INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

An investigation of therapeutic rapport through prosody in brief psychodynamic psychotherapy

Carolina De Pasquale[1], *Charlie Cullen*[2], *Brian Vaughan*[1]

[1]Technological University Dublin, Ireland
[2]University of the West of Scotland, UK

carolina.depasquale@dit.ie

Abstract

Therapeutic alliance, a concept closely related to rapport, is one of the most important variables in psychotherapy. High degrees of synchrony/coordination in the therapeutic session are considered to contribute to rapport, and have received attention in the psychotherapy literature.

Coordinative behaviours are observable in speech, and they manifest in phenomena such as prosodic accommodation, a dynamic phenomenon closely related to conversational success.

A preliminary investigation of interpersonal prosodic dynamics in psychotherapy was performed on a database obtained in collaboration with the University of Padua, consisting of 16 recordings making up the entire course of a brief psychodynamic psychotherapy intervention for a 25 year old female volunteer and a 41 years old male psychotherapist.

The data was analysed with Time Aligned Moving Averages, a method commonly used in interpersonal speech research. Issues of data sparsity are discussed, and preliminary results on the relationship between empathy and anxiety with interpersonal speech dynamics are presented.

Index Terms: speech prosody, therapeutic alliance, behavioural signal processing, interpersonal synchrony

1. Introduction

In psychotherapy, one of the most important variables that impact therapeutic success is therapeutic alliance, which can not only predict outcome, but also compliance and dropout rates [1].

The role of rapport in psychotherapy has long been investigated in the literature, where it has been shown to play a crucial role in the outcome of the therapy and the development of therapeutic alliance [2]. Further, non-verbal coordination between the individuals involved in the interaction has a strong impact on the development of rapport and alliance, and while it has not been investigated long in the field of clinical psychology research, findings and methodologies developed outside of the clinical field, and particularly in communication science, might help analyse psychotherapeutic interactions.

Drawing on a pilot study [3] that investigated prosodic dynamics in psychiatric interactions, this study expanded the investigation on the co-creation of therapeutic rapport (a concept akin to alliance [4]) with a larger database of psychotherapeutic sessions.

1.0.1. Empathy in psychotherapy

Therapeutic alliance can be defined as the existence of a relationship of mutual confidence and regard between the patient and the therapist, and is affected by therapist characteristics such as empathy, openness, and warmth [4].

High degrees of synchrony in the therapeutic session can be considered a measure of successful relationship. Geerts et al. [5] observed that a higher degree of convergence (that is, two people becoming more similar to each other) between patients and therapists corresponded both to higher patients' satisfaction and lower risk of recurring depression. Similarly, Ramseyer and Tschacher [6] found the amount of synchrony between patient and therapist predicted relationship quality and treatment outcome.Weiste and Peräkylä [7] point out that how therapists address and respond to their client can be as important as the content of the utterance, however, most of the research on therapeutic alliance has focused on posture, gestures, and non verbal non vocal behaviour. The relationship between vocal behaviour and perceived empathy, rapport, and therapy outcome has recently been the subject of interest: Imel et al. [8] found that there was a higher degree of correlation between patients' and therapists mean f_0 when empathy was high, suggesting a high degree of rapport in the session.

1.1. Interaction Dynamics

Individuals engaged in interactions adapt their behaviour to each other for a variety of reasons: [9] argue that high amounts of coordination lead to heightened rapport, which they consider to be the most important element in a successful interaction.

Coordinative behaviours are observable in speech, and they manifest in phenomena such as prosodic convergence and divergence, backchannels, overlaps, and turn taking behaviour [10]. Prosodic accommodation, as one of these coordinative phenomena, is of particular interest due to its similarity to interpersonal synchrony as studied in psychotherapy research: speakers engaged in an interaction have a tendency to adapt pitch and intonation contours, voice intensity levels, speech rate and timing [11, 12], and so on.

Recent studies suggest that accommodation is both linear and dynamic, where some features exhibit a steady linear convergence/divergence dynamic while others fluctuate over the course of one or several interactions [10, 13]: patterns of similarity (synchrony, convergence) and anti-similarity (anti-synchrony, divergence) do not increase during the course of an interaction, but rather seem to be linked to the information structures of the dialogue and appear to reveal moments of engagement.

The goal of this paper is to analyse psychotherapy sessions with time aligned moving windows (TAMA), a commonly used method in speech science literature, and determine whether the method can be used to gather more information on the interpersonal dynamics that take place during the session.

1.2. Methodologies

Due to the multidisciplinarity of the field of interpersonal behaviour study, the literature on methodologies and protocols is varied and the measures on which the analysis is focused are

http://dx.doi.org/10.21437/Interspeech.2019-2551

very diverse. In the studies that focus on speech characteristics as a way to investigate dyadic interactions, rolling windows are often employed to examine the dynamics as they evolve in time during the conversation [10, 14]. The present study was conducted on a database obtained through collaboration, where the audio was recorded with a single microphone (Panasonic RR-US510) placed on a table between the speakers, and the analysis focuses on rolling windows, as it is a method that allows for the investigation of a wide temporal span in the conversation, rather than binding the analysis to adjacent utterances.

1.2.1. TAMA

A common rolling windows method method used in interpersonal speech research is the Time Aligned Moving Average (TAMA) [15]. This method employs a series of overlapping windows used to extract averages and smooth out the features' contours (see fig. 1). In TAMA a window of a certain time duration, which has to be determined according to the characteristics of the conversation, is analysed; a statistic of the prosodic values is calculated for each window, after which the window is moved in such a way that the new time window is mostly overlapping with the old window: this allows for a smoother contour that is still able to capture the dynamics with a high degree of accuracy; the larger the window, the smoother the contour, and the smaller the window, the higher the accuracy.

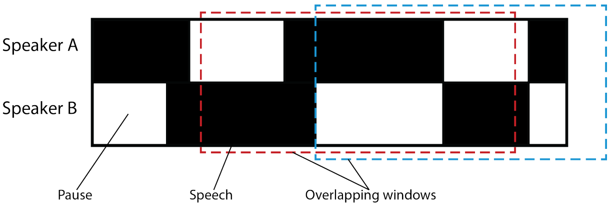

Figure 1: ***Time Aligned Moving Window.*** *The diagram illustrates the TAMA method, showing how the moving windows operate.*

2. Study

The audio was obtained as part of a collaboration with the University of Padua, where it was collected and analysed as part of a doctoral research project [16], with the aim of investigating the role of physiological synchrony in the development of rapport between the client and the therapist, assessed by self report empathy scales.

The database was selected because of the large amount of information available in it; moreover the collaboration with a research group in a psychology department offers domain expertise and theoretical validation of the psychotherapeutic process captured.

2.1. Data Description

The data consists of 16 recordings, each 45 minutes long, making up the entire course of a brief psychodynamic psychotherapy intervention for a total of 720 minutes of raw audio. The therapist is a 41 years old male psychotherapist, and the patient is a 25 year old female volunteer who met the following criteria: neurotic or high-functioning borderline level of personality organization, absence of previous psychiatric symptomatology, no pharmacological treatment; the criteria were assessed by the therapist as part of the routine anamnesis performed upon intake of new clients (see [16] for a detailed explanation).

Each session was video and audio recorded and self-assessment questionnaires were collected during the study: both the therapist and the patient compiled an empathy scale questionnaire, EUS (Barrett-Lennard relationship inventory, empathic understanding subscale), after each session, and the patient also compiled an anxiety questionnaire STAI (State-Trait Anxiety Inventory).

Lastly, word by word timestamped annotations were performed by the researchers.

2.2. Data Issues

Raw acoustic data recorded in naturalistic settings is often not ideal for analysis. The chief issue was the presence of a large amount of extraneous noise and speaker overlap: the audio analysed in this study was recorded on a single microphone, which resulted on both speakers being on the same track.

A cursory examination of the raw audio showed that speaker turns were longer than what would be expected in a standard conversation, with longer silences between turns. Computational analysis of the data confirmed that the speech data in this study was sparser than conversational speech data: these psychotherapy sessions feature very few overlaps and turn changes are often delimited by pauses or silences.

Data sparsity adds complexity to the analysis, as much of the methodology relies on audio chunks from different speakers being adjacent, or close enough to be captured by the same window. When the acoustic data from the two speakers is not adjacent, smaller windows will be incapable of capturing both speakers at the same time, therefore failing to perform a correlation. However, windows that are too large capture too much variation in the individual acoustic features, resulting in non monotonic data that cannot be analysed with standard correlations. Therefore a balance had to be found between accuracy and capturing enough speech from both individuals.

2.2.1. Speaker Separation

Speaker separation was a necessary step before feature extraction to ensure that each signal was processed separately and vocal feature characteristics were analysed individually.

Since word by word time stamped transcriptions of the sessions were available, speakers were separated with a semi-automatic procedure. The semi-automatic separation used the speaker and time information in the annotations to make decisions on turns: the tool searched for the first instance of a speaker (e.g. the patient), stored the relevant time stamp as the turn start, and then kept going through the data until the speaker changed; when a new speaker (e.g. the therapist) was encountered, the time stamp was stored as the start of the second speaker's turn and the first speaker's turn was considered finished.

Overlaps were not clearly marked in the transcriptions, so they were deducted by the time-stamp of words: when two words by different speakers were uttered simultaneously, that was considered an overlap.

Manual corrections had to be made both to the time-stamps and to the separated audio to account for discrepancies between expected results and actual results. This was a necessary step: when comparing results from automatically separated tracks and manually corrected tracks, the number and strength of correlations was much higher in the audio processed through purely automatic methods, which points to artefacts.

2.3. Methodology

2.3.1. Feature Extraction and Smoothing

Features were extracted with the script prosodyShs.config, one of the standard scripts in the openSMILE toolkit. The script returns one measure every 10 milliseconds for fundamental frequency, voicing probability, and vocal energy.

The time series was processed through smoothing windows that compute a smooth trajectory from statistics to reduce granularity. The window size was based on the mean length of speakers's turns: the mean turn length of patient's speech was 26.7 seconds, while the doctor's mean was 12.4 seconds. For the smoothing function, windows of 10 seconds with a 5 seconds step were chosen by rounding down and halving each mean turn length. The features returned by the data windowing were f_0 median, standard deviation, and slope, and intensity median, standard deviation, and slope.

2.3.2. TAMA Based Method

A custom tool was built in Python and used for the analysis. It performed a TAMA based analysis similar to [10], but allowed for greater flexibility with missing data.

To deal with the high amount of gaps, a minimum overlap percentage (set to 30%) was determined for the Spearman to be executed; where the overlap between the two signals was less than the minimum overlap percentage the calculation returned a gap (for an example of the correlations and gaps as a time series see fig. 2.

Based on turn sizes, the size of the correlation window was chosen by rounding up the mean turn lengths and doubling the longest to get at least two in one window: the result is a window of 60 seconds and a step of 15 (which captures 12 values from the smoothing window). The correlations are calculated on the smooth trajectories yielded by the smoothing function; each feature returned by the windowing function was processed through the rolling Spearman window, where client's prosodic values were correlated to therapist's prosodic values to investigate adaptation throughout the conversation.

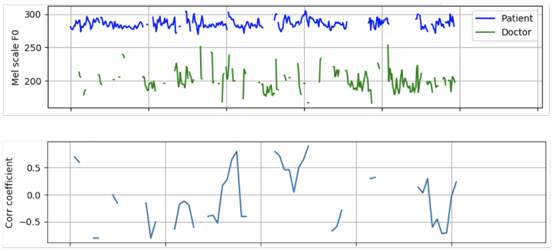

Figure 2: ***A sample session.*** *The graph clearly shows large gaps in the correlation data (bottom), which are due to the long turns and relatively few turn changes (top)*

2.3.3. Fake versus real conversations

In order to determine whether the observed dynamics in acoustic data were meaningful as opposed to a coincidental phenomena, fake conversations were created by shuffling each speaker's data independently and re-pairing the speakers after shuffling: this created interactions that never actually took place to be used as baseline for between speaker dynamics. This procedure is widely used in the literature to test for the presence of real dyadic dynamics as opposed to spurious manifestations of synchrony [17, 10].

3. Results

The result showed clear dynamic patterns, with various gaps throