frontiers in Psychology

METHODS
published: 28 July 2020
doi: 10.3389/fpsyg.2020.01726

Supervised Speaker Diarization Using Random Forests: A Tool for Psychotherapy Process Research

Lukas Fürer[1]*, Nathalie Schenk[1], Volker Roth[2], Martin Steppan[1], Klaus Schmeck[1] and Ronan Zimmermann[1,3]

[1] Clinic for Children and Adolescents, University Psychiatric Clinic, Basel, Switzerland, [2] Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland, [3] Division of Clinical Psychology and Psychotherapy, Faculty of Psychology, University of Basel, Basel, Switzerland

Speaker diarization is the practice of determining who speaks when in audio recordings. Psychotherapy research often relies on labor intensive manual diarization. Unsupervised methods are available but yield higher error rates. We present a method for supervised speaker diarization based on random forests. It can be considered a compromise between commonly used labor-intensive manual coding and fully automated procedures. The method is validated using the EMRAI synthetic speech corpus and is made publicly available. It yields low diarization error rates (M: 5.61%, STD: 2.19). Supervised speaker diarization is a promising method for psychotherapy research and similar fields.

Keywords: supervised speaker diarization, psychotherapy process measure, dyadic audio analysis, EMRAI speech corpus, random forest

OPEN ACCESS

Edited by:
Giuseppe Sartori,
University of Padua, Italy

Reviewed by:
Cristina Mazza,
Sapienza University of Rome, Italy
Graziella Orrù,
University of Pisa, Italy

****Correspondence:***
Lukas Fürer
lukas.fuerer@posteo.net

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *10 February 2020*
Accepted: *23 June 2020*
Published: *28 July 2020*

Citation:
Fürer L, Schenk N, Roth V, Steppan M, Schmeck K and Zimmermann R (2020) Supervised Speaker Diarization Using Random Forests: A Tool for Psychotherapy Process Research. Front. Psychol. 11:1726. doi: 10.3389/fpsyg.2020.01726

INTRODUCTION

Human interaction is organized by interpersonal coordination that manifests itself in temporally coordinated behavior. Interpersonal coordination can be broadly grouped into behavior matching and interpersonal synchrony, which involve the rhythmic and "smooth meshing of interaction" over time (Bernieri and Rosenthal, 1991). During the dyadic interactions of psychotherapy, patients and therapists have been shown to synchronize in verbal, non-verbal, and physiological behavior (Marci et al., 2007; Ramseyer and Tschacher, 2008; Lord et al., 2015; Koole and Tschacher, 2016; Kleinbub, 2017). A growing body of empirical research has associated the degree to which interpersonal synchrony is present during therapy with therapeutic outcome (Ramseyer and Tschacher, 2014), empathy (Marci et al., 2007; Imel et al., 2014; Lord et al., 2015), the formation of the therapeutic relationship (Ramseyer and Tschacher, 2011), personality traits (Tschacher et al., 2018), and emotion regulation (Galbusera et al., 2019; Soma et al., 2019). Due to their integrative value, processes of interpersonal synchrony have thus moved to the center of attention of psychotherapy research and related fields (Ramseyer and Tschacher, 2006). In the case of non-verbal movement synchrony, motion energy analysis has become a widespread tool to quantify movement from video (Ramseyer, 2013). It is made available through standalone software (Ramseyer, 2019), a MATLAB implementation (Altmann, 2013), and an R-package for synchronization analysis and visualization (Kleinbub and Ramseyer, 2019). This allows researchers to engage non-verbal synchrony in an automized, objective, reproducible, and non-labor-intensive fashion in their respective setting and has accelerated research on non-verbal movement synchrony in the clinical dyad (Delaherche et al., 2012). In the same line, autonomic measures (heart rate, skin conductance, breathing) applied in the field of interpersonal physiology (Kleinbub, 2017) also benefit from accessible measurement in the naturalistic setting (Weippert et al., 2010; Pijeira-Díaz et al., 2016; Barrios et al., 2019). In contrast, studies on vocal quality or vocal coordination have not gained the same amount of attention (Imel et al., 2014; Reich et al., 2014; Tomicic et al., 2017; Soma et al., 2019; Zimmermann et al., 2020). This is somewhat surprising because audio recordings are a widely used tool for educational, scientific, and supervisory activities (Aveline, 1992) and, in comparison to video or physiological measures, are non-invasive and inexpensive to attain in high quality. However, while the processing of non-verbal movement or physiological measures is facilitated through software solutions and devices, post-processing of audio for quantitative statistics can be strenuous due to speaker diarization (Anguera et al., 2012).

