; London, 2nd ed. edition.

Beckham, E. (1992). Predicting patient dropout in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, 29(2):177.

Bell, L., Gustafson, J., and Heldner, M. (2003). Prosodic adaptation in human-computer interaction. In *Proceedings of ICPHS*, volume 3, pages 833–836. Citeseer.

Biswas, A., Yılmaz, E., de Wet, F., van der Westhuizen, E., and Niesler, T. (2019). Semi-supervised acoustic model training for five-lingual code-switched ASR. *arXiv preprint arXiv:1906.08647*.

Black, M., Katsamanis, A., Baucom, B., Lee, C., Lammert, A., Christensen, A., Georgiou, P., and Narayanan, S. (2013). Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech communication*, 55(1):1–21.

Boateng, G., Sels, L., Kuppens, P., Hilpert, P., and Kowatsch, T. (2020). Speech emotion recognition among couples using the peak-end rule and transfer learning. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 17–21.

Bone, D., Lee, C., Potamianos, A., and Narayanan, S. (2014). An investigation of vocal arousal dynamics in child-psychologist interactions using synchrony measures and a conversation-based model. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Booth, B., Mundnich, K., and Narayanan, S. (2018). Fusing annotations with majority vote triplet embeddings. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 83–89.

Bordin, E. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252.

Bower, P. and Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency: narrative literature review. *The British Journal of Psychiatry*, 186(1):11–17.

Breznitz, Z. (1992). Verbal indicators of depression. *The Journal of general psychology*, 119(4):351–363.

British Psychological Society (2001). Treatment choice in psychological therapies and counselling: Evidence based clinical practice guideline. *Department of Health*.

Bryan, C., Baucom, B., Crenshaw, A., Imel, Z., Atkins, D., Clemans, T., Leeson, B., Burch, S., Mintz, J., and Rudd, D. (2018). Associations of patient-rated emotional bond and vocally encoded emotional arousal among clinicians and acutely suicidal military personnel. *Journal of consulting and clinical psychology*, 86(4):372.

Bucci, W. (1997). *Psychoanalysis and cognitive science: A multiple code theory*. Guilford Press.

Bucci, W. and Maskit, B. (2005). Building a weighted dictionary for referential activity. *Computing attitude and affect in text*, pages 49–60.

Bucci, W. and Maskit, B. (2007). Beneath the surface of the therapeutic interaction: The psychoanalytic method in modern dress. *Journal of the American Psychoanalytic Association*, 55(4):1355–1397.

Bucci, W., Maskit, B., and Hoffman, L. (2012). Objective measures of subjective experience: The use of therapist notes in process-outcome research. *Psychodynamic psychiatry*, 40(2):303–340.

Cairney, J., Corna, L., Veldhuizen, S., Herrmann, N., and Streiner, D. (2008). Comorbid depression and anxiety in later life: patterns of association, subjective well-being, and impairment. *The American journal of geriatric psychiatry*, 16(3):201–208.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., and Snyder, P. (2004). Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56(1):30–35.

Carvalho, A., Firth, J., and Vieta, E. (2020). Bipolar disorder. *New England Journal of Medicine*, 383(1):58–66.

Charny, J. (1966). Psychosomatic manifestations of rapport in psychotherapy. *Psychosomatic medicine*, 28(4):305–315.

Chartrand, T. and Bargh, J. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

Chekroud, A., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., et al. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, 20(2):154–170.

Chen, S. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Chen, Z., Flemotomos, N., Ardulov, V., Creed, T., Imel, Z., Atkins, D., and Narayanan, S. (2021). Feature fusion strategies for end-to-end evaluation of cognitive behavior therapy sessions. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1836–1839. IEEE.

Christian, C., Barzilai, E., Nyman, J., and Negri, A. (2021). Assessing key linguistic dimensions of ruptures in the therapeutic alliance. *Journal of Psycholinguistic Research*, 50(1):143–153.

Churiwala, S. (2019). *An introduction to machine learning*. Springer, Cham, Switzerland.

Clark, D. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *International review of psychiatry*, 23(4):318–327.

Cohn, J., Cummins, N., Epps, J., Goecke, R., Joshi, J., and Scherer, S. (2018). Multimodal assessment of depression from behavioral signals. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pages 375–417.

Constantino, M., Boswell, J., Bernecker, S., and Castonguay, L. (2013). Context-responsive psychotherapy integration as a framework for a unified clinical science: Conceptual and empirical considerations. *Journal of Unified Psychotherapy and Clinical Science Volume*, 2(1):1–20.

Corrales-Astorgano, M., Martínez-Castilla, P., Escudero-Mancebo, D., Aguilar, L., González-Ferreras, C., and Cardeñoso-Payo, V. (2019). Automatic assessment of prosodic quality in down syndrome: Analysis of the impact of speaker heterogeneity. *Applied sciences*, 9(7):1440.

Cournoyer, L., Brochu, S., Landry, M., and Bergeron, J. (2007). Therapeutic alliance, patient behaviour and dropout in a drug rehabilitation programme: the moderating effect of clinical subpopulations. *Addiction*, 102(12):1960–1970.

Coutinho, J., Ribeiro, E., Sousa, I., and Safran, J. (2014). Comparing two methods of identifying alliance rupture events. *Psychotherapy*, 51(3):434.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

Crangle, C., Wang, R., Perreau, M., Nguyen, M., Nguyen, D., and Suppes, P. (2019). Machine learning for the recognition of emotion in the speech of couples in psychotherapy using the stanford suppes brain lab psychotherapy dataset. *arXiv preprint arXiv:1901.04110*.

Creaner, M. (2013). *Getting the Best Out of Supervision in Counselling & Psychotherapy: A Guide for the Supervisee*. SAGE.

Cully, J. and Teten, A. (2008). *A Therapist's Guide to Brief Cognitive Behavioral Therapy*. South Central Mental Illness Research, Education and Clinical Center (MIRECC).

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., and Schuller, B. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 478–484.

Cummins, N., Amiriparian, S., Ottl, S., Gerczuk, M., Schmitt, M., and Schuller, B. (2018). Multimodal bag-of-words for cross domains sentiment analysis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4954–4958. IEEE.

Cummins, N., Epps, J., and Ambikairajah, E. (2013a). Spectro-temporal analysis of speech affected by depression and psychomotor retardation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Cummins, N., Epps, J., Breakspear, M., and Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*.

Cummins, N., Epps, J., Sethu, V., Breakspear, M., and Goecke, R. (2013b). Modeling spectral variability for the classification of depressed speech. In *INTERSPEECH*, pages 857–861.

Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., and Epps, J. (2013c). Diagnosis of depression by behavioural signals: a multimodal approach. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Dahl, H. (1978). A new psychoanalytic model of motivation: Emotions as appetites and messages. *Psychoanalysis and Contemporary Thought*, 1(3):373–408.

Daly, A., Llewelyn, S., McDougall, E., and Chanen, A. (2010). Rupture resolution in cognitive analytic therapy for adolescents with borderline personality disorder. *Psychology and Psychotherapy: Theory, Research and Practice*, 83(3):273–288.

Darby, J. and Hollien, H. (1977). Vocal and speech patterns of depressive patients. *Folia Phoniatrica et Logopaedica*, 29(4):279–291.

De Boer, J., Voppel, A., Brederoo, S., Schnack, H., Truong, K., Wijnen, F., and Sommer, I. (2021). Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. *Psychological medicine*, pages 1–11.

De Wet, F., Louw, P., and Niesler, T. (2006). The design, collection and annotation of speech databases in south africa. *Proc. of the Pattern Recognition Association of South Africa (PRASA 2006)*, pages 1–5.

Devanand, D., Sano, M., Tang, M., Taylor, S., Gurland, B., Wilder, D., Stern, Y., and Mayeux, R. (1996). Depressed mood and the incidence of alzheimer's disease in the elderly living in the community. *Archives of general psychiatry*, 53(2):175–182.

Diniz, B., Butters, M., Albert, S., Dew, M., and Reynolds, C. (2013). Late-life depression and risk of vascular dementia and Alzheimer's disease: systematic review and meta-analysis of community-based cohort studies. *The British Journal of Psychiatry*, 202(5):329–335.

Doorn, K., Kamsteeg, C., Bate, J., and Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, 31(1):92–116.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J., Devillers, L., Abrilian, S., Batliner, A., et al. (2007). The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*, pages 488–500. Springer.

Drucker, H., Burges, C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*.

Dwyer, D., Falkai, P., and Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology*, 14:91–118.

Ekman, P., Friesen, W., and Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings*, volume 11. Elsevier.

Ekman, P. and Oster, H. (1979). Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554.

El Ayadi, M., Kamel, M., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587.

Ellgring, H. and Scherer, K. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20(2):83–110.

Elsey, C., Drew, P., Jones, D., Blackburn, D., Wakefield, S., Harkness, K., Venneri, A., and Reuber, M. (2015). Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient education and counseling*, 98(9):1071–1077.

Eubanks, C., Lubitz, J., Muran, C., and Safran, J. (2019). Rupture Resolution Rating System (3RS): Development and validation. *Psychotherapy Research*, 29(3):306–319.

Eubanks, C., Muran, C., and Safran, J. (2018). Alliance rupture repair: A meta-analysis. *Psychotherapy*, 55(4):508.

Eubanks, C. F., Muran, J. C., and Safran, J. D. (2015). Rupture Resolution Rating System (3RS): Manual. *Unpublished manuscript, Mount Sinai-Beth Israel Medical Center, New York*.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., et al. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Fairburn, C. and Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour research and therapy*, 49(6-7):373–378.

Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C. (2008). Liblinear: A library for large linear classification. *the Journal of machine Learning research*, 9:1871–1874.

Fiedler, F. (1950). The concept of an ideal therapeutic relationship. *Journal of Consulting Psychology*, 14(4):239–245.

Firth, N., Barkham, M., Kellett, S., and Saxon, D. (2015). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. *Behaviour Research and Therapy*, 69:54–62.

Flemotomos, N., Martinez, V., Gibson, J., Atkins, D., Creed, T., and Narayanan, S. (2018). Language features for automated evaluation of cognitive behavior psychotherapy sessions. In *INTERSPEECH*, pages 1908–1912.

Flint, A., Black, S., Campbell-Taylor, I., Gailey, G., and Levinton, C. (1993). Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of psychiatric research*, 27(3):309–319.

France, D., Shiavi, R., Silverman, S., Silverman, M., and Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837.

Freud, S. (1958). On beginning the treatment (Further recommendations on the technique of psycho-analysis I). In *The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XII (1911-1913): The Case of Schreber, Papers on Technique and Other Works*, pages 121–144.

Fuller, B., Horii, Y., and Conner, D. (1992). Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety. *Research in nursing & health*, 15(5):379–389.

Gakuto, K., Kartik, A., and Brian, K. (2019). IBM Research advances in end-to-end speech recognition at INTERSPEECH 2019. https://www.ibm.com/blogs/research/2019/10/end-to-end-speech-recognition/.

Gale, R., Chen, L., Dolata, J., Van Santen, J., and Asgari, M. (2019). Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques. In *INTERSPEECH*, volume 2019, page 11. NIH Public Access.

Ge, W. and Yu, Y. (2017). Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1086–1095.

Georgescu, A., Cucu, H., and Burileanu, C. (2019). Kaldi-based DNN architectures for speech recognition in romanian. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE.

Ghahremani, P., Manohar, V., Hadian, H., Povey, D., and Khudanpur, S. (2017). Investigation of transfer learning for ASR using LF-MMI trained neural networks. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 279–286. IEEE.

Gideon, J., Schatten, H., McInnis, M., and Provost, E. (2019). Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In *INTERSPEECH*.

Goberman, A., Hughes, S., and Haydock, T. (2011). Acoustic characteristics of public speaking: Anxiety and practice effects. *Speech communication*, 53(6):867–876.

Golinko, E. and Zhu, X. (2019). Generalized feature embedding for supervised, unsupervised, and online learning tasks. *Information Systems Frontiers*, 21(1):125–142.

Google (2020). Cloud Speech-to-Text .

Greenberg, L., Watson, J., Elliot, R., and Bohart, A. (2001). Empathy. *Psychotherapy: Theory, research, practice, training*, 38(4):380.

Gregory, S. (1983). A quantitative analysis of temporal symmetry in microsocial relations. *American Sociological Review*, pages 129–135.

Gregory, S. (1990). Analysis of fundamental frequency reveals covariation in interview partners' speech. *Journal of Nonverbal Behavior*, 14(4):237–251.

Gregory, S. and Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of personality and social psychology*, 70(6):1231.

Grimm, M. and Kroschel, K. (2005). Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 381–385. IEEE.

Grimm, M., Kroschel, K., and Narayanan, S. (2007). Support vector regression for automatic recognition of spontaneous emotions in speech. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4. IEEE.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, pages 865–868. IEEE.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3).

Hagenaars, M. and Van Minnen, A. (2005). The effect of fear on paralinguistic aspects of speech in patients with panic disorder with agoraphobia. *Journal of Anxiety Disorders*, 19(5):521–537.

Hall, J., Harrigan, J., and Rosenthal, R. (1995). Nonverbal behavior in clinician—patient interaction. *Applied and Preventive Psychology*, 4(1):21–37.

