Article:
Attas, D., Power, N., Smithies, J. et al. (2022) Automated detection of the competency of delivering guided self-help for anxiety via speech and language processing. Applied Sciences, 12 (17). 8608. ISSN: 2076-3417

https://doi.org/10.3390/app12178608

applied sciences

MDPI

Article

Automated Detection of the Competency of Delivering Guided Self-Help for Anxiety via Speech and Language Processing

Dalia Attas [1,*], Niall Power [2], Jessica Smithies [3], Charlotte Bee [3], Vikki Aadahl [3], Stephen Kellett [4], Chris Blackmore [5] and Heidi Christensen [1,*]

1 Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
2 Derbyshire Community Health Services NHS Foundation Trust, Chesterfield DE45 1AD, UK
3 Improving Access to Psychological Therapies, Pennine Care NHS Foundation Trust, Ashton-under-Lyne OL6 7SR, UK
4 Rotherham Doncaster and South Humber NHS Foundation Trust and Clinical and Applied Psychology Unit, Department of Psychology, University of Sheffield, Sheffield DN4 8QN, UK
5 School of Health and Related Research (ScHARR), University of Sheffield, Sheffield S1 4DA, UK
* Correspondence: deattas@uqu.edu.sa (D.A.); heidi.christensen@sheffield.ac.uk (H.C.)

Abstract: Speech and language play an essential role in automatically assessing several psychotherapeutic qualities. These automation procedures require translating the manual rating qualities to speech and language features that accurately capture the assessed psychotherapeutic quality. Speech features can be determined by analysing recordings of psychotherapeutic conversations (acoustics), while language-based analyses rely on the transcriptions of such psychotherapeutic conversations (linguistics). Guided self-help is a psychotherapeutic intervention that mainly relay on therapeutic competency of practitioners. This paper investigates the feasibility of automatically analysing guided self-help sessions for mild-to-moderate anxiety to detect and predict practitioner competence. This analysis is performed on sessions drawn from a patient preference randomised controlled trial using actual patient-practitioner conversations manually rated using a valid and reliable measure of competency. The results show the efficacy and potential of automatically detecting practitioners' competence using a system based on acoustic and linguistic features extracted from transcripts generated by an automatic speech recogniser. Feature extraction, feature selection and classification or regression have been implemented as blocks of the prediction model. The Lasso regression model 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.

Keywords: competency; guided self-help sessions; automatic speech recognition; machine learning; speech processing; language processing

Citation: Attas, D.; Power, N.; Smithies, J.; Bee, C.; Aadahl, V.; Kellett, S.; Blackmore, C.; Christensen, H. Automated Detection of the Competency of Delivering Guided Self-Help for Anxiety via Speech and Language Processing. *Appl. Sci.* **2022**, *12*, 8608. https://doi.org/10.3390/app12178608

Academic Editor: Jorge Martin-Gutierrez

Received: 30 June 2022
Accepted: 26 August 2022
Published: 28 August 2022

1. Introduction

The automatic detection of any human behaviour in recorded speech requires analysis of the speech and language for important features that can accurately detect the chosen behaviours. The speakers' expressed behaviours might be revealed in their speech (acoustics) or in their language use. Representing these by acoustic and language-based features could allow for the modelling and detection of several human behaviours automatically. The acoustic features represent the physical characteristics of the speech signal, such as the loudness, frequency and amplitude of the signal [1]. 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 [2].

Treatment competence during any psychological intervention refers to the levels of the clinician's knowledge of the disorder and the intervention being delivered, the skill with which the intervention is delivered, consideration of how appropriate the intervention is for the problems being addressed [3] and that the model-specific change methods are delivered to the levels required for the patient to make the expected clinical progress [4].

The delivery of competent treatment depends on the understanding and interpretation of communicative behaviours observed by the therapists or practitioners in the sessions. In making these decisions, the therapist or the practitioner could search for different signs in the conversation, such as acoustic and linguistic cues. In the case of psychological interventions, usually therapists or practitioners record sessions with patients using an audio or video recorder. These recordings are useful for clinical supervision and training purposes. However, this procedure is time-consuming and costly to do manually and it is generally difficult for the practitioner to accurately observe some acoustic signs in the moment of a session, such as tracking a patient's pitch or speaking rate, due to the multiple demands of the situation. Detecting these types of cues using automatic speech processing could potentially be more efficient and accurate and so contribute to better clinical supervision and effective and safe delivery of interventions. Such a system would leverage recent progress and maturing of signal processing and machine learning algorithms that are widely applied to recognise human-centred information based on features extracted from audio signals [5–7].

Psychotherapeutic competence is one of the behaviours that can influence the quality of therapy delivered and this is manually rated via the manual analysis of recordings of sessions and the use of observational rating scales (i.e., typically by trained raters such as clinical supervisors or clinical trainers). Competency concerns the standard of delivering a psychological treatment in a theoretically consistent matter to achieve its expected therapeutic effects [8]. Competency ratings are made based on the presence and quality of certain pre-determined features and an example of this is the Low-Intensity Cognitive Behavioural Therapy (LI-CBT) treatment competency scale [9]. The LI-CBT treatment competency scale is a valid and reliable, observational and scaled measure used for rating practitioners during treatment sessions. The scale consists of six items including: focusing the session, enabling engagement within the guided self-help approach, interpersonal competencies, ability to gather information: specific to change, use of self-help change methods and shared decision making [10]. Low-intensity psychological interventions are designed for mild-to-moderate intensity psychological disorders (typically anxiety and depression). They are defined as being brief in duration and utilising psychoeducational materials to facilitate change within a sound therapeutic alliance. Low-intensity psychological interventions are often delivered on the telephone for ease of access. Concerns have been raised relating to manual competency rating processes, due to it being very time-consuming, the variability of ratings and the bias of ratings applied by supervisors who might know the practitioners [11,12]. Therefore, there is a distinct need for an automatic tool to predict practitioners' competency ratings to improve the efficiency of the process, reduce bias and improve patient care. In additional, the practitioner's skills, such as competence, have been found to contribute positively to the therapeutic alliance. The practitioner's personal qualities and techniques have a positive influence on the understanding or repair of ruptures. That is, any deterioration in the therapeutic alliance [13,14]. A system for the automatic prediction of the practitioner's competence scores could therefore assist in defining the existence of a positive or ruptured therapeutic alliance in the session.

This paper investigates the automatic detection of psychotherapeutic competence. It explores both the problem of classifying the level of competency as well as that of predicting the competency measure as a total and item-based scores. The use of both acoustic and linguistic features is adopted in the study to explore the features that truly map to the treatment competency in the session. Experiments have been conducted using a dataset of real audio recordings. Both the patient's and practitioner's speech are used for the feature extraction stage in the machine learning pipeline. The features that achieved high performance in the system are explored based on the therapeutic alliance perspective in the study. The study therefore investigates the potential efficacy of implementing an automatic system for detecting the practitioner's competence in real clinical settings by the use of acoustic and linguistic features.

The system will use both acoustic and linguistic features. Our earlier results reported in [15] demonstrated the feasibility of predicting the continuous emotional labels from the guided self-help sessions dataset used in this study using acoustic features with respect to practitioner' competency. The acoustic features that demonstrated efficient prediction results in the study were the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) and the Bag of Audio Words (BoAWs). Therefore, those acoustic features are adopted in this paper as a suitable indicator for predicting the practitioner's competency measure.

The speakers' language in sessions could also reveal the emotions present during the session, along with the strength of those emotions. The use of language-based (linguistic) features to capture speakers' emotions could also improve the accuracy of the prediction of practitioner competency. The Referential Activity (RA) is a dictionary-based model that estimates the degree to which language is associated with emotions. It was built to analyse the language of therapeutic sessions and how it is connected to clinical evaluation. It depends on a set of functional stages that describe 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 [16,17]. Therefore, the RA model scores have been implemented as the language-based (linguistic) features in this study. The linguistic features could provide complementary information to the acoustic features. Consequently, the fusion of acoustic and linguistic features mentioned earlier is examined to predict the practitioner's competency.

To the best of our knowledge, this is the first work to predict the competency of delivery of a low-intensity psychological intervention automatically using real patient-practitioner conversations drawn from treatment sessions. These interventions are being increasingly used in public health systems due to the low cost of training and delivery. Furthermore, this is the first study that investigated the RA measures as language-based features extracted from the practitioner's and patient's speech to predict the practitioner's competency ratings. The study investigates the potential efficacy of implementing an automatic system for detecting the practitioner's competence in real clinical settings using acoustic and linguistic features.

A growing body of literature recognises the importance of acoustic features in the field of therapy. A study 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. 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, such as mean, minimum and maximum. The results showed that the acoustic features outperformed the behavioural codes in predicting the state of the marriage [18]. Another study 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 women; 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 [19]. In a recent study related to competency, automatic nonverbal analysis of a presenter's competency and behaviour was conducted based on real-world audiovisual recordings [20]. 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 found to be the most dominant nonverbal cues used to estimate presentation competence. Due to the high correlation between emotions and competency, especially in the rating criteria specified for the treatment competency raters, the use of acoustic features is selected as an appropriate approach for detecting the practitioner's competence. The eGeMAPS and BoAWs are considered from the two feature sets that show reliable results in the field of time-continuous recognition of dimensional emotions [5,15,21–24].

Several studies have explored the use of linguistic features in the field of therapeutic alliance. A study examined how to indicate the correlation between emotional elaboration and therapeutic alliance within a single session [25]. In the study, they allocated 40 patients with varying diagnoses to be videotaped, transcribed and analysed using linguistic measures of the referential process using the Discourse Attributes Analysis Program (DAAP), followed by a human-centred scoring with the Working Alliance Inventory (WAI) for every 5-min interval. The results showed that if the patients' ratings indicated more emotional engagement with their experience and were 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 has investigated the connection between non-verbal emotional experiences and how verbal language is affected by ruptures during treatment [26]. They employed linguistic measures of the referential process in association with measures of the therapeutic alliance. A scored measure was utilised 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 compared to non-ruptured segments and an increase in a measure of distancing. The practitioners showed similar patterns to the patients during rupture, in addition to self-disclosure, increased attempts at emotional control, increased references to bodily experiences and increased natural affect words. From the recent studies that helped identify the psychological orientation related to COVID-19, a study focused on recent COVID-19 tweets to apply for sentiment classification works, especially in the Nepali language. They proposed different CNN models for the sentiment classification of Nepali tweets using other feature extraction methods. The proposed methods showed stable and promising classification performance after being validated against traditional machine learning and state-of-the-art techniques [27].

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 [28,29]. Overall, the studies highlighted in this section demonstrated the beneficial effects of the acoustic and linguistic features and their fusion in the counselling domains. In this paper, to predict the 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 classification and prediction model.

Several research studies in psychotherapy tend to employ Automatic Speech Recognition (ASR) to investigate automatic approaches to detect measures related to the quality of therapy and therapeutic outcomes. The recognition accuracy is affected by several characteristics such as the rate of speech, interpretations of prosodic features presented in the pitch, power, intonation, stress, speaker age, and variations in pronunciation. It also can be affected when the same speaker speaks the same word [30]. 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 with ranges reported around 30–40% Word Error Rate (WER) compared to other ASR systems trained on controlled recording conditions with a large training data size [31]. Several studies in the literature confirmed these results, such as a study that 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 empa-

thy [32]. 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%. Another study developed an automatic pipeline that transcribed 225 Cognitive Behaviour Therapy (CBT) sessions' recordings and extracted linguistic features for behavioural coding of the sessions [33]. They adopted an ASR system based on the Kaldi pipeline. The reported WER in the study was 44.01%. A recent study developed an automated competency rating tool for processing audio recordings from 188 motivational interviewing sessions in a real-world clinical setting. The recordings were manually coded using behaviour codes from utterance and session levels. The investigated automatic system was composed of several modules; one of those modules was ASR based on a Time-Delay Neural Network (TDNN) [34]. They reported a WER for different test sets of 37.1% and 30.4% for automatically transcribing the sessions using the Kaldi tool, ignoring the speaker labels and concatenating all the utterances of the session and a WER of 38.1% and 31.6% using machine-generated segments.

The remainder of the paper is organised as follows. First, the dataset used in the paper experiments is described in Section 2. The proposed method is described in Section 3, which includes the extracted features and the used machine learning pipeline. Experimental results are shown in Section 4. Finally, conclusions are drawn in Section 5.

2. GSHTS Dataset Description and Processing

2.1. General Description

The recordings of the guided self-help sessions (GSHTS) used in this paper originate from research by [35] aimed at comparing the efficiency and clinical durability of two treatments for anxiety disorders delivered in the National Health Service in the UK. A total of 54 session recordings were collected for the study in 2019 and 2020; each includes a conversation between a practitioner and a patient during a guided self-help treatment session. The practitioners delivered low-intensity interventions for mild-to-moderate anxiety. During guided self-help, practitioners are highly active and guide the patients through the treatment, which is in contrast to more traditional therapy styles [36]. The low-intensity sessions typically last between 30–35 min. The total duration of sessions analysed is approximately 27 h, with a total of 9271 speaking turns (in the following referred to as segments) and a total of 264,069 number of words. Table 1 shows the patient demographics and therapy session information of the dataset using rounded values. As part of the data collection took place during the COVID-19 pandemic (and subsequent lock-down periods), some sessions are recorded face-to-face (in-person sessions) whilst the remainder of the sessions were conducted via the telephone.

Each session is analysed based on speaking duration in seconds and the spoken words of each speaker, either practitioner or patient. Figure 1 shows the total number of seconds and the total number of words per session for each speaker. Knowing the speaking turn lengths for the patient and practitioner may aid in determining any impact of this on the extracted features. It is clear from the figure that during most of the session time, the practitioner would speak more than the patient in terms of the number of words and time. This is consistent with the highly active psychoeducational guided self-help approach.

Table 1. Patient demographics and therapy session information for the GSHTS 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 (mins)	1634	30	20	52
Number of words	264,069	4890	354	5462
Number of segments	9271	172	35	194
Time talking per session (mins)				
Patient	644	12	3	22
Practitioner	927	18	8	28
Words spoken per session				
Patient	104,981	1944	354	3997
Practitioner	159,088	2946	1351	5462
Number of segments per session (N)				
Patient	4526	84	35	193
Practitioner	4745	88	36	194

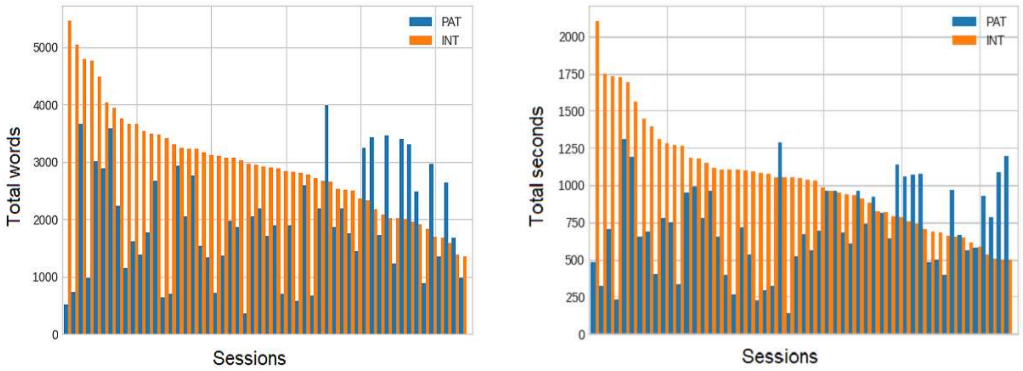

Figure 1. Dataset details per session for each speaker (PAT: patient, INT: practitioner) (**Left**) Number of words. (**Right**) Total seconds.

2.2. The LI-CBT Treatment Competency Scale

There are 51 sessions in the dataset that has associated mood outcome scores: the Patient Health Questionnaire (PHQ-9), the Generalised Anxiety Disorder (GAD-7) and the LI-CBT treatment competency scale. The PHQ-9 consists of nine depression questions, each scored from zero to three, summing up to 27 points. 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. The LI-CBT treatment competency scale is a scaled measure used to rate practitioners' skillfulness during the treatment of patients with mild-moderate depression and anxiety disorders. It consists of six items that enable treatment session raters to examine a range of competencies [10]:

- 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 0 to 6 based on a competence level, for each competency item. The total score may range from zero to 36 and the psychometric reliability and validity of the measure have been established.

The focusing the session competency item concentrates on rating the practitioner's ability to develop and subsequently adhere to an agenda for the treatment session. The continued engagement competencies item assesses the continuous engagement of the patient in the process of change in a collaborative manner. The practitioner should guarantee that the patient's progress is acknowledged by reflection and summaries. The interpersonal competencies item depends on ratings of the practitioner'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 practitioner's competence in gathering information from the patient concerning the changes made and improvements achieved over the course of treatment in a positive and considerate manner. The within session self-help change method item evaluates the practitioner's ability to select the appropriate treatment method for the patient based on their problem and following established treatment principles. The scale item planning and shared decision making competencies determines the practitioner's competence in planning 'homework' actions related to the patient's needs, session content and stage of treatment, taking into consideration the appropriate evidence base. Additionally, practitioners should ensure that planning and associated decisions are made collaboratively in a patient centred style [10].

In terms of using these recordings for automatic processing of the speech, several challenges should be noted. The practitioner and the patient may experience several inner states or behaviours during the session that would be reflected in their speech acoustics, such as the practitioner's empathetic state or the patient's anxious emotions. During episodes of extreme emotional states, the ability to hear some words would be challenging, such as if the patient was crying. Sessions recorded in a real-world environment could also include background noises such as doors slamming, chairs moving, clocks ticking, practitioners typing, patients coughing and room reverberation. Such background noises make it challenging to recognise words. Finally, due to the Coronavirus Disease 2019 (COVID-19) pandemic, some sessions were recorded over a mobile phone which affects the acoustic quality of the recorded signal.

2.3. Data Processing

To aid the training of the ASR system, the sessions were sent for transcription using a third-party agency. Each speaker's turn in the conversation is labelled in the transcription with the speaker's ID (PAT for patient and INT for practitioner). Furthermore, the dataset has been annotated (each speaking turn was annotated with the speaking start and end times in minutes done by the third-party agency) and adjusted (the second phase of annotation involved checking and adjusting the speaking start and end times of each speaking turn in seconds and eliminating the overlapping speech done by this study main author). The transcription process involved labelling the common filler words from several transcribers, with the finalised filler words being, umm, oh, aa, um, ah, hmm and mm. The time alignment annotations for each segment's start and end times have been reviewed and adjusted manually by the paper's primary author to produce more accurate timestamps for each segment in the dataset using the 'Transcriber' software [37]. It is important to be able to distinguish the practitioner's from the patient's speaking turn due to differences in the acoustics and language-based characteristics expressed by each speaker in the sessions. Some patients suffer from high depression or anxiety, which could affect their use of words. For the automatic prediction of the competency measures experiment, each patient and practitioner segment is extracted separately and then concatenated based on each speaker for each session which helps in obtaining each speaker's acoustic and linguistic feature set separately.

2.4. Experimental Setup

The dataset has been partitioned into training and test sets for building the ASR system based on the segments (utterances). The training set represented 80% of the data which is 43 sessions with around 1307 min and 7417 utterances, and the test set represented 20% of the data which is 11 sessions with around 327 min and 1854 utterances. For interactional data such as this, the speech segments with overlapping speech is often removed before analysis, as they represent particularly challenging segments to transcribe [38]. For this reason, the overlapped speech has been eliminated from the dataset for the ASR experiment, and the data included in Table 1 exclude overlapping speech.

3. Methods

3.1. ASR

Speech recognition is considered a particular form of pattern recognition. The optimal goal of ASR is to find the word sequence given the input speech signal [39]. The success of ASR systems depends on the quality of the system's modelling is. Kaldi is an open-source speech recognition toolkit developed in C++. It provides several recipes for creating a ASR systems by utilising famous databases such as LibriSpeech and WSJ [40]. For the acoustic modelling part in ASR, the main extracted features are Mel filterbanks and MFCCs. They are extracted from window frames of 25 ms with a shift of 10 ms. In addition, 100-dimensional i-vectors has extracted capturing speaker and environment information (noise).

The LibriSpeech recipe has been adopted for the training of the ASR system. The recipe enhances the acoustic model training by introducing further acoustic models. The language model has been built using the SRI Language Modeling (SRILM) toolkit as advised in Kaldi [40]. Kaldi requires specific data preparation for the sessions' recordings and transcriptions. For that reason, several data processing stages are 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.

The GSHTS ASR system has been trained using the 54 manually transcribed sessions (as described in Section 2) using the LibriSpeech recipe, and involving Hidden Markov Model with Gaussian Mixture Model (HMM-GMM) and TDNN models plus transfer learning. Using the weight transfer method, a transfer learning approach has been applied to adapt the LibriSpeech model to the GSHTS dataset. Following the LibriSpeech recipe, the LibriSpeech corpus has been used to train a base TDNN acoustic model. High-dimensional MFCCs features have been extracted from the GSHTS 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 GSHTS dataset such that both structure and weights have been transferred from the training of the LibriSpeech corpus. Afterwards, the training stage is conducted using two epochs on the training set. The language model used for the transfer learning approach has been trained as a 4 g with Turing smoothing interpolated with the 4 g language model from the LibriSpeech corpus following the technique used in [41,42], who reported that the best performance is 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 GSHTS dataset has been organised into six equal folds, each containing nine sessions, on account of the relatively small overall dataset size. The micro-average WER has been estimated for the cross-validation considering the number of word imbalances between the

different folds. The final model has been used for generating the transcriptions used by the wider system for automatically detecting the treatment competency in the sessions.

3.2. Automatic Detection of the Treatment Competency

The pipeline for the automatic detection of the treatment competency system consists of: feature extraction, feature selection and classification/regression. The prediction of the treatment competency scores is considered a regression problem and the classification of the treatment competency score levels is considered a classification problem.

3.2.1. Feature Extraction

The extracted features consists of acoustic and linguistic features for each speaker individually. The acoustic features extracted are eGeMAPS and BoAWs using the same settings described in [15]. The eGeMAPS comprises 88 features covering the following acoustic features: spectral, cepstral, prosodic and voice quality information. Those feature groups are presented using a set of descriptors called Low-Level Descriptors (LLDs). The spectral related LLDs are Alpha ratio, Hammarberg index, Energy slope 0–500 Hz and 500–1500 Hz, and Spectral flux. The voice quality related LLDs are First, Second and Third Formant, Harmonic difference H1-H2, and Harmonic difference H1-A3. The prosodic related LLDs are loudness and F0. The cepstral related LLDs are the MFCCs from 1 to 4. In addition, 13 MFCCs and their deltas and delta deltas are computed using a set of acoustic LLDs. The MFCCs consider the human perception sensitivity at predefined frequencies, which converts the conventional frequency to Mel Scale frequency [43]. 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 [44]. The features are extracted using the openSMILE toolkit [45,46].

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 [47]. The BoAWs are extracted using the open-source toolkit openXBOW [47].

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.

The linguistic features are extracted based on the Discourse Attributes Analysis Program (DAAP) dictionaries scores. The 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. As shown in Figure 2, the RA model that is the base for the used linguistic features consists of three main categories: unweighted dictionaries, weighted dictionaries and dictionary covariations. The weighted dictionaries 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, which are as follows [16,48]:

- The Weighted Referential Activity Dictionary (WRAD) is an 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 [49].

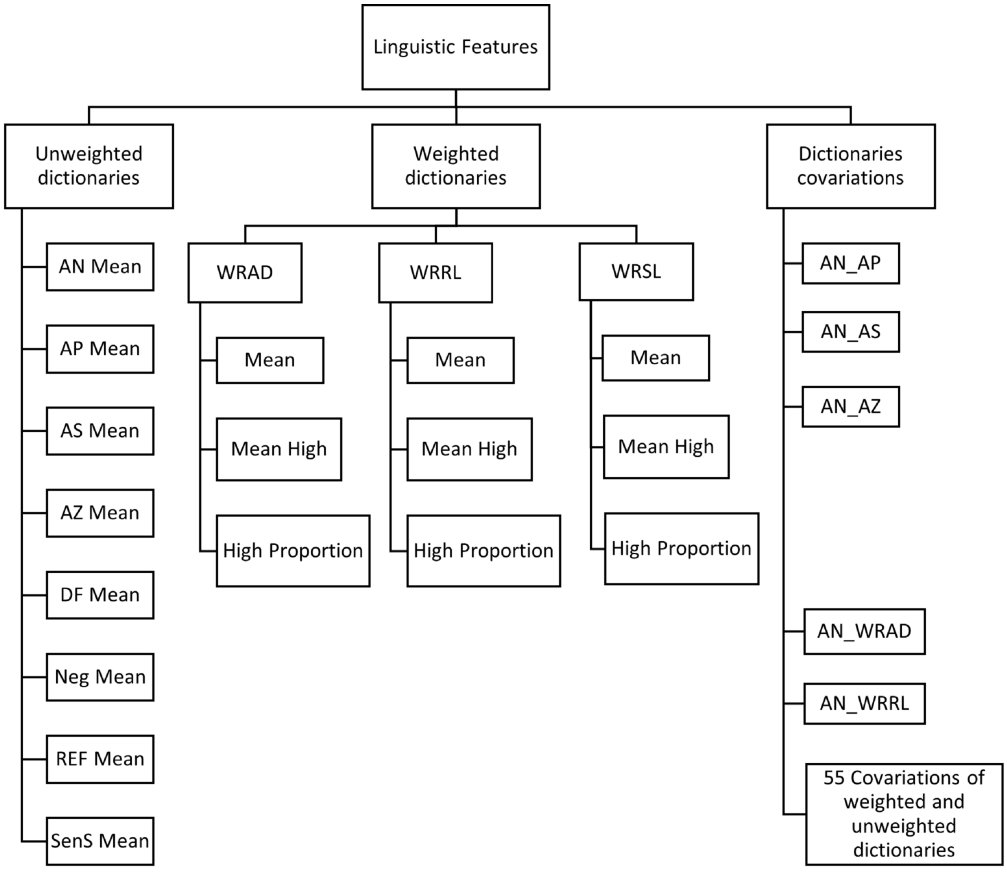

Figure 2. Extracted DAAP linguistic features tree diagram.

The unweighted dictionaries 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 [16,48]:

- 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 DAAP calculations depend on reading the words in each speaking turn (segment), comparing each word with the considered dictionary and assigning 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 [50].

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 [51]. 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 AN and AP covariation scores, AN and WRAD covariation scores, and WRAD and WRRL covariation scores. The filler words existed in the GSHTS dataset manual transcriptions described in Section 2.3 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. This is the first study that used this type of linguistic feature to train an ML model for predicting a practitioner’s competency measure.

3.2.2. Feature Selection

After combining the acoustic and linguistic feature vectors, the total features used in the classification and prediction system is composed of 520 features for both speakers. The Recursive Feature Elimination (RFE) approach has been applied for the feature selection phase in the experiment. 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 [52,53]. This approach has been proven to be successful for similar sparse data domains [54].

The features selection step involved using a cross-validation technique adopting the approach of Recursive Feature Elimination Cross-Validation (RFECV). The reason behind implementing RFE with cross-validation (RFECV) 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 RFECV has been implemented using the Scikit-learn library in Python [52,53]. 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. The number of folds selected in the cross-validation process has been fixed at six folds, which takes into account that the dataset is relatively small and imbalanced.

3.2.3. Estimating Treatment Competency; Classification and Prediction

Several classifiers have been evaluated to determine an efficient model for classifying the competency rating levels, such as Support Vector Machine (SVM), Decision Tree Classifier, Random Forest Classifier, Ada-Boost Classifier and Gradient Boosting Classifier. In addition, 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.

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 have been used to predict the competency rating in total and per item on the competency rating scale. Furthermore, the practitioner's level of competence has been classified based on various levels of the patient's depression and anxiety. The best acoustic and linguistic features have been reported based on the total ratings and each item in the competency rating scale for patients and practitioners.

4. Results and Discussion

4.1. ASR Results

Table 2 presents the results of the different ASR systems using 54 sessions for the HMM-GMM, TDNN, transfer learning and cross-validation.

Table 2. ASR system results using transfer learning and cross-validation techniques; for each model the data used for training, testing and transfer learning has been specified.

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

As presented in Table 2, we initially started by building a HMM-GMM model using only the GSHTS dataset and got an initial WER of 67.18%. Using instead an TDNN model yielded a WER of 47.18%. The following step was to train the TDNN model using the complete Librispeech database and test the model on the GSHTS dataset without any adaptation to investigate the use of a more common, and much larger, database in the ASR field, which resulted in a WER of 70.30%. This high WER can be attributed to the mismatch between the two datasets. In particular, the presence of more accented English in the GSHTS dataset rather than the read English speech of audiobooks in Librispeech [55]. After applying transfer learning using the acoustic model trained on the LibriSpeech corpus and adapted to the GSHTS dataset, the results gained were a WER of 35.86%, which showed better results than the initial system's results. These initial models were all trained using a fixed partitioning of the dataset into a training and a test part. The typical procedure when using sparse domain datasets, such as GSHTS, is to apply the cross-validation techniques. For our data the results were WERs with an average absolute lower value of 1.85% with WERs of 65.49%, 45.29% and 33.88% on the HMM-GMM, TDNN and transfer learning systems, respectively. This reflects that, when using cross-validation techniques, more data can be set aside for the training in each fold, resulting in more powerful models. After performing transfer learning, the final results fall in the range of what has been achieved in the literature. As discussed in Section 1, the range of the WER achieved after applying ASR

systems in therapeutic domains [32–34] is usually between 30% and 40% which is aligned with the results we gained on the guided self-help session recordings dataset.

4.2. Treatment Competency: Classification and Prediction

4.2.1. Prediction Results

This section presents the prediction results using the Pearson Correlation Coefficient (R), Mean Absolute error (MAE) and Root Mean Squared Error (RMSE). The results is reported for the total treatment competency rating and each rating item mentioned in Section 2.1. Furthermore, results from evaluating the best acoustic and linguistic features for predicting the total competency rating as well as for each rating item is presented in this section.

Table 3 presents the total practitioner's competency rating prediction results using a number of different regressors for manual and automatic transcriptions. A grid search resulted in the following parameters being set for the SVR and Lasso systems, which achieved higher performance rates: SVR with a C of 0.1 and epsilon of 1.00 with linear kernel, and Lasso with alpha of 0.03. Overall, the Lasso and SVR performed better than Linear Regression, Elastic Net, Decision Tree Regressor, Ada-Boost Regressor and Gradient Boosting Regressor. The best results are achieved for when using the manual transcripts; in particular, the Lasso regressor outperform that the SVR and exhibits 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 following investigation in this section concentrates on the use of Lasso and the manual transcripts.

Table 3. The total practitioner'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 actual versus predicted total practitioner's competency ratings are presented in Figure 3 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.

The achieved results indicate that it is feasible to use the practitioner's competency ratings as input for competency ratings prediction models, which eventually could help in assessing the quality of the therapy delivered to patients.

The results of the practitioner's competency measures based on each item as defined in the manual ([10]) are displayed in Table 4. It can be seen from the results in the table that the *focusing the session* and the *interpersonal competencies* rating items are easier to predict with R 0.89 and 0.87, respectively. This indicates the feasibility of detecting the practitioner's empathy, which is one of the rating specifications connected to the *interpersonal competencies* rating item. On the other hand, the *information gathering* and the *planning and shared decision-making competencies* rating items are harder to predict with R 0.75 and 0.73, respectively.

As this is the first study to investigate the practitioner's competence, it is of interest to explore the results based on the manual transcripts, especially looking at which features could accurately describe the competence from an acoustic and linguistic perspective. In the table, the number of features output from using the RFE feature selection methods (described in Section 3.2.1) is indicated. The number of features selected to predict the

continued engagement competencies rating item is the lowest number of features in comparison to the other items.

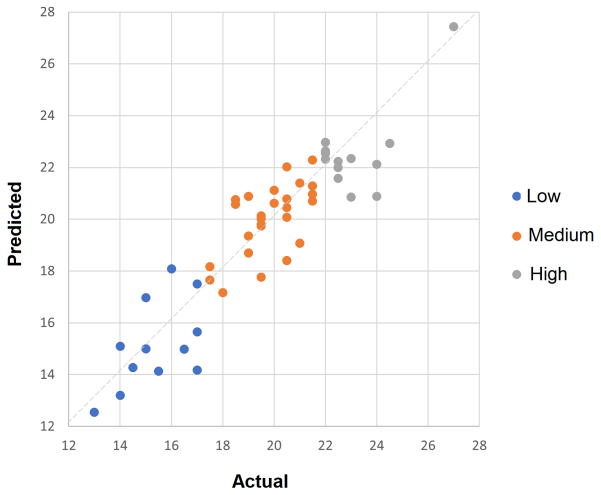

Figure 3. The actual versus predicted total practitioner's competence measure highlighting each level of competence.

Table 4. 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

In general, applying the RFE method revealed the best, selected acoustic and linguistic features for predicting the total practitioner's competency measure, as presented in Figure 4 for the patient's (PAT) and practitioner's (INT) speech. As shown in the figure, the best eGeMAPS features for the patient is the standard deviation of the third formant frequency and this feature could reflect on the vowel pronunciation in the patient's speech. For the practitioner, the mean of the third MFCC coefficient has been selected as the most optimal feature for predicting the competency, which could reflect on the practitioner's voice timbre [56]. The best linguistic feature for the patient is the MHigh of WRAD, which relates to the patient being in the RA symbolising phase, verbally immersed in a memory or a dream. For the practitioner, the best linguistic features are the WRRL MHigh, the AP and AS covariation, the AZ and WRSL covariation and the WRAD and WRSL covariation. The former best features can be divided into two groups based on the practitioner's levels of competence: the high competency practitioners would be expected to 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 practitioners express with a natural affect language (AZ) neither positive nor negative located either in the arousal phase of the referential process (WRSL) or in the symbolising phase (WRAD), which means that the practitioner in those phases is not reflecting or engaging with the patient and this is indicated by the low to medium ratings of competency. Measuring the covariation between those dictionaries could capture the low to medium level of practitioner's competence, as found in the selected features by the prediction model. It is more difficult to provide an explanation for the best BoAWs features, because each selected word corresponds to a combination of MFCC acoustic features.

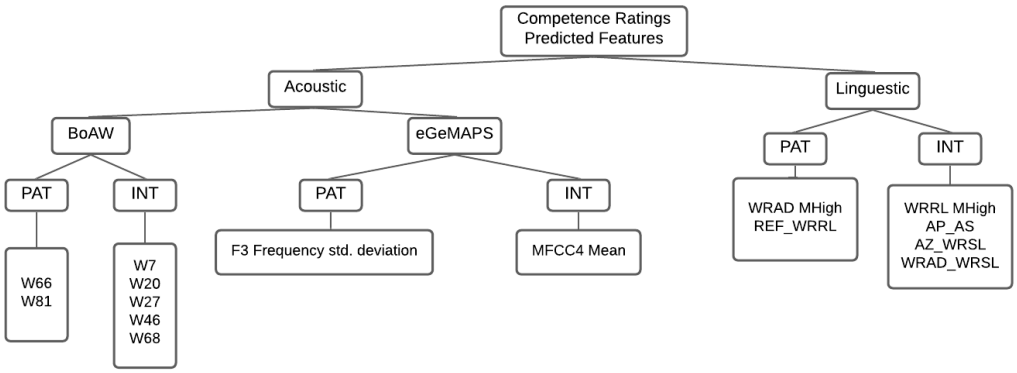

Figure 4. The best acoustic and linguistic features for predicting the total competency measure (std. deviation = standard deviation, PAT = patient, INT = practitioner).

The best eGeMAPS features for predicting the full ratings and each rating item has been presented in Figure 5 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 [57]. Furthermore, the best eGeMAPS features are composed of more patient related features compared to the practitioner related features. This could contribute to the ability of acoustic features to reveal the patient behaviours that could not be captured by language.

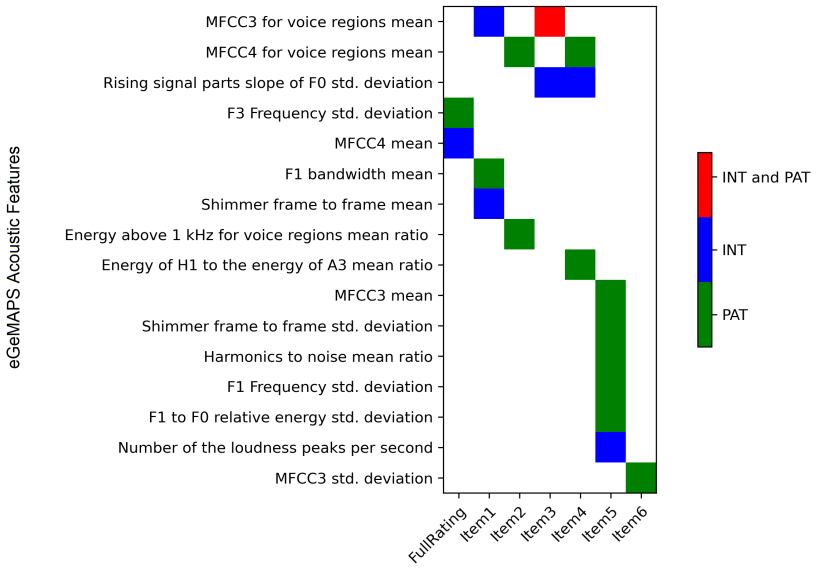

Figure 5. The best eGeMAPS features represented on the horizontal axis for predicting the full competency rating or each rating item as indicated on the vertical axis. The rating item numbers are based on the numbers in Table 4. The green blocks indicate the selected features for the patient's segments, the blue blocks indicate the selected features for the practitioner's segments and the red blocks indicate the selected features for the full session, including the patient's and practitioner'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 6. The most common linguistic features are: AZ and WRSL covariation, WRAD and WRSL covariation, SenS and WRSL covariation, AN and SenS covariation, WRSL mean, AZ mean and Neg mean. These features could indicate moments of rupture in the sessions based on the work of [26], such that WRSL could indicate a

decrease in the emotional engagement for the patient and self-disclose for the practitioner. Furthermore, the SenS score could map to the practitioner'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 practitioner's speech is higher than the ones related to the patient because the most dominant approach for the practitioner is to communicate and express verbally with the patient in the session.

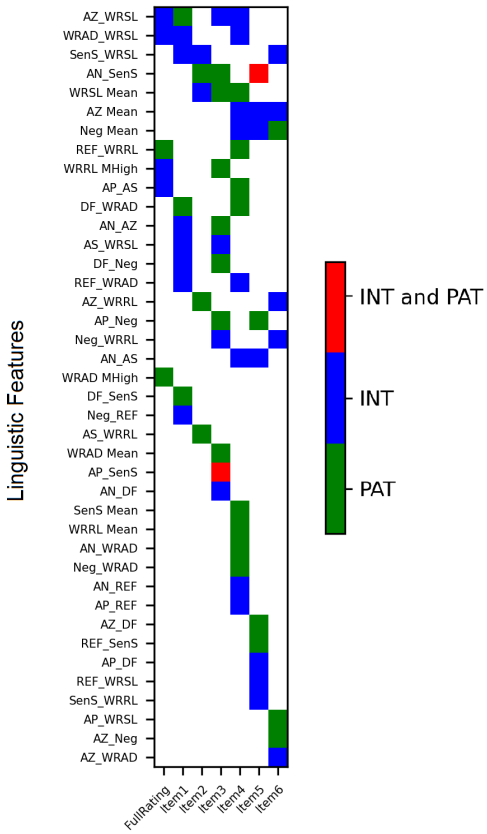

Figure 6. The best linguistic features are represented on the horizontal axis, for predicting the full competency rating or each rating item represented on the vertical axis. The rating item numbers are based on the numbers in Table 4. The green blocks indicate the selected features for the patient's segments, the blue blocks indicate the selected features for the practitioner's segments and the red blocks indicate the selected features for the full session, including the patient's and practitioner's segments.

4.2.2. Classification Results

The classification results are reported using the accuracy, precision, and recall scores. The main levels of competence to be classified are low, medium, and high. 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 5 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 5. 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 study also investigated the relationship between the patient's current level of depression and anxiety and the practitioner's level of competence in that session. For that reason, the practitioner's levels of competence have been classified using manual transcriptions based on the different levels of depression and anxiety. Figure 7 shows the classification results based on the patient's level of anxiety using the GAD-7 scores. In contrast, Figure 8 shows the classification results based on the patients' level of depression using the PHQ-9 scores. Due to a large number of occurrences of medium levels of competence in the GSHTS 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 3. Furthermore, it is clear from the results that the model is capable of classifying higher levels of competence in higher levels of depression and anxiety, which also relates to the higher prevalence of these types of data samples in the dataset.

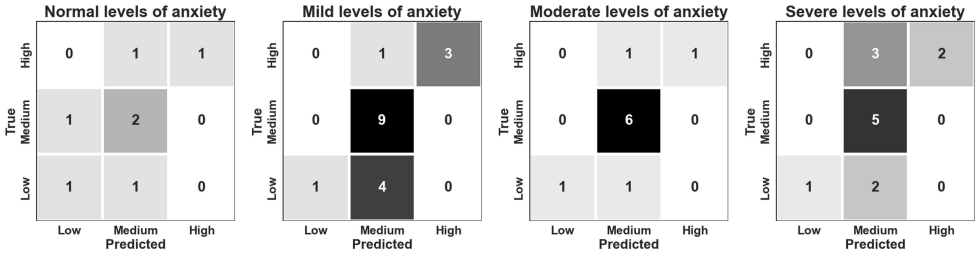

Figure 7. The confusion matrix results of the practitioner's level of competence (low, medium, high) based on the patient's level of anxiety (normal, mild, moderate and severe).

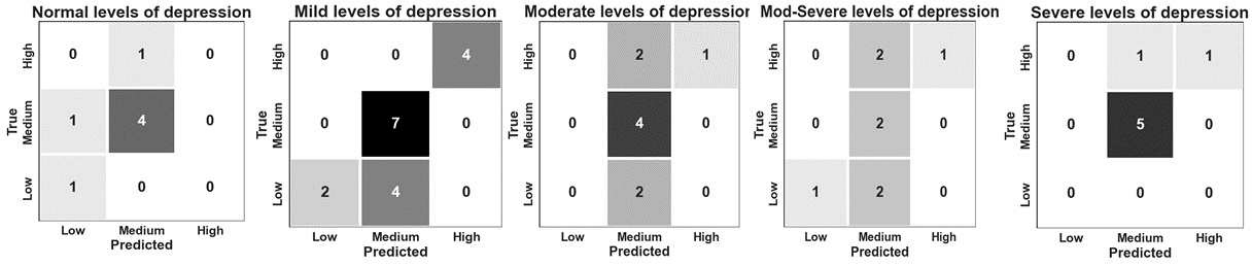

Figure 8. The confusion matrix results of the practitioner's levels of competence (low, medium, high) based on the patient's level of depression (normal, mild, moderate, moderately severe, and severe).

5. Conclusions

In this article, we presented and analysed a processing pipeline that is capable of automatically detecting practitioner competence by evaluating the recordings of guided self-help sessions for patients with mild-to-moderate anxiety.

The pipeline mentioned above consisted of two main sub-modules: ASR and the automatic detection of the practitioner's competence in recordings of real guided self-help sessions. The ASR system achieved relatively low performance, which is a reflection of the challenging data and not out of line with similar systems on other health domain data.

Despite that, the automatic transcriptions achieved from the ASR system are considered a major part of this study, and the corresponding prediction results are comparable to those achieved with the manual transcriptions with a good level of accuracy. As for the sub-module related to the automatic detection of the practitioner's competence, we focused on predicting and classifying the treatment competency measures as discrete scores and as levels of competence. Both the accuracy (90.20%) and the correlation coefficients (0.92) results denote higher classification and prediction results on the manual transcriptions, which declare the ability of the machine learning models to detect the practitioner's competence in real settings with the use of both acoustic and language-based features. Furthermore, the prediction of the treatment competency measures is deployed based on each item on the rating scale. The results showed that the items that depend on more personal qualifications related to the practitioners gained higher prediction results than the other items. That is a confirmed sign of the practitioner's language showing signs of competence in the session. In addition, the results are presented based on the features that gained higher performance results. Those features indicated signs of a ruptured therapeutic alliance based on what was reported in [26].

To conclude, the deployment of the system developed here in real-world busy clinical settings and the context of a pragmatic randomised clinical trial could enhance the therapeutic experience for the patient and guarantee fast and low-cost feedback for clinical supervisors. Such feedback would be invaluable both in training qualified practitioners and effectively supervising qualified practitioners, improving the delivery of interventions, and potentially positively impacting the patients' experience and outcomes.

In future work, the real deployment of such a tool will give insights into its acceptability in routine clinical services. This could include conducting a follow-up study to explore practitioners' attitudes towards using such a tool to engage various stakeholder groups, such as supervisors, practitioners, and even patients.

Author Contributions: Conceptualization, D.A. and H.C.; methodology, D.A.; software, D.A.; validation, D.A., H.C. and C.B. (Chris Blackmore); formal analysis, D.A.; investigation, D.A.; resources, J.S., C.B. (Charlotte Bee), V.A. and S.K.; data curation, D.A. and N.P.; writing—original draft preparation, D.A.; writing—review and editing, D.A., S.K., C.B. (Chris Blackmore) and H.C.; visualization, D.A.; supervision, H.C.; project administration, H.C.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: The research of the principal investigator has been supported by the Saudi Ministry of Higher Education. The clinical trial was funded by Association of Cognitive Analytic Therapists and Catalyse.

Institutional Review Board Statement: The original trial was funded by Association of Cognitive Analytic Therapy (ACAT) and a charity called Catalyse.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable

Acknowledgments: The authors acknowledge the support of the Association of Cognitive Analytic Therapists and Catalyse.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Wickramasinghe, N.; Geisler, E. *Encyclopedia of Healthcare Information Systems*; IGI Global: London, UK, 2008.
2. Stegmann, G.; Hahn, S.; Liss, J.; Shefner, J.; Rutkove, S.; Kawabata, K.; Bhandari, S.; Shelton, K.; Duncan, C.; Berisha, V. Repeatability of commonly used speech and language features for clinical applications. *Digit. Biomarkers* **2020**, *4*, 109–122. [CrossRef] [PubMed]
3. Waltz, J.; Addis, M.; Koerner, K.; Jacobson, N. Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *J. Consult. Clin. Psychol.* **1993**, *61*, 620. [CrossRef] [PubMed]

4. McLeod, B.D.; Southam-Gerow, M.A.; Jensen-Doss, A.; Hogue, A.; Kendall, P.C.; Weisz, J.R. Benchmarking treatment adherence and therapist competence in individual cognitive-behavioral treatment for youth anxiety disorders. *J. Clin. Child Adolesc. Psychol.* **2019**, *48*, S234–S246. [CrossRef] [PubMed]
5. Ringeval, F.; Schuller, B.; Valstar, M.; Cowie, R.; Kaya, H.; Schmitt, M.; Amiriparian, S.; Cummins, N.; Lalanne, D.; Michaud, A.; et al. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, Seoul, Korea, 22 October 2018; ACM: New York, NY, USA, 2018; pp. 3–13.
6. Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* **2011**, *53*, 1062–1087. [CrossRef]
7. Liang, Y.; Zheng, X.; Zeng, D. A survey on big data-driven digital phenotyping of mental health. *Inf. Fusion* **2019**, *52*, 290–307. [CrossRef]
8. Kohrt, B.; Jordans, M.; Rai, S.; Shrestha, P.; Luitel, N.; Ramaiya, M.; Singla, D.; Patel, V. Therapist competence in global mental health: Development of the ENhancing Assessment of Common Therapeutic factors (ENACT) rating scale. *Behav. Res. Ther.* **2015**, *69*, 11–21. [CrossRef]
9. Kellett, S.; Simmonds-Buckley, M.; Limon, E.; Hague, J.; Hughes, L.; Stride, C.; Millings, A. Defining the Assessment and Treatment Competencies to Deliver Low-Intensity Cognitive Behavior Therapy: A Multi-Center Validation Study. *Behav. Ther.* **2021**, *52*, 15–27. [CrossRef]
10. Kellett, S.; Simmonds-Buckley, M.; Limon, E.; Hague, J.; Hughes, L.; Stride, C.; Millings, A. Low Intensity Cognitive Behavioural Competency Scale Manual. 2021, *Unpublished document*.
11. Fairburn, C.; Cooper, Z. Therapist competence, therapy quality, and therapist training. *Behav. Res. Ther.* **2011**, *49*, 373–378. [CrossRef]
12. Watkins, C.E., Jr. Educating psychotherapy supervisors. *Am. J. Psychother.* **2012**, *66*, 279–307. [CrossRef]
13. Ackerman, S.; Hilsenroth, M. A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clin. Psychol. Rev.* **2003**, *23*, 1–33. [CrossRef]
14. Weck, F.; Richtberg, S.; Jakob, M.; Neng, J.M.; Höfling, V. Therapist competence and therapeutic alliance are important in the treatment of health anxiety (hypochondriasis). *Psychiatry Res.* **2015**, *228*, 53–58. [CrossRef] [PubMed]
15. Attas, D.; Kellett, S.; Blackmore, C.; Christensen, H. Automatic Time-Continuous Prediction of Emotional Dimensions during Guided Self Help for Anxiety Disorders. In Proceedings of the FRIAS Junior Researcher Conference: Human Perspectives on Spoken Human-Machine Interaction (SpoHuMa21), Online, 15–17 November 2021.
16. Bucci, W.; Maskit, B. Beneath the surface of the therapeutic interaction: The psychoanalytic method in modern dress. *J. Am. Psychoanal. Assoc.* **2007**, *55*, 1355–1397. [CrossRef]
17. Mergenthaler, E.; Bucci, W. Linking verbal and non-verbal representations: Computer analysis of referential activity. *Br. J. Med Psychol.* **1999**, *72*, 339–354. [CrossRef] [PubMed]
18. Nasir, M.; Baucom, B.; Georgiou, P.; Narayanan, S. Predicting couple therapy outcomes based on speech acoustic features. *PLoS ONE* **2017**, *12*, e0185123. [CrossRef] [PubMed]
19. Amir, N.; Mixdorff, H.; Amir, O.; Rochman, D.; Diamond, G.; Pfitzinger, H.; Levi-Isserlish, T.; Abramson, S. Unresolved anger: Prosodic analysis and classification of speech from a therapeutic setting. In Proceedings of the Speech Prosody 2010-Fifth International Conference, Chicago, IL, USA, 10–14 May 2010.
20. Sümer, Ö.; Beyan, C.; Ruth, F.; Kramer, O.; Trautwein, U.; Kasneci, E. Estimating Presentation Competence using Multimodal Nonverbal Behavioral Cues. *arXiv* **2021**, arXiv:2105.02636.
21. Ringeval, F.; Marchi, E.; Grossard, C.; Xavier, J.; Chetouani, M.; Cohen, D.; Schuller, B. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In Proceedings of the INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association (ISCA), San Francisco, CA, USA, 8–12 September 2016; pp. 1210–1214.
22. Mencattini, A.; Mosciano, F.; Comes, M.; Di Gregorio, T.; Raguso, G.; Daprati, E.; Ringeval, F.; Schuller, B.; Di Natale, C.; Martinelli, E. An emotional modulation model as signature for the identification of children developmental disorders. *Sci. Rep.* **2018**, *8*, 14487. [CrossRef]
23. Gideon, J.; Schatten, H.; McInnis, M.; Provost, E. Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019.
24. Zhang, Z.; Lin, W.; Liu, M.; Mahmoud, M. Multimodal deep learning framework for mental disorder recognition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 344–350.
25. Atta, N.; Christopher, C.; Mariani, R.; Belotti, L.; Andreoli, G.; Danskin, K. Linguistic features of the therapeutic alliance in the first session: A psychotherapy process study. *Res. Psychother. Psychopathol. Process. Outcome* **2019**, *22*, 374.
26. Christian, C.; Barzilai, E.; Nyman, J.; Negri, A. Assessing key linguistic dimensions of ruptures in the therapeutic alliance. *J. Psycholinguist. Res.* **2021**, *50*, 143–153. [CrossRef]
27. Sitaula, C.; Basnet, A.; Mainali, A.; Shahi, T.B. Deep learning-based methods for sentiment analysis on Nepali COVID-19-related tweets. *Comput. Intell. Neurosci.* **2021**, *2021*, 2158184. [CrossRef]
28. Wiegersma, S.; Nijdam, M.; van Hessen, A.; Truong, K.; Veldkamp, B.; Olff, M. Recognizing hotspots in Brief Eclectic Psychotherapy for PTSD by text and audio mining. *Eur. J. Psychotraumatol.* **2020**, *11*, 1726672. [CrossRef]

29. Tavabi, L.; Stefanov, K.; Zhang, L.; Borsari, B.; Woolley, J.; Scherer, S.; Soleymani, M. Multimodal Automatic Coding of Client Behavior in Motivational Interviewing. In Proceedings of the 2020 International Conference on Multimodal Interaction, Virtual, 25–29 October 2020; pp. 406–413.
30. Bhardwaj, V.; Ben Othman, M.T.; Kukreja, V.; Belkhier, Y.; Bajaj, M.; Goud, B.S.; Rehman, A.U.; Shafiq, M.; Hamam, H. Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review. *Appl. Sci.* **2022**, *12*, 4419. [CrossRef]
31. Kodish-Wachs, J.; Agassi, E.; Kenny, P., III; Overhage, J.M. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. *Amia Annu. Symp. Proc.* **2018**, *2018*, 683. [PubMed]
32. Xiao, B.; Imel, Z.; Georgiou, P.; Atkins, D.; Narayanan, S. "Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS ONE* **2015**, *10*, e0143055. [CrossRef] [PubMed]
33. Chen, Z.; Flemotomos, N.; Ardulov, V.; Creed, T.; Imel, Z.; Atkins, D.; Narayanan, S. Feature fusion strategies for end-to-end evaluation of cognitive behavior therapy sessions. In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual Conference, 1–5 November 2021; pp. 1836–1839.
34. Flemotomos, N.; Martinez, V.R.; Chen, Z.; Singla, K.; Ardulov, V.; Peri, R.; Caperton, D.D.; Gibson, J.; Tanana, M.J.; Georgiou, P.; et al. Automated evaluation of psychotherapy skills using speech and language technologies. *Behav. Res. Methods* **2021**, *54*, 690–711. [CrossRef]
35. Kellett, S.; Bee, C.; Aadahl, V.; Headley, E.; Delgadillo, J. A pragmatic patient preference trial of cognitive behavioural versus cognitive analytic guided self-help for anxiety disorders. *Behav. Cogn. Psychother.* **2020**, *49*, 1–8. [CrossRef]
36. Firth, N.; Barkham, M.; Kellett, S.; Saxon, D. Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. *Behav. Res. Ther.* **2015**, *69*, 54–62. [CrossRef]
37. Barras, C.; Geoffrois, E.; Wu, Z.; Liberman, M. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Commun.* **2001**, *33*, 5–22. [CrossRef]
38. Renals, S.; Swietojanski, P. Distant speech recognition experiments using the AMI Corpus. In *New Era for Robust Speech Recognition*; Springer: Cham, Switzerland, 2017; pp. 355–368.
39. Wang, Y. *Automatic Speech Recognition Model for Swedish Using Kaldi*; KTH School of Electrical Engineering and Computer Science: Stockholm, Sweden, 2020.
40. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
41. Mirheidari, B.; Blackburn, D.; O'Malley, R.; Venneri, A.; Walker, T.; Reuber, M.; Christensen, H. Improving Cognitive Impairment Classification by Generative Neural Network-Based Feature Augmentation. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 2527–2531.
42. Mirheidari, B.; Pan, Y.; Blackburn, D.; O'Malley, R.; Christensen, H. Identifying Cognitive Impairment Using Sentence Representation Vectors. In Proceedings of the INTERSPEECH 2021, Brno, Czechia, 30 August–3 September 2021; pp. 2941–2945.
43. Sitaula, C.; He, J.; Priyadarshi, A.; Tracy, M.; Kavehei, O.; Hinder, M.; Withana, A.; McEwan, A.; Marzbanrad, F. Neonatal bowel sound detection using convolutional neural network and Laplace hidden semi-Markov model. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1853–1864. [CrossRef]
44. Eyben, F.; Scherer, K.; Schuller, B.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.; Epps, J.; Laukka, P.; Narayanan, S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
45. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
46. Valstar, M.; Gratch, J.; Schuller, B.; Ringeval, F.; Lalanne, D.; Torres Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; Pantic, M. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016; pp. 3–10.
47. Schmitt, M.; Schuller, B. OpenXBOW: Introducing the passau open-source crossmodal Bag-of-Words Toolkit. *J. Machine Learn. Res.* **2017**, *18*, 1–5.
48. Maskit, B. Overview of computer measures of the referential process. *J. Psycholinguist. Res.* **2021**, *50*, 29–49. [CrossRef] [PubMed]
49. Tocatly, K.; Bucci, W.; Maskit, B. *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; City University of New York: New York, NY, USA, 2019.
50. Maskit, B. The Discourse Attributes Analysis Program (DAAP) (Series 8) [Computer Software]. 2012. Available online: http://www.thereferentialprocess.org/dictionary-measures-and-computer-programs (accessed on 22 September 2021).
51. Murphy, S.; Maskit, B.; Bucci, W. 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, Denver, CO, USA, 5 June 2015; pp. 80–88.
52. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
53. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

54. Mirheidari, B.; Blackburn, D.; O'Malley, R.; Walker, T.; Venneri, A.; Reuber, M.; Christensen, H. Computational Cognitive Assessment: Investigating the Use of an Intelligent Virtual Agent for the Detection of Early Signs of Dementia. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
55. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
56. De Boer, J.; Voppel, A.; Brederoo, S.; Schnack, H.; Truong, K.; Wijnen, F.; Sommer, I. Acoustic speech markers for schizophrenia-spectrum disorders: A diagnostic and symptom-recognition tool. *Psychol. Med.* **2021**, *51*, 1–11. [CrossRef] [PubMed]
57. Corrales-Astorgano, M.; Martínez-Castilla, P.; Escudero-Mancebo, D.; Aguilar, L.; González-Ferreras, C.; Carde noso-Payo, V. Automatic assessment of prosodic quality in down syndrome: Analysis of the impact of speaker heterogeneity. *Appl. Sci.* **2019**, *9*, 1440. [CrossRef]