SPEAKER DIARIZATION IN PSYCHOTHERAPY RESEARCH

Speaker diarization is the practice of determining who speaks when (Anguera et al., 2012). In other words, diarization means creating a feature stream indicating speaker identity over time. Diarization in psychotherapy research is currently practiced in two different ways. On one side researchers rely on manual annotation of speaker identity, being time intensive but accurate (Imel et al., 2014; Reich et al., 2014; Soma et al., 2019). On the other side researchers rely on unsupervised automated methods, presenting with a minor work intensity but also with higher error rates (Xiao et al., 2015; Nasir et al., 2017a,b). The term "unsupervised" indicates that the system is not given prior knowledge as to how the speakers are embodied in the audio features. Mostly, the audio stream is segmented into speaker homogenous segments, which then are clustered (Tranter and Reynolds, 2006). In the field of psychotherapy research, studies have used unsupervised methods producing diarization error rates above 10%. For example, Xiao et al. (2015) used automatic speech recognition in motivational interviewing to produce text-based empathy scores of sessions and compare them with human empathy ratings. They employed a clustering based unsupervised diarization procedure that produced an error of 18.1%. Nasir et al. (2017b) predicted the outcome of couple therapy using speech features. The audio stream was segmented to indicate speaker changes based on generalized likelihood ratio criteria, which then are clustered to provide speaker-homogenous segments. Average pitch information in these segments are then used to provide a speaker annotation (wife or husband). They report a diarization error rate of 27.6%. While fully automated diarization procedures are appealing, diarization error rates can substantially be improved when introducing a learning step into the procedure, based on a small quantity of pure data (Sinclair and King, 2013). This relates to the idea of supervised machine learning. A recent study on a new fully supervised speaker diarization method using recurrent neural networks reported an error rate of 7.6% on a corpus of telephone calls (Zhang et al., 2019).

As described, regarding diarization practices in psychotherapy research, researchers tend to rely either on manual coding, which makes research very cost intensive, or they resort to fully automized unsupervised methods. In order to overcome this obstacle and to accelerate scientific undertakings on audio recordings in psychotherapeutic settings, we introduce a method for supervised speaker diarization, developed to work for standard single microphone audio recordings of dyadic talk psychotherapies. Considering the workload, the supervised method is a compromise between work intensive manual annotation and error prone unsupervised methods. It involves creating a learning set and introducing a learning step prior to automatically diarizing the whole data set.

AIM OF THIS STUDY

The aim of this study is to present a supervised method for dyadic speaker diarization based on a random forest algorithm. The method is tested using a freely available speech corpus. In the future, this will allow testing alternative methods and refinements of the current method on the same data set. The code has been made publicly available (Fürer, 2020). The procedure has been aggregated to one function and the preparations to run the function have been documented. We hope that this allows researchers with minimal coding experience or unfamiliar with MATLAB to carry out analyses on their own. The method is conceptualized in MATLAB and relies on readily available components (Segbroeck et al., 2013; Giannakopoulos and Pikrakis, 2014). We hypothesize that the method will produce diarization error rates comparable to current supervised diarization methods employed in other fields (below 10% per dyad; Zhang et al., 2019). Based on using random forest algorithm, we further hypothesize that the dyadic out-of-bag error rate (explained below) will positively correlate with the dyadic diarization error calculated on a test set. In future studies, this would allow quality checks on a dyadic level without producing a separate test set.

METHODS

Random Forest

The presented method for supervised speaker diarization in dyadic psychotherapy is based on a random forest algorithm. While machine learning methods in general have gained attention in psychological research (Orrù et al., 2020), random forests can be considered a rather understandable machine learning algorithm that has already found its way into psychotherapy research (Imel et al., 2015; Masías et al., 2015; Husain et al., 2016; Sun et al., 2017; Wallert et al., 2018; Zilcha-Mano, 2019; Rubel et al., 2020; Zimmermann et al., 2020). The random forest algorithm is a machine learning classifier based on decision trees (Kotsiantis, 2013). The random forest combines a certain amount of decision trees in a single prediction model and is consequently also called an ensemble learner. It can be employed for regression or classification problems. When confronted with classification problems, the decision is a majority vote over all trees in the ensemble, which, in ensemble format, provides greater accuracy (Breiman, 2001). Major advantages of the random forest algorithm are that it is insensitive to multicollinearity in the input data and to variables that do not contribute to the classification strength (Imel et al., 2015). In our setting this is of importance since we don't know which variables will be important for which dyad, and it is assumed that speech features may be highly correlated. The "random" in random forest refers to the usage of a random subsample of variables and a random subsample of data entries in the learning set when growing each tree (Husain et al., 2016). The process of randomly selecting a subsample of data entries without replacement for the training of each tree is called bagging (Breiman, 2001). This bagging process allows for the calculation of an out-of-bag error rate, which can be considered an estimate for the generalization error (Breiman, 1996). For each entry in the learning set the trees not using this specific entry for learning can be identified. They are called the out-of-bag classifier. The out-of-bag error is the error produced by the out-of-bag classifier, estimated using only the learning set. Given our use case, the possibility to estimate the generalization error with only the learning set is useful: If we apply this method to new and real psychotherapy audio and calculate the out-of-bag error on the learning data, we can estimate the overall strength of the prediction in each dyad, informing us for which dyads the diarization worked well and for which it didn't. We therefore report the correlation of the dyadic out-of-bag error with the dyadic speaker error (explained below) calculated in the separate test set.

Supervised Diarization in Dyadic Psychotherapy

The dyadic nature of talk therapy allows for an assumption to simplify the otherwise more complicated diarization process: the number of speakers is known, two in this case. Relating to the idea of supervised learning, here, a classifier is given prior knowledge as to how the two speakers are embodied in the input features (supervised diarization). Fortunately, inside the context of psychotherapy research, the classifiers do not have to be generalizable to different dyads, but rather, multiple classifiers can be trained, each one specialized to diarize one dyad only. The necessary steps involve: (1) creation of a learning set for each dyad (human coder), (2) automatic silence detection, (3) automatic voice activity detection, (4) feature extraction, (5) learning to provide a dyadic classifier, (6) prediction in one dyad, and (7) data aggregation. The steps are explained below.

EMRAI Synthetic Diarization Corpus

The supervised diarization method is tested on the EMRAI Synthetic Diarization Corpus (Edwards et al., 2018). This corpus is based on the LibriSpeech Corpus (Panayotov et al., 2015), namely recordings of English audiobooks. The manual labeling of audio data for training purposes is extremely time intensive. Thus, the authors of the corpus have synthetically created both 2-person and 3-person "dialogues" with and without overlap by sequentially arranging spoken parts. The EMRAI synthetic diarization corpus thereby offers an opportunity for testing diarization systems built for the context of the dyadic conversations as given in talk therapy.

Silence Detection and Voice Activity Detection

For silence detection, an algorithm calculates an individual intensity threshold value for each session recording. For more information, please refer to the source code (Fürer, 2020). The result of silence detection is a vector indicating silence and non-silence windows in the audio file. In a second step, voice activity detection is performed using a robust and competitive voice activity detection system for MATLAB developed by Segbroeck et al. (2013). This differentiates between voice and noise in the non-silence windows. Voice activity detection was performed over the whole audio, not only in non-silence windows. The procedure feeds contextually expanded spectral cues related to speech (spectral shape, spectro-temporal modulations, harmonicity, and the spectral variability) to a standard Multilayer Perceptron classifier (Segbroeck et al., 2013).

Feature Extraction

In order to allow the classifier to accurately differentiate between patient and therapist speech, appropriate features need to be extracted from the audio file. We aimed at using an existing and open source MATLAB library to make the procedure replicable by others. Features are provided by the MATLAB Audio Analysis Library and its function "stFeatureExtraction" (Theodoros and Aggelos, 2014). The function yields a total of 35 audio features: energy, zero-crossing rate, entropy of energy, two spectral centroids, spectral entropy, spectral flux, spectral flux roll-off, 13 Mel-frequency coefficients, 12 chroma vectors, harmonic ratio, and mean fundamental frequency. All audio features and their calculations are described in detail in the introductory publication accompanying the library (Giannakopoulos and Pikrakis, 2014). Here, we will focus our description on the Mel-frequency coefficients (MFCCs), since they are crucial features for speaker diarization (Friedland et al., 2009). The calculation of MFCCs takes into account that our perception of the frequency spectrum is not linear (Goldstein, 2010). We perceive differences in lower frequencies as more predominant than differences in higher frequencies. This non-linear relationship is represented by the mel scale, a function which, informed by psychoacoustics, mimics the human auditory system (Zhou et al., 2011). First, the audio signal is represented in the frequency domain by calculating the log discrete Fourier transform. The power spectrum then is submitted to a mel-scale filter bank consisting of overlapping triangular bandpass filters. Their bandwidth and spacing are given by a linear mel scale interval (Umesh et al., 1999). That way, the frequency spectrum is filtered (warped) in the same way, as it is thought to be filtered in the auditory system. MFCCs are then provided as the discrete cosine transform of the mel-filtered log power spectrum, providing coefficients in the time scale (Kathania et al., 2019). The authors of the MATLAB Audio Analysis Library have calculated MFCCs according to Slaney (1998).

In addition to the features provided by the MATLAB Audio Analysis Library, we calculated HF500, being a voice quality ratio between high spectral energy (above 500 Hz up to 3500 Hz) and low spectral energy (80–500 Hz). It has extensively been used in arousal quantification from speech (Bone et al., 2014; Chen et al., 2016).

All features, including silence and voice activity detection, are calculated in non-overlapping windows of 0.1 s in width, and all features have been used in training and predicting the diarization models.

Learning and Classification

The described features are then used to train a random forest classifier per dyad to predict speaker identity based of the features using the available speaker annotations from the learning set. Only spoken parts (labeled with person-1 or person-2 speech, no silence) were introduced to the learning set. To illustrate, the learning set would contain the timestamps (start of utterance and stop of utterance) and a variable of speaker identity of all utterances for person-1 and person-2 in the first 10 min of the recording. Using the corresponding features, an ensemble of 500 trees is trained for each dyad, using Breiman's algorithm (Breiman, 2001). All EMRAI dialogues of length bigger than 20 min (n = 107) were selected. The first 10 min of each dialogue were chosen for learning purposes, while minutes 10–20 were used as a test set. This simulates the creation of a learning set in a naturalistic setting. Using 10 min of audio in the learning material means that each speaker is represented by less than 5 min of speech (Mean: 4.17 min, Std: 0.38). After training, the classifier is then used to predict speaker identity in the independent test set (minutes 10–20 of the respective recording), resulting in classifications of either person-1-speech or person-2-speech. Please note that the classifier would also yield a decision for actual silence windows; it was not trained to discriminate between silence, noise, and spoken parts. This requires aggregating information to a final decision.

Data Aggregation

After classification is acquired, three information streams must be aggregated in order to produce a final diarization vector. Results of silence detection (silence or no silence), voice activity detection (voiced or unvoiced), and random forest based diarization (person-1 or person-2 speech) are combined to a feature stream of 0.1s segments of either non-speech, person-1-speech, or person-2-speech according to the following rules: Windows classified as silence by the silence detection remain unchanged. Non-silence windows, however, are replaced by the information stream of the voice activity detection resulting in a combined stream indicating silence, non-speech/noise, and speech. The windows classified as speech are then replaced by the person-1-speech and person-2-speech labels obtained by the respective classifier. The resulting vector contains the labels "non-speech" (silence or noise), "person-1-speech," and "person-2-speech."

Error Reporting and Data Set

The performance of a speaker diarization method is assessed via the diarization error rate (Barras et al., 2006), a measure comprised of the sum of the following elements: (1) *speaker error* (SpE, percentage of times the wrong speaker is predicted), (2) *missed speech* (MSp, percentage of times silence is predicted instead of speech), (3) *false alarm speech* (FASp, percentage of times speech is predicted instead of silence), and (4) *overlap error* (percentage of times overlapped speech is not assigned to one of the respective speakers). Given our choice of using 2-person non-overlapping speech, the diarization error rate (DER) is reported as the sum of the first three errors (Reynolds and Torres-Carrasquillo, 2005). SpE, MSp, and FASp are reported as mean values with standard deviations over all dyads, same-sex dyads, and different-sex dyads. The sampling frequencies (fs) of the corpus and our prediction stream were different (fs corpus = 100, fs prediction stream = 10) insofar as 10 windows at a time of the corpus are summarized to match one window of our prediction. Transitional windows, where more than one classification was present in the corpus windows to be summarized (both speech and silence), are excluded from the analyses.

We also hypothesized that the dyadic out-of-bag error would be a useful measure to control for the quality of the diarization (speaker annotation) in any specific dyad. We report the correlation between the dyadic out-of-bag error and the dyadic SpE.

RESULTS

Total DER, Speaker Error, Missed Speech, False-Alarm Speech

Table 1 provides an overview over error rates. Although total mean DER can be considered low, there are differences between dyads, as already implied by higher error rates for same sex dyads than different sex dyads, $t(61) = 4.16$, $p = 1.01\text{e-}04$. While the error produced through silence detection and voice activity detection (MSp + FASp) seems to show high stability throughout dyads (Mean: 3.11, Std: 1.27), SpE is more prone to vary over dyads (Mean: 2.50, Std: 2.12). This is confirmed by the correlation of the total DER and the SpE, $r(105) = 0.83$, $p = 1.24\text{e-}28$. This implies that the variability in total DER is mainly produced by the SpE. Forty of 107 dyads presented a total DER below 5%. Ninety-one of 107 dyads had a total DER below 7.5%. Five dyads showed total DER above 10%. Hence there are dyads for which the method had somewhat increased DER (16 dyads with DER above 7.5%). Mean FASp error rates are below 1%, mean MSp error rates are located just above 2%.

As expected, the dyadic out-of-bag error did correlate positively with the dyadic SpE, providing evidence for the usefulness of the out-of-bag error to estimate the quality of the diarization for specific dyads ($r(105) = 0.85$, $p = 1.65\text{e-}30$, see **Figure 1**).

TABLE 1 | Mean *error rates* (Std) in percent over all dyads (n = 107), same-sex dyads (n = 44), and different-sex dyads (n = 63).

	Total DER	Speaker Error	MSp	FASp
All dyads	5.61 (2.19)	2.50 (2.12)	2.60 (1.07)	0.51 (0.88)
Same-sex dyads	6.48 (2.57)	3.60 (2.58)	2.47 (0.98)	0.41 (0.75)
Different-sex dyads	5.01 (1.64)	1.74 (1.28)	2.69 (1.13)	0.57 (1.04)

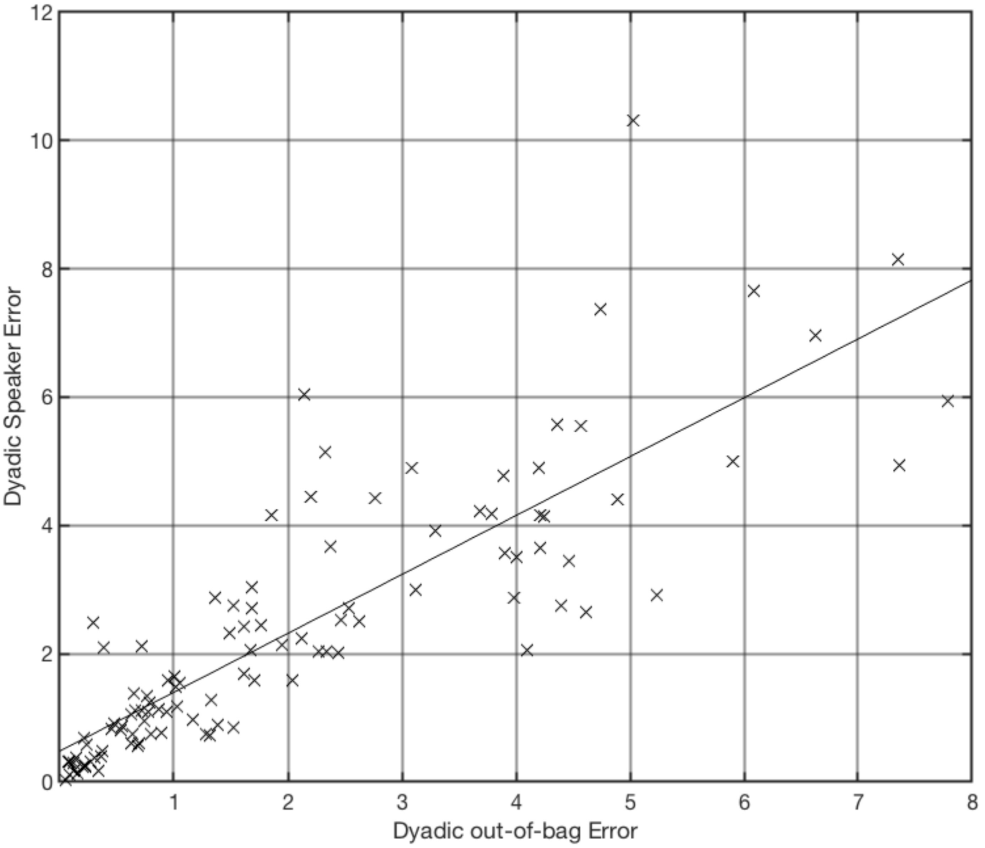

FIGURE 1 | Dyadic speaker error (e.g., speaker-1 predicted instead of speaker-2 and vice versa) in percent against dyadic out-of-bag error over all dyads (n = 107), r(105) = 0.85, p = 1.65e-30.

DISCUSSION

Speaker diarization in psychotherapy research has both been performed in a manual, time-consuming (Imel et al., 2014; Reich et al., 2014; Soma et al., 2019), and, alternatively, unsupervised, automated fashion (Xiao et al., 2015; Nasir et al., 2017a,b). For certain scientific contexts, a supervised procedure can be favorable, as it greatly reduces effort (manpower, time, and costs) compared to manual diarization and yields low error rates. In this study, we have described a method for supervised speaker diarization feasible for the dyadic nature of talk therapy. The method requires that the user manually creates a learning set of approximately 5 min cumulative length per speaker. A random forest classifier is trained from the learning set, one for each dyad, using speech features extracted by the MATLAB Audio Analysis Library (Giannakopoulos and Pikrakis, 2014). The classifier is then set out to diarize the whole amount of data (sessions) of this respective dyad. The distinction between voiced and unvoiced windows is made using an already existing procedure for voice activity detection by Segbroeck et al. (2013) and a custom silence detection algorithm. The method is made publicly available (Fürer, 2020). A major advantage of the study is that an open source speech corpus was used to present first results of the proposed method. The availability of the corpus allows other researchers to present results of other methods on the same data set or allows the test of the impact of improvements to the here proposed method.

The method shows satisfying diarization error rates (Mean: 5.61, Std: 2.19), comparable to other fully supervised methods (Zhang et al., 2019). Error rates are higher for same sex dyads (Mean: 6.48, Std: 2.57) than for different sex dyads (Mean: 5.01, Std: 1.64), t(61) = 4.16, p = 1.01e-04. This result is expected. The classifier is faced with features of higher degree in similarity when dealing with same-sex dyads, resulting in higher error rates. The difference of diarization error in those groups is mainly due to speaker error (confusion of the classifier toward the distinction of speaker one and speaker two). Speaker error and the total diarization error correlate with r(105) = 0.83, p = 1.24e-28, indicating that the total diarization error is mainly produced by speaker error, while missed speech and false alarm speech errors are more stable across dyads. While false alarm speech rates are substantially low (below 1%), missed speech rates are located around 2.5%. Low false alarm speech rates reflect the additional use of silence detection, which has shown to be very robust in differentiating silence windows from non-silence windows. *Post hoc* analyses for silence detection over the whole test set (all dyads together) reveal a miss rate (non-silence predicted instead of silence) of 0.70% and a false alarm rate (silence predicted instead of non-silence) of 1.51%. Considering the synthetic nature of the corpus used in this study, where the audio files of the corpus contain only silence or speech (no noise), voice activity detection may seem needless besides silence detection. In an environment, where one can be sure that no noise occurs (only speech or silence), the sole use of silence detection can be considered favorable. For later use of the method on naturalistic data, however, where noise may well be part of the equation, voice activity detection is indispensable and is therefore introduced as well. Both silence detection and voice activity detection have been incorporated in the code published (Fürer, 2020).

For 16 (out of 107) dyads, total diarization error exceeds 7.5%. When working with real psychotherapy data, it would be practical to be able to identify these dyads without creating a separate test set. Therefore, we tested whether the out-of-bag error presents a good estimate for the dyadic speaker error. The correlation showed to be high, $r(105) = 0.85$, $p = 1.65\text{e-}30$. We argue that the out-of-bag error can be used to make assumptions toward the quality of diarization, maybe leading to the exclusion of specific dyads, for reasons of error management. It is encouraged for future research to include the out-of-bag error as moderator variable to control for noise.

LIMITATIONS AND CHALLENGES

In comparison to manual annotations of speaker identity, unsupervised and supervised procedures of speaker diarization will be error prone. It is therefore important for future studies of this realm to report how and to what extent diarization errors influence the research findings at hand. As we reported, the out-of-bag error can be used for this purpose. However, there are no clear guidelines, for example, indicating the need to exclude a dyad for reasons of untolerable diarization error. Consequently, researchers are encouraged to at least publish diarization error rates and to test whether study results correlate with the diarization errors found.

Further, applying machine learning methods to psychotherapeutic data involves experience in programming. Proximity to data scientific or machine learning colleagues is not always guaranteed for workgroups invested in psychotherapy research. It was therefore important to us to publish the code used in this study (Fürer, 2020). The procedure is summarized to one function and an extensive explanation of preprocessing steps is given, in order to make it applicable by users with minimal coding experience.

While using a speech corpus may allow testing future improvements, future studies should invest in testing the proposed procedure on real psychotherapy data in order to clarify concerns toward the validity of results. For an application example of the procedure we refer the reader to the study of Zimmermann et al. (2020), which is using the method presented here to analyze the impact of silence across speaker switching patterns in psychotherapy sessions. Dyadic out-of-bag errors were comparable to the errors found here (Mean: 5.3%, Std = 3.3).

In light of the growing interest in interpersonal processes in psychotherapy, the supervised diarization applied in the study at hand may facilitate the exploration of dyadic vocal and conversational processes that may be linked to change processes, treatment outcome, diagnoses, and patient characteristics. Also, it may facilitate process research to uncover trajectories of variables of interest based on audio recordings. By catalyzing studies concerned with speech and conversational measures, psychotherapy research will gain in rater-independent, objective measures that can widely be used by various research groups and thereby provide results that are comparable and reproducible.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

LF, RZ, and VR contributed to the conception, method, and design of the study. LF performed the analysis and wrote the manuscript with support and revisions of NS, MS, KS, and RZ. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the University of Basel and the Psychiatric Hospitals of the University of Basel.

ACKNOWLEDGMENTS

This study is performed as part of the EAST psychotherapy study [clinicaltrials.gov: NCT02518906; Zimmermann et al. (2018)].

REFERENCES

Altmann, U. (2013). *Synchronisation Nonverbalen Verhaltens: Weiterentwicklung und Anwendung Zeitreihenanalytischer Identifikationsverfahren*. Berlin: Springer.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: a review of recent research. *IEEE Trans. Audio Speech Lang. Proc.* 20, 356–370. doi: 10.1109/TASL.2011.2125954

Aveline, M. (1992). The use of audio and videotape recordings of therapy sessions in the supervision and practice of dynamic psychotherapy. *Br. J. Psychother.* 8, 347–358. doi: 10.1111/j.1752-0118.1992.tb01198.x

Barras, C., Zhu, X., Meignier, S., and Gauvain, J.-L. (2006). Multistage speaker diarization of broadcast news. *IEEE Trans. Audio Speech Lang. Proc.* 14, 1505–1512. doi: 10.1109/TASL.2006.878261

Barrios, L., Oldrati, P., Santini, S., and Lutterotti, A. (2019). "Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis," in *Proceedings of the 13th EAI International Conference* *on Pervasive Computing Technologies for Healthcare - Pervasive Health*, New York, NY.

Bernieri, F. J., and Rosenthal, R. (1991). "Interpersonal coordination: behavior matching and interactional synchrony," in *Fundamentals of Nonverbal Behavior*, eds R. Feldman and B. Rimé (New York, NY: Cambridge University Press), 401–432.

Bone, D., Lee, C.-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*. Shanghai: INTERSPEECH.

Breiman, L. (1996). *Out-Of-Bag Estimation*.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Chen, H.-Y., Liao, Y.-H., Jan, H.-T., Kuo, L.-W., and Lee, C.-C. (2016). "A Gaussian mixture regression approach toward modeling the affective dynamics between acoustically-derived vocal arousal score (VC-AS) and internal brain fMRI bold signal response," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal synchrony: a survey of evaluation methods across disciplines. *IEEE Trans. Affect. Comput.* 3, 349–365. doi: 10.1109/T-AFFC.2012.12

Edwards, E., Brenndoerfer, M., Robinson, A., Sadoughi, N., Finley, G. P., Korenevsky, M., et al. (2018). "A free synthetic corpus for speaker diarization research," in *Speech and Computer*, eds A. Karpov, O. Jokisch, and R. Potapova (Cham: Springer), 113–122. doi: 10.1007/978-3-319-99579-3_13

Friedland, G., Vinyals, O., Huang, Y., and Muller, C. (2009). Prosodic and other long-term features for speaker diarization. *IEEE Trans. Audio Speech Lang. Proc.* 17, 985–993. doi: 10.1109/TASL.2009.2015089

Fürer, L. (2020). *Supervised Dyadic Speaker Diarization (Version v1) [Computer software]*. San Francisco: GitHub.

Galbusera, L., Finn, M. T. M., Tschacher, W., and Kyselo, M. (2019). Interpersonal synchrony feels good but impedes self-regulation of affect. *Sci. Rep.* 9:14691. doi: 10.1038/s41598-019-50960-0

Giannakopoulos, T., and Pikrakis, A. (2014). *Introduction To Audio Analysis: A MATLAB Approach*, 1st Edn, Cambridge, MA: Academic Press.

Goldstein, E. B. (2010). "Sound, the auditory system, and pitch perception," in *Sensation and Perception*, eds J.-D. Hague, and J. A. Perkins, 8th Edn (Wadsworth CENGAGE Learning), 490.

Husain, W., Xin, L. K., Rashid, N. A., and Jothi, N. (2016). "Predicting Generalized Anxiety Disorder among women using random forest approach," in *Proceedings of the 2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, Kuala Lumpur.

Imel, Z. E., Barco, J. S., Brown, H. J., Baucom, B. R., Baer, J. S., Kircher, J. C., et al. (2014). The association of therapist empathy and synchrony in vocally encoded arousal. *J. Counsel. Psychol.* 61, 146–153. doi: 10.1037/a0034943

Imel, Z. E., Steyvers, M., and Atkins, D. C. (2015). Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. *Psychotherapy* 52, 19–30. doi: 10.1037/a0036841

Kathania, H. K., Shahnawazuddin, S., Ahmad, W., and Adiga, N. (2019). "On the role of linear, mel and inverse-mel filterbank in the context of automatic speech recognition," in *Proceedings of the 2019 National Conference on Communications (NCC)*, Bangalore.

Kleinbub, J. R. (2017). State of the art of interpersonal physiology in psychotherapy: a systematic review. *Front. Psychol.* 8:2053. doi: 10.3389/fpsyg.2017.02053

Kleinbub, J. R., and Ramseyer, F. (2019). *RMEA: Synchrony in Motion Energy Analysis (MEA) Time-Series (Version 1.1.0) [Computer Software]*. Available online at: https://cran.r-project.org/web/packages/rMEA/index.html (accessed July 17, 2020).

Koole, S. L., and Tschacher, W. (2016). Synchrony in psychotherapy: a review and an integrative framework for the therapeutic alliance. *Front. Psychol.* 7:862. doi: 10.3389/fpsyg.2016.00862

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artif. Intellig. Rev.* 39, 261–283. doi: 10.1007/s10462-011-9272-4

Lord, S. P., Sheng, E., Imel, Z. E., Baer, J., and Atkins, D. C. (2015). More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behav. Ther.* 46, 296–303. doi: 10.1016/j.beth.2014.11.002

Marci, C. D., Ham, J., Moran, E., and Orr, S. P. (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *J. Nerv. Ment. Dis.* 195, 103–111. doi: 10.1097/01.nmd.0000253731.71025.fc

Masías, V. H., Krause, M., Valdés, N., Pérez, J. C., and Laengle, S. (2015). Using decision trees to characterize verbal communication during change and stuck episodes in the therapeutic process. *Front. Psychol.* 6:379. doi: 10.3389/fpsyg.2015.00379

Nasir, M. D., Baucom, B. R., Bryan, C. J., Narayanan, S. S., and Georgiou, P. (2017a). Complexity in speech and its relation to emotional bond in therapist-patient interactions during suicide risk assessment interviews. *Interspeech* 2017, 3296–3300. doi: 10.21437/Interspeech.2017-1641

Nasir, M. D., Baucom, B. R., Georgiou, P., and Narayanan, S. (2017b). Predicting couple therapy outcomes based on speech acoustic features. *PLoS One* 12:e0185123. doi: 10.1371/journal.pone.0185123

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., and Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Front. Psychol.* 10:2970. doi: 10.3389/fpsyg.2019.02970

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane.

Pijeira-Díaz, H. J., Drachsler, H., Järvelä, S., and Kirschner, P. A. (2016). "Investigating collaborative learning success with physiological coupling indices based on electrodermal activity," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK*, Amsterdam.

Ramseyer, F. (2013). "Synchronized movement in social interaction," in *Proceedings of the 2013 Inputs-Outputs Conference on An Interdisciplinary Conference on Engagement in HCI and Performance - Inputs*, New York, NY.

Ramseyer, F. (2019). Motion Energy Analysis (MEA). A primer on the assessment of motion from video. *J. Counsel. Psychol.* 67:536.

Ramseyer, F., and Tschacher, W. (2006). Synchrony: a core concept for a constructivist approach to psychotherapy. *Construct. Hum. Sci.* 11, 150–171.

Ramseyer, F., and Tschacher, W. (2008). "Synchrony in dyadic psychotherapy sessions," in *Simultaneity*, eds S. Vrobel, O. E. Rössler, and T. Marks-Tarlow (Singapore: World Scientific), 329–347. doi: 10.1142/9789812792426_0020

Ramseyer, F., and Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. *J. Consult. Clin. Psychol.* 79, 284–295. doi: 10.1037/a0023419

Ramseyer, F., and Tschacher, W. (2014). Nonverbal synchrony of head- and body-movement in psychotherapy: different signals have different associations with outcome. *Front. Psychol.* 5:979. doi: 10.3389/fpsyg.2014.00979

Reich, C. M., Berman, J. S., Dale, R., and Levitt, H. M. (2014). Vocal synchrony in psychotherapy. *J. Soc. Clin. Psychol.* 33, 481–494. doi: 10.1521/jscp.2014.33.5.481

Reynolds, D. A., and Torres-Carrasquillo, P. (2005). "Approaches and applications of audio diarization," in *Proceedings of the (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA.

Rubel, J. A., Zilcha-Mano, S., Giesemann, J., Prinz, J., and Lutz, W. (2020). Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—A demonstration. *Psychother. Res.* 30, 300–309. doi: 10.1080/10503307.2019.1597994

Segbroeck, M. V., Tsiartas, A., and Narayanan, S. (2013). *A Robust Front end for VAD: Exploiting Contextual, Discriminative and Spectral Cues of Human Voice. INTERSPEECH*. Shanghai: Shanghai International Convention Center.

Sinclair, M., and King, S. (2013). "Where are the challenges in speaker diarization?," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC.

Slaney, M. (1998). *Auditory Toolbox, Version 2. Technical Report*. Palo Alto, CA: Interval Research Corproation.

Soma, C. S., Baucom, B. R. W., Xiao, B., Butner, J. E., Hilpert, P., Narayanan, S., et al. (2019). Coregulation of therapist and client emotion during psychotherapy. *Psychother. Res.* 30, 591–603. doi: 10.1080/10503307.2019.1661541

Sun, B., Zhang, Y., He, J., Yu, L., Xu, Q., Li, D., et al. (2017). "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC*, Cham.

Theodoros, G., and Aggelos, P. (2014). *Introduction to Audio Analysis—1st Edn*. Available online at: https://www.elsevier.com/books/introduction-to-audio-analysis/giannakopoulos/978-0-08-099388-1 (accessed July 17, 2020).

Tomicic, A., Pérez, J. C., Martínez, C., and Rodríguez, E. (2017). Vocalization-silence dynamic patterns: a system for measuring coordination in psychotherapeutic dyadic conversations. *Rev. Latinoam. Psicol.* 49, 48–60. doi: 10.1016/j.rlp.2016.09.004

Tranter, S. E., and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Trans. Audio Speech Lang. Proc.* 14, 1557–1565. doi: 10.1109/TASL.2006.878256

Tschacher, W., Ramseyer, F., and Koole, S. L. (2018). Sharing the now in the social present: duration of nonverbal synchrony is linked with personality. *J. Pers.* 86, 129–138. doi: 10.1111/jopy.12298

Umesh, S., Cohen, L., and Nelson, D. (1999). "Fitting the mel scale," in *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ.

Wallert, J., Gustafson, E., Held, C., Madison, G., Norlund, F., von Essen, L., et al. (2018). Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial. *J. Med. Internet Res.* 20:e10754. doi: 10.2196/10754

Weippert, M., Kumar, M., Kreuzfeld, S., Arndt, D., Rieger, A., and Stoll, R. (2010). Comparison of three mobile devices for measuring R-R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system. *Eur. J. Appl. Physiol.* 109, 779–786. doi: 10.1007/s00421-010-1415-9

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., and Narayanan, S. S. (2015). "Rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS One* 10:e0143055. doi: 10.1371/journal.pone.0143055

Zhang, A., Wang, Q., Zhu, Z., Paisley, J., and Wang, C. (2019). "Fully supervised speaker diarization," in *Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton.

Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., and Shamma, S. (2011). "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, Waikoloa, HI.

Zilcha-Mano, S. (2019). Major developments in methods addressing for whom psychotherapy may work and why. *Psychother. Res.* 29, 693–708. doi: 10.1080/10503307.2018.1429691

Zimmermann, R., Fürer, L., Schenk, N., Koenig, J., Roth, V., Schlüter-Müller, S., et al. (2020). Silence in the psychotherapy of adolescents with borderline personality pathology. *Pers. Disord. Theor. Res. Treat.* doi: 10.1037/per0000402

Zimmermann, R., Krause, M., Weise, S., Schenk, N., Fürer, L., Schrobildgen, C., et al. (2018). A design for process-outcome psychotherapy research in adolescents with borderline personality pathology. *Contemp. Clin. Trials Commun.* 12, 182–191. doi: 10.1016/j.conctc.2018.10.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.