Han, J., Zhang, Z., Cummins, N., Ringeval, F., and Schuller, B. (2017). Strength modelling for real-world automatic continuous affect recognition from audiovisual signals. *Image and Vision Computing*, 65:76–86.

Hatcher, R. (2015). Interpersonal competencies: Responsiveness, technique, and training in psychotherapy. *American Psychologist*, 70(8):747.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., and Dean, J. (2013). Multilingual acoustic models using distributed deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8619–8623. IEEE.

Helfer, B., Quatieri, T., Williamson, J., Mehta, D., Horwitz, R., and Yu, B. (2013). Classification of depression state based on articulatory precision. In *INTERSPEECH*, pages 2172–2176.

Hepple, J. (2004). Psychotherapies with older people: an overview. *Advances in psychiatric treatment*, 10(5):371–377.

Herkov, M. (2018). What is psychotherapy? https://psychcentral.com/lib/what-is-psychotherapy/.

Hirschfeld, R. (1994). Major depression, dysthymia and depressive personality disorder. *British Journal of Psychiatry*, 165(S26):23–30.

Hoffman, L., Algus, J., Braun, W., Bucci, W., and Maskit, B. (2013). Treatment notes: Objective measures of language style point to clinical insights. *Journal of the American Psychoanalytic Association*, 61(3):535–568.

Hofmann, S., Gerlach, A., Wender, A., and Roth, W. (1997). Speech disturbances and gaze behavior during public speaking in subtypes of social phobia. *Journal of Anxiety Disorders*, 11(6):573–585.

Hölzer, M., Pokorny, D., Kächele, H., and Luborsky, L. (1997). The verbalization of emotions in the therapeutic dialogue–a correlate of therapeutic outcome? *Psychotherapy Research*, 7(3):261–273.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., and Krajewski, J. (2014). Automatic modelling of depressed speech: relevant features and relevance of gender.

Horvath, A., Del Re, A., Flückiger, C., and Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, 48(1):9.

Horvath, A. and Greenberg, L. (1989). Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223.

Horvath, A. and Greenberg, L. (1994). *The working alliance: Theory, research, and practice*, volume 173. John Wiley & Sons.

Horwitz, R., Quatieri, T., Helfer, B., Yu, B., Williamson, J., and Mundt, J. (2013). On the relative importance of vocal source, system, and prosody in human depression. In *2013 IEEE International Conference on Body Sensor Networks*, pages 1–6. IEEE.

Huahu, X., Jue, G., and Jian, Y. (2010). Application of speech emotion recognition in intelligent household robot. In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, volume 1, pages 537–541. IEEE.

Huang, J., Kuchaiev, O., O'Neill, P., Lavrukhin, V., Li, J., Flores, A., Kucsko, G., and Ginsburg, B. (2020). Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. *arXiv preprint arXiv:2005.04290*.

Imel, Z., Barco, J., Brown, H., Baucom, B., Baer, J., Kircher, J., and Atkins, D. (2014). The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of counseling psychology*, 61(1):146.

Jacobson, S. (2015). What is cognitive analytic therapy? *Harley Therapy Counselling Blog*.

Johnson, J., Burlingame, G., Olsen, J., Davies, R., and Gleave, R. (2005). Group climate, cohesion, alliance, and empathy in group psychotherapy: Multilevel structural equation models. *Journal of counseling psychology*, 52(3):310.

Jordan, M. and Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Jurafsky, D. and Martin, J. (2013). *Speech and language processing: Pearson new international edition*. Pearson.

Kang, B., Jeon, H., and Park, J. (2020). Speech recognition for task domains with sparse matched training data. *Applied Sciences*, 10(18):6155.

Kaya, H., Fedotov, D., Dresvyanskiy, D., Doyran, M., Mamontov, D., Markitantov, M., Akdag Salah, A., Kavcar, E., Karpov, A., and Salah, A. (2019). Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 27–35.

Kellett, S., Bee, C., Aadahl, V., Headley, E., and Delgadillo, J. (2020). A pragmatic patient preference trial of cognitive behavioural versus cognitive analytic guided self-help for anxiety disorders. *Behavioural and Cognitive Psychotherapy*, page 1–8.

Kellett, S., Simmonds-Buckley, M., Limon, E., Hague, J., Hughes, L., Stride, C., and Millings, A. (2021a). Defining the assessment and treatment competencies to deliver low-intensity cognitive behavior therapy: A multi-center validation study. *Behavior Therapy*, 52(1):15 – 27.

Kellett, S., Simmonds-Buckley, M., Limon, E., Hague, J., Hughes, L., Stride, C., and Millings, A. (2021b). Low intensity cognitive behavioural competency scale manual, unpublished document.

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M., and Cleder, C. (2019). Automatic speech emotion recognition using machine learning. In *Social media and machine learning*. IntechOpen.

Kiss, G. and Vicsi, K. (2017). Mono-and multi-lingual depression prediction based on speech processing. *International Journal of Speech Technology*, 20(4):919–935.

Kohrt, B., Jordans, M., Rai, S., Shrestha, P., Luitel, N., Ramaiya, M., Singla, D., and Patel, V. (2015). Therapist competence in global mental health: development of the enhancing assessment of common therapeutic factors (ENACT) rating scale. *Behaviour research and therapy*, 69:11–21.

Koolagudi, S. and Rao, S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117.

Koole, S. and Tschacher, W. (2016). Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance. *Frontiers in psychology*, 7:862.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B., et al. (2019). SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*.

Kraepelin, E. (1921). Manic depressive insanity and paranoia. *The Journal of Nervous and Mental Disease*, 53(4):350.

Kroenke, K., Spitzer, R., and Williams, J. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kullmann, E. (2016). Speech to text for swedish using kaldi.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., and Stober, S. (2017). Transfer learning for speech recognition on a budget. *arXiv preprint arXiv:1706.00290*.

Lambert, M. and Bergin, A. (1994). The effectiveness of psychotherapy. *Handbook of psychotherapy and behavior change*, 4:143–189.

Lamel, L., Gauvain, J., and Adda, G. (2002). Unsupervised acoustic model training. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–877. IEEE.

Lee, C., Busso, C., Lee, S., and Narayanan, S. (2009). Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *Tenth Annual Conference of the International Speech Communication Association*.

Lee, F., Hull, D., Levine, J., Ray, B., and McKeown, K. (2019). Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 12–23.

Lee, S., Suh, S., Kim, T., Kim, K., Lee, K., Lee, J., Han, G., Hong, J., Han, J., Lee, K., et al. (2021). Screening major depressive disorder using vocal acoustic features in the elderly by sex. *Journal of Affective Disorders*, 291:15–23.

Leigh, B. and Milgrom, J. (2008). Risk factors for antenatal depression, postnatal depression and parenting stress. *BMC psychiatry*, 8(1):1–11.

Li, A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., Song, Z., Ruhi, U., Venkataramani, V., and Chen, X. (2000). CASS: A phonetically transcribed corpus of mandarin spontaneous speech. In *Sixth International Conference on Spoken Language Processing*.

Li, J. (2021). Recent advances in end-to-end automatic speech recognition. *arXiv preprint arXiv:2111.01690*.

Li, X. and Kivlighan, D. (2020). Examining therapy dynamics and session outcome using differential equations model and multilevel data disaggregation. *Psychotherapy Research*, 30(5):604–621.

Liang, Y., Zheng, X., and Zeng, D. (2019). A survey on big data-driven digital phenotyping of mental health. *Information Fusion*, 52:290–307.

Lin, J., Wu, C., and Wei, W. (2011). Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE Transactions on Multimedia*, 14(1):142–156.

Lo, T. and Chen, B. (2019). Semi-supervised training of acoustic models leveraging knowledge transferred from out-of-domain data. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1400–1404. IEEE.

Lord, S., Sheng, E., Imel, Z., Baer, J., and Atkins, D. (2015). More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

Low, D., Bentley, K., and Ghosh, S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116.

Ma, J., Matsoukas, S., Kimball, O., and Schwartz, R. (2006). Unsupervised training on large amounts of broadcast news data. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 3, pages 1056–1059. IEEE.

Malandrakis, N., Potamianos, A., Evangelopoulos, G., and Zlatintsi, A. (2011). A supervised approach to movie emotion tracking. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2376–2379. IEEE.

Manea, L., Gilbody, S., and McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the patient health questionnaire (PHQ-9): a meta-analysis. *Cmaj*, 184(3):E191–E196.

Mariooryad, S. and Busso, C. (2013). Exploring cross-modality affective reactions for audiovisual emotion recognition. *IEEE Transactions on affective computing*, 4(2):183–196.

Martin, D., Garske, J., and Davis, K. (2000). Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology*, 68(3):438.

Martinez, V., Flemotomos, N., Ardulov, V., Somandepalli, K., Goldberg, S., Imel, Z., Atkins, D., and Narayanan, S. (2019). Identifying therapist and client personae for therapeutic alliance estimation. In *INTERSPEECH*, pages 1901–1905.

Maskit, B. (2012). The Discourse Attributes Analysis Program (DAAP) (Series 8) [Computer software]. http://www.thereferentialprocess.org/dictionary-measures-and-computer-programs.

Maskit, B. (2021). Overview of computer measures of the referential process. *Journal of Psycholinguistic Research*, 50(1):29–49.

Maskit, B., Shanahan, J., Qu, Y., and Wiebe, J. (2005). A weighted dictionary for referential activity. *Computing Attitude and Affect in Text*, pages 49–60.

Matassoni, M., Gretter, R., Falavigna, D., and Giuliani, D. (2018). Non-native children speech recognition through transfer learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6229–6233. IEEE.

Maurer, R. and Tindall, J. (1983). Effect of postural congruence on client's perception of counselor empathy. *Journal of counseling psychology*, 30(2):158.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

Meadows, J. and Kellett, S. (2017). Development and evaluation of cognitive analytic guided self-help (CAT-SH) for use in IAPT services. *Behavioural and cognitive psychotherapy*, 45(3):266–284.

Mencattini, A., Mosciano, F., Comes, M., Di Gregorio, T., Raguso, G., Daprati, E., Ringeval, F., Schuller, B., Di Natale, C., and Martinelli, E. (2018). An emotional modulation model as signature for the identification of children developmental disorders. *Scientific reports*, 8(1):1–12.

Mergenthaler, E. and Bucci, W. (1999). Linking verbal and non-verbal representations: Computer analysis of referential activity. *British Journal of Medical Psychology*, 72(3):339–354.

Metallinou, A., Katsamanis, A., and Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152.

Meyer, J. (2019). Multi-task and transfer learning in low-resource speech recognition.

Michie, S., Atkins, L., West, R., et al. (2014). The behaviour change wheel. *A guide to designing interventions. 1st ed. Great Britain: Silverback Publishing*, pages 1003–1010.

Milton, J. (2004). *Making sense of psychotherapy and psychoanalysis*. Mind Publications.

Milton, J., Polmear, C., and Fabricius, J. (2011). *A short introduction to psychoanalysis*. SAGE.

Miner, A., Haque, A., Fries, J., Fleming, S., Wilfley, D., Wilson, T., Milstein, A., Jurafsky, D., Arnow, B., Agras, S., et al. (2020). Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ digital medicine*, 3(1):1–8.

Mirheidari, B., Blackburn, D., O'Malley, R., Venneri, A., Walker, T., Reuber, M., and Christensen, H. (2020). Improving cognitive impairment classification by generative neural network-based feature augmentation. In *INTERSPEECH*, pages 2527–2531.

Mirheidari, B., Blackburn, D., O'Malley, R., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2019). Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Mirheidari, B., Blackburn, D., Reuber, M., Walker, T., and Christensen, H. (2016). Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of INTERSPEECH*, pages 1220–1224. ISCA.

Mirheidari, B., Pan, Y., Blackburn, D., O'Malley, R., and Christensen, H. (2021). Identifying cognitive impairment using sentence representation vectors. *Proc. INTERSPEECH 2021*, pages 2941–2945.

Moscati, A., Flint, J., and Kendler, K. (2016). Classification of anxiety disorders comorbid with major depression: common or distinct influences on risk? *Depression and anxiety*, 33(2):120–127.

Muliyala, K. and Varghese, M. (2010). The complex relationship between depression and dementia. *Annals of Indian Academy of Neurology*, 13(Suppl2):S69.

Munder, T., Schlipfenbacher, C., Toussaint, K., Warmuth, M., Anderson, T., and Gumz, A. (2019). Facilitative interpersonal skills performance test: Psychometric analysis of a german language version. *Journal of Clinical Psychology*, 75(12):2273–2283.

Mundt, J., Snyder, P., Cannizzaro, M., Chappie, K., and Geralts, D. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of neurolinguistics*, 20(1):50–64.

Muran, C., Safran, J., Eubanks, C., and Gorman, B. (2018). The effect of alliance-focused training on a cognitive-behavioral therapy for personality disorders. *Journal of consulting and clinical psychology*, 86(4):384.

Murphy, S., Maskit, B., and Bucci, W. (2015). Putting feelings into words: Cross-linguistic markers of the referential process. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 80–88.

Mustafa, M., Salim, S., Mohamed, N., Al-Qatab, B., and Siong, C. (2014). Severity-based adaptation with limited data for ASR to aid dysarthric speakers. *PloS one*, 9(1):e86285.

Narayanan, S. and Georgiou, P. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, 101(5):1203–1233.

Nasir, M., Baucom, B., Georgiou, P., and Narayanan, S. (2017). Predicting couple therapy outcomes based on speech acoustic features. *PloS one*, 12(9):e0185123.

Newson, J. (2018). The challenges of mental health diagnosis. https://sapienlabs.org/the-challenges-of-mental-health-diagnosis/.

NHS (2018a). Medical psychotherapy. https://www.healthcareers.nhs.uk/explore-roles/doctors/roles-doctors/psychiatry/medical-psychotherapy.

NHS (2018b). Psychiatry. https://www.nhs.uk/conditions/psychiatry/.

NHS (2018c). Self-help therapies. https://www.nhs.uk/conditions/stress-anxiety-depression/self-help-therapies/.

NHS (2021). The improving access to psychological therapies manual. https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/. The National Collaborating Centre for Mental Health.

NHS (2021). Psychological therapies, Annual report on the use of IAPT services-England. https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services/annual-report-2020-21.

Nicolaou, M., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105.

Nilsonne, A. (1988). Speech characteristics as indicators of depressive illness. *Acta Psychiatrica Scandinavica*, 77(3):253–263.

Novotney, S. and Schwartz, R. (2009). Analysis of low-resource acoustic model self-training. In *Tenth Annual Conference of the International Speech Communication Association*.

Nwe, T., Foo, S., and De Silva, L. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623.

O'Donohue, W. and Fisher, J. (2012). *Cognitive behavior therapy: Core principles for practice*. John Wiley & Sons.

Ozdas, A., Shiavi, R., Silverman, S., Silverman, M., and Wilkes, M. (2004). Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE transactions on Biomedical engineering*, 51(9):1530–1540.

Özkanca, Y., Demiroğlu, C., Besirli, A., and Celik, S. (2018). Multi-lingual depression-level assessment from conversational speech using acoustic and text features. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. International Speech Communication Association.

Özseven, T., Düğenci, M., Doruk, A., and Kahraman, H. (2018). Voice traces of anxiety: acoustic parameters affected by anxiety disorder. *Archives of Acoustics*, pages 625–636.

Pan, S. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Pandit, V., Cummins, N., Schmitt, M., Hantke, S., Graf, F., Paletta, L., and Schuller, B. (2018). Tracking authentic and in-the-wild emotions using speech. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE.

Parlato-Oliveira, E., Saint-Georges, C., Cohen, D., Pellerin, H., Pereira, I., Fouillet, C., Chetouani, M., Dommergues, M., and Viaux-Savelon, S. (2021). "Motherese" prosody in fetal-directed speech: An exploratory study using automatic social signal processing. *Frontiers in Psychology*, 12:649.

Partila, P., Voznak, M., and Tovarek, J. (2015). Pattern recognition methods and features selection for speech emotion recognition system. *The Scientific World Journal*, 2015.

Pecht, M. and Kang, M. (2018). *Prognostics and health management of electronics : fundamentals, machine learning, and internet of things*. John Wiley Sons, Hoboken, New Jersey, second edition. edition.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12.

Pennebaker, J., Boyd, R., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC 2015. Technical report.

Perfect, D., Jackson, C., Pybis, J., and Hill, A. (2016). Choice of therapies in IAPT: An overview of the availability and client profile of step 3 therapies. *British Association of Counselling & Psychotherapy*.

Perlis, R. (2005). Misdiagnosis of bipolar disorder. *The American journal of managed care*, 11(9 Suppl):S271–4.

Piselli, A., Halgin, R., and MacEwan, G. (2011). What went wrong? Therapists' reflections on their role in premature termination. *Psychotherapy Research*, 21(4):400–415.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Power, N. (2021). *Investigating the Relationship Between Therapist Competence and Patient Outcome in Adult Psychological Interventions*. PhD thesis, University of Sheffield.

Quatieri, T. and Malyska, N. (2012). Vocal-source biomarkers for depression: A link to psychomotor activity. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Ramseyer, F. and Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 79(3):284.

Rao, S., Koolagudi, S., and Vempada, R. (2013). Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, 16(2):143–160.

Reich, C., Berman, J., Dale, R., and Levitt, H. (2014). Vocal synchrony in psychotherapy. *Journal of Social and Clinical Psychology*, 33(5):481–494.

Renals, S. and Swietojanski, P. (2017). Distant speech recognition experiments using the AMI Corpus. *New Era for Robust Speech Recognition*, pages 355–368.

Ringeval, F., Marchi, E., Grossard, C., Xavier, J., Chetouani, M., Cohen, D., and Schuller, B. (2016). Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In *Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association (ISCA)*, pages 1210–1214.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., et al. (2018). AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 3–13. ACM.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., and Pantic, M. (2015a). AVEC 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1335–1336.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., and Pantic, M. (2017). AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9.

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., and Pantic, M. (2015b). AV+ EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th international workshop on audio/visual emotion challenge*, pages 3–8.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

Rohde, P., Lewinsohn, P., and Seeley, J. (1991). Comorbidity of unipolar depression: II. Comorbidity with other mental disorders in adolescents and adults. *Journal of abnormal psychology*, 100(2):214.

Rollnick, S. and Miller, W. (1995). What is motivational interviewing? *Behavioural and cognitive Psychotherapy*, 23(4):325–334.

Russell, J. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Safran, J. (1990a). Towards a refinement of cognitive therapy in light of interpersonal theory: I. Theory. *Clinical Psychology Review*, 10(1):87–105.

Safran, J. (1990b). Towards a refinement of cognitive therapy in light of interpersonal theory: II. Practice. *Clinical Psychology Review*, 10(1):107–121.

Safran, J. (1993). Breaches in the therapeutic alliance: An arena for negotiating authentic relatedness. *Psychotherapy: Theory, Research, Practice, Training*, 30(1):11.

Safran, J. (2009). Psychoanalytic therapy over time.

Safran, J. and Muran, C. (2006). Resolving therapeutic impasses.

Sahraeian, R. and Van Compernolle, D. (2016). Using weighted model averaging in distributed multilingual DNNs to improve low resource ASR. *Procedia Computer Science*, 81:152–158.

Salary, S. and Moghadam, M. (2013). Relationship between depression and cognitive disorders in women affected with dementia disorder. *Procedia-Social and Behavioral Sciences*, 84:1763–1769.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. ii—recent progress. *IBM Journal of research and development*, 11(6):601–617.

Scheflen, A. (1963). Communication and regulation in psychotherapy. *Psychiatry*, 26(2):126–136.

Scherer, S., Stratou, G., Gratch, J., and Morency, L. (2013). Investigating voice quality as a speaker-independent indicator of depression and PTSD. In *INTERSPEECH*, pages 847–851.

Schmidtke, K., Pohlmann, S., and Metternich, B. (2008). The syndrome of functional memory disorder: definition, etiology, and natural course. *The American Journal of Geriatric Psychiatry*, 16(12):981–988.

Schmitt, M., Ringeval, F., and Schuller, B. (2016). At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech. In *INTERSPEECH*, pages 495–499.

Schmitt, M. and Schuller, B. (2017). OpenXBOW: introducing the passau open-source crossmodal Bag-of-Words Toolkit.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011a). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication*, 53(9-10):1062–1087.

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., and Konosu, H. (2009). Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774.

Schuller, B., Rigoll, G., and Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE international conference on acoustics, speech, and signal processing*, volume 1, pages I–577. IEEE.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. (2011b). AVEC 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer.

Schuller, B., Valster, M., Eyben, F., Cowie, R., and Pantic, M. (2012). AVEC 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456.

Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S., et al. (2018). Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *Proc. INTERSPEECH*, pages 2808–2812.

Sexton, H., Hembre, K., and Kvarme, G. (1996). The interaction of the alliance and therapy microprocess: A sequential analysis. *Journal of Consulting and Clinical Psychology*, 64(3):471.

Shafran, R., Myles-Hooton, P., Bennett, S., and Öst, L. (2021). The concept and definition of 'low intensity' cognitive behaviour therapy. *Behaviour Research and Therapy*, page 103803.

Shaharin, R., Prodhan, U., and Rahman, M. (2014). Performance study of TDNN training algorithm for speech recognition. *International Journal of Advanced Research in Computer Science & Technology*, 2(4):90–95.

Sharf, J., Primavera, L., and Diener, M. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, 47(4):637.

Shaw, B. and Dobson, K. (1988). Competency judgments in the training and evaluation of psychotherapists. *Journal of Consulting and Clinical Psychology*, 56(5):666.

Sheehan, D., Lecrubier, Y., Sheehan, H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., and Dunbar, G. (1998). The mini-international neuropsychiatric interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of clinical psychiatry*.

Silber-Varod, V., Kreiner, H., Lovett, R., Levi-Belz, Y., and Amir, N. (2016). Do social anxiety individuals hesitate more? the prosodic profile of hesitation disfluencies in social anxiety disorder individuals. In *Proceedings of Speech Prosody*, volume 2016, pages 1211–1215.

Singla, K., Chen, Z., Flemotomos, N., Gibson, J., Can, D., Atkins, D., and Narayanan, S. (2018). Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. In *INTERSPEECH*, pages 3413–3417.

Spencer-Oatey, H. (2005). politeness, face and perceptions of rapport: unpackaging their bases and interrelationships. *Politeness Res. Lang. Behav. Cult.*

Spitzer, R., Kroenke, K., Williams, J., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, 166(10):1092–1097.

Stassen, H. et al. (1993). Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of psychiatric research*, 27(3):289–307.

Stegmann, G., Hahn, S., Liss, J., Shefner, J., Rutkove, S., Kawabata, K., Bhandari, S., Shelton, K., Duncan, C., and Berisha, V. (2020). Repeatability of commonly used speech and language features for clinical applications. *Digital biomarkers*, 4(3):109–122.

Sumali, B., Mitsukura, Y., Liang, K., Yoshimura, M., Kitazawa, M., Takamiya, A., Fujita, T., Mimura, M., and Kishimoto, T. (2020). Speech quality feature analysis for classification of depression and dementia patients. *Sensors*, 20(12):3599.

Sümer, Ö., Beyan, C., Ruth, F., Kramer, O., Trautwein, U., and Kasneci, E. (2021). Estimating presentation competence using multimodal nonverbal behavioral cues. *arXiv preprint arXiv:2105.02636*.

Taguchi, T., Tachikawa, H., Nemoto, K., Suzuki, M., Nagano, T., Tachibana, R., Nishimura, M., and Arai, T. (2018). Major depressive disorder discrimination using vocal acoustic features. *Journal of affective disorders*, 225:214–220.

Tasnim, M. and Stroulia, E. (2019). Detecting depression from voice. In *Canadian Conference on Artificial Intelligence*, pages 472–478. Springer.

Tavabi, L., Stefanov, K., Zhang, L., Borsari, B., Woolley, J., Scherer, S., and Soleymani, M. (2020). Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 406–413.

Thomas, S., Seltzer, M., Church, K., and Hermansky, H. (2013). Deep neural network features and semi-supervised training for low resource speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6704–6708. IEEE.

Tickle-Degnen, L. and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293.

Tocatly, K., Bucci, W., and Maskit, B. (2019). Developing a Preliminary Measure of the Arousal Function of the Referential Process [Poster presentation]. Research Day Colloquium at the City College of New York's Clinical Psychology Doctoral Program .

Tracy, J., Özkanca, Y., Atkins, D., and Ghomi, R. (2020). Investigating voice as a biomarker: Deep phenotyping methods for early detection of parkinson's disease. *Journal of biomedical informatics*, 104:103362.

Trepka, C., Rees, A., Shapiro, D., Hardy, G., and Barkham, M. (2004). Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy and Research*, 28(2):143–157.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M., Schuller, B., and Zafeiriou, S. (2016). 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)*, pages 5200–5204. IEEE.

Tseng, S., Chakravarthula, S., Baucom, B., and Georgiou, P. (2016). Couples behavior modeling and annotation using low-resource lstm language models. In *INTERSPEECH*, pages 898–902.

Tversky, A. (1969). Intransitivity of preferences. *Psychological review*, 76(1):31.

Upadhyaya, P., Farooq, O., Abidi, M., and Varshney, Y. (2017). Continuous hindi speech recognition model based on kaldi ASR toolkit. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 786–789. IEEE.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10. ACM.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). AVEC 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM.

Vicsi, K., Sztahó, D., and Kiss, G. (2012). Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 511–515. IEEE.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339.

Waltz, J., Addis, M., Koerner, K., and Jacobson, N. (1993). Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *Journal of consulting and clinical psychology*, 61(4):620.

Wang, D., Shangguan, Y., Yang, H., Chuang, P., Zhou, J., Li, M., Venkatesh, G., Kalinli, O., and Chandra, V. (2021). Noisy training improves E2E ASR for the edge. *arXiv preprint arXiv:2107.04677*.

Wang, J., Sui, X., Zhu, T., and Flint, J. (2017). Identifying comorbidities from depressed people via voice analysis. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 986–991. IEEE.

Wang, Y. (2020). Automatic Speech Recognition Model for Swedish using Kaldi.

Watkins Jr, C. E. (2012). Educating psychotherapy supervisors. *American Journal of Psychotherapy*, 66(3):279–307.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063.

Weck, F., Richtberg, S., Jakob, M., Neng, J. M., and Höfling, V. (2015). Therapist competence and therapeutic alliance are important in the treatment of health anxiety (hypochondriasis). *Psychiatry Research*, 228(1):53–58.

Weeks, J., Lee, C., Reilly, A., Howell, A., France, C., Kowalsky, J., and Bush, A. (2012). "The Sound of Fear": Assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder. *Journal of anxiety disorders*, 26(8):811–822.

Weiste, E. and Peräkylä, A. (2014). Prosody and empathic communication in psychotherapy interaction. *Psychotherapy Research*, 24(6):687–701.

Whisman, M. (1993). Mediators and moderators of change in cognitive therapy of depression. *Psychological bulletin*, 114(2):248.

Whiteside, S. (1998). Simulated emotions: an acoustic study of voice and perturbation measures. In *Fifth International Conference on Spoken Language Processing*.

Wickramasinghe, N. and Geisler, E. (2008). *Encyclopedia of healthcare information systems*. IGI Global.

Wiegersma, S., Nijdam, M., van Hessen, A., Truong, K., Veldkamp, B., and Olff, M. (2020). Recognizing hotspots in brief eclectic psychotherapy for PTSD by text and audio mining. *European Journal of psychotraumatology*, 11(1):1726672.

Wierzbicki, M. and Pekarik, G. (1993). A meta-analysis of psychotherapy dropout. *Professional Psychology: Research and Practice*, 24(2):190.

Wolberg, L. (1988). *The technique of psychotherapy*. Grune & Stratton, Inc/Harcourt, Bra.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., and Cowie, R. (2008). Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. 9th INTERSPEECH 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, pages 597–600.

Wollmer, M., Schuller, B., Eyben, F., and Rigoll, G. (2010). Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing*, 4(5):867–881.

World Health Organization (2017). Depression and other common mental disorders: global health estimates.

Wu, D., Parsons, T., Mower, E., and Narayanan, S. (2010). Speech emotion estimation in 3D space. In *2010 IEEE International Conference on Multimedia and Expo*, pages 737–742. IEEE.

Xia, W., Gibson, J., Xiao, B., Baucom, B., and Georgiou, P. (2015). A dynamic model for behavioral analysis of couple interactions using acoustic features. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Xiao, B., Bone, D., Segbroeck, M., Imel, Z., Atkins, D., Georgiou, P., and Narayanan, S. (2014). Modeling therapist empathy through prosody in drug addiction counseling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Xiao, B., Georgiou, P., Imel, Z., Atkins, D., and Narayanan, S. (2013). Modeling therapist empathy and vocal entrainment in drug addiction counseling. In *INTERSPEECH*, pages 2861–2865.

Xiao, B., Huang, C., Imel, Z., Atkins, D., Georgiou, P., and Narayanan, S. (2016). A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science*, 2:e59.

Xiao, B., Imel, Z., Atkins, D., Georgiou, P., and Narayanan, S. (2015a). Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Xiao, B., Imel, Z., Georgiou, P., Atkins, D., and Narayanan, S. (2015b). " Rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS one*, 10(12):e0143055.

Yacoub, S., Simske, S., Lin, X., and Burns, J. (2003). Recognition of emotions in interactive voice response systems. In *INTERSPEECH*.

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629.

Yang, T., Wu, C., Huang, K., and Su, M. (2017). Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):895–906.

Yi, J., Tao, J., Wen, Z., and Bai, Y. (2018). Language-adversarial transfer learning for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):621–630.

Yoon, B. (2009). Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415.

Yoon, W., Cho, Y., and Park, K. (2007). A study of speech emotion recognition and its application to mobile services. In *International Conference on Ubiquitous Intelligence and Computing*, pages 758–766. Springer.

Yoshioka, T., Erdogan, H., Chen, Z., Xiao, X., and Alleva, F. (2018). Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks. *arXiv preprint arXiv:1810.03655*.

Young, V. and Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112. PMID: 20698428.

Yu, L., Lee, L., Hao, S., Wang, J., He, Y., Hu, J., Lai, K., and Zhang, X. (2016). Building chinese affective resources in Valence-Arousal dimensions.

Zavaliagkos, G., Siu, M., Colthurst, T., and Billa, J. (1998). Using untranscribed training data to improve performance. In *Fifth International Conference on Spoken Language Processing*.

Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2008). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

Zhang, Z., Lin, W., Liu, M., and Mahmoud, M. (2020). Multimodal deep learning framework for mental disorder recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 344–350. IEEE.

Zheng, A. and Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. O'Reilly Media, Inc.

Appendix A

A Sample of PHQ-9

Nine-symptom Checklist

Name _____ Date _____

Over the *last 2 weeks*, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

(For office coding: Total Score _____ = _____ + _____ + _____)

If you checked off *any* problems, how *difficult* have these problems made it for you to do your work, take care of things at home, or get along with other people?

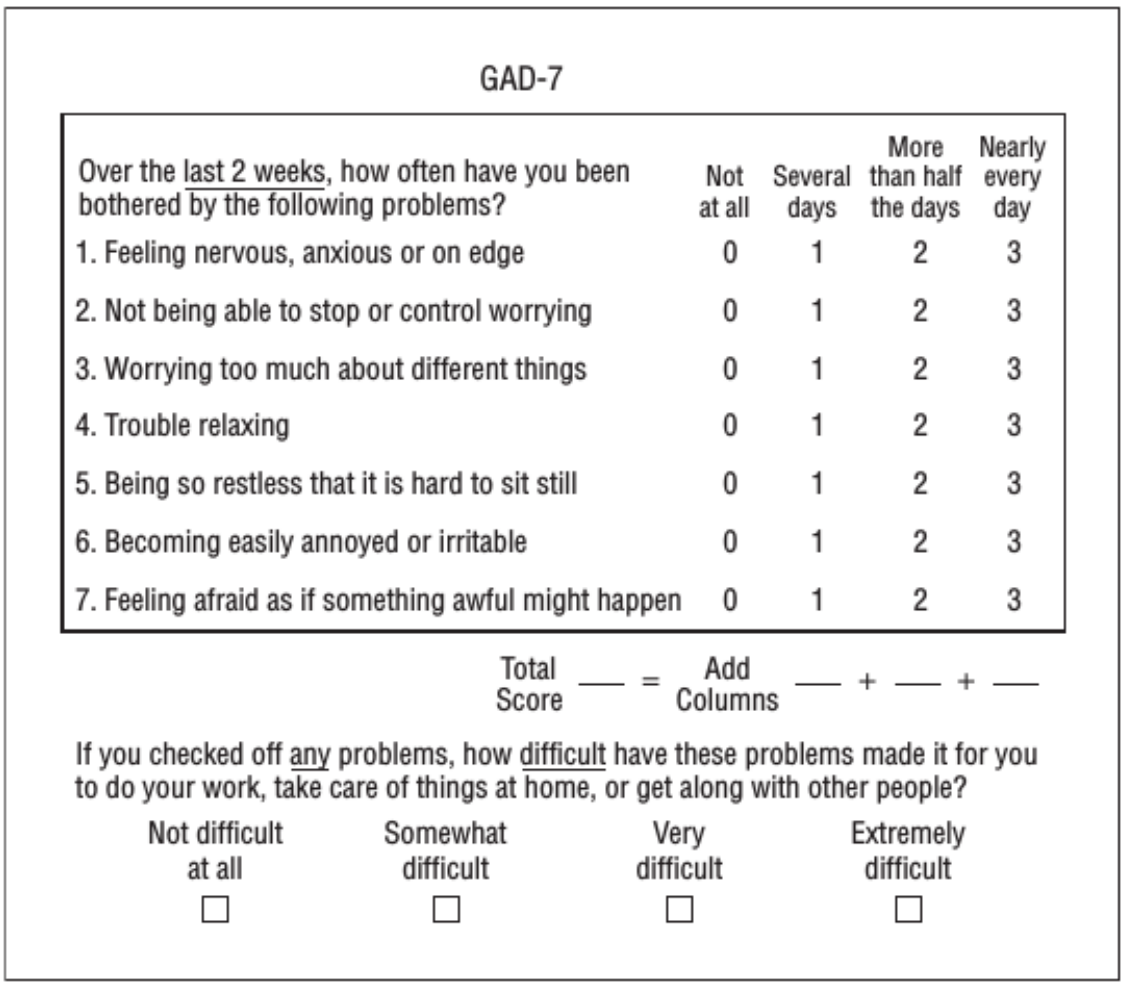
Not difficult at all	Somewhat difficult	Very difficult	Extremely difficult
☐	☐	☐	☐

From the Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PRIME-MD PHQ). The PHQ was developed by Drs. Robert L. Spitzer, Janet BW Williams, Kurt Kroenke, and colleagues. For research information, contact Dr. Spitzer at *rls8@columbia.edu*. PRIME-MD is a trademark of Pfizer Inc.

Fig. A.1 PHQ-9 Questionnaire (Kroenke et al., 2001)

Appendix B

A Sample of GAD-7



GAD-7

Over the last 2 weeks, how often have you been bothered by the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Feeling nervous, anxious or on edge	0	1	2	3
2. Not being able to stop or control worrying	0	1	2	3
3. Worrying too much about different things	0	1	2	3
4. Trouble relaxing	0	1	2	3
5. Being so restless that it is hard to sit still	0	1	2	3
6. Becoming easily annoyed or irritable	0	1	2	3
7. Feeling afraid as if something awful might happen	0	1	2	3

Total Score ___ = Add Columns ___ + ___ + ___

If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all	Somewhat difficult	Very difficult	Extremely difficult
☐	☐	☐	☐

Fig. B.1 GAD-7 Questionnaire (Spitzer et al., 2006)

Appendix C

A Sample of the LI-CBT Treatment Competency Scale

CBT-GSH PATIENT PREFERENCE TRIAL TREATMENT SESSION **COMPETENCY RATING:_____/36**

Rater name: Date: Tape identifier:

Focusing the CBT-GSH session						
	INCOMPETENT	**NOVICE**	**ADVANCED BEGINNER**	**COMPETENT**	**PROFICIENT**	**EXPERT**
Agrees collaborative agenda with patient						
Subsequent adherence to that agenda						
Overall Section Competency Rating	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Continued Engagement in CBT-GSH						
	INCOMPETENT	**NOVICE**	**ADVANCED BEGINNER**	**COMPETENT**	**PROFICIENT**	**EXPERT**
Collaborative approach concerning change or difficulties						
Acknowledges progress or difficulties by use of simple reflections						
Acknowledges progress or difficulties by use of complex reflections						
Use of capsule summaries regarding progress or difficulties						
Use of major section summaries						
Ratio of questions to feedback to facilitate change						
Overall Section Competency Rating	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Interpersonal Competencies in the session						
	INCOMPETENT	**NOVICE**	**ADVANCED BEGINNER**	**COMPETENT**	**PROFICIENT**	**EXPERT**
Empathises through verbal communication						
Non-verbal behaviour ***(please note do not score if the session is audio and do not let this alter the main rating in this section)***						
Encouraging and reinforcing						
Warmth and compassion						
Pacing of the session						
Overall Section Competency Rating	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Information Gathering Competencies Specific to Change						
COM-B		**INCOMPETENT**	**NOVICE**	**ADVANCED BEGINNER**	**COMPETENT**	**PROFICIENT**	**EXPERT**
CAPABILITY OPPORTUNITY MOTIVATION	**Questioning skills**						
	Problem statement review ***This only gets completed at session 3 of CBT-GSH and is not then present in subsequent sessions, so do not let this alter the main rating in this section.***						
	Review of goal progress ***(please note may be absent and do not let this alter the main rating in this section)***						
	Homework review						
	Risk review ***(please note can be a very quick check in for a 3 score)***						
	Review of medication ***(please note can be a very quick check in for a 3 score)***						
	Outcome monitoring						
Overall Section Competency Rating	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Within Session Self-Help Change Method Competencies						
COM-B		**INCOMPETENT**	**NOVICE**	**ADVANCED BEGINNER**	**COMPETENT**	**PROFICIENT**	**EXPERT**
CAPABILITY OPPORTUNITY MOTIVATION	**Rationale for treatment (e.g. introduced or reiterated) ↓**						
	Adherence to principles of PWP intervention Displays fidelity to low intensity treatment (eg. using psychoeducational materials in session) ↓						
	Appropriateness of PWP intervention Relevant PWP intervention appropriate for patient, stage of intervention and presenting problem ↓						
	Change method (e.g. use of diaries, ABC or 5-areas conceptualisation to drive the low intensity change methods such as BA, CT ect)						
Overall Section Competency Rating	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Planning and Shared Decision Making Competencies						
COM-B		**INCOMPETENT**	**NOVICE**	**ADVANCED BEGINNER**	**COMPETENT**	**PROFICIENT**	**EXPERT**
C O M	**Agrees next steps of treatment and the between session work ↓**						
	Defines and agrees the implementation plan for the between session work ↓						
	Session review and ending						
Overall Section Competency Rating	**0**	**1**	**2**	**3**	**4**	**5**	**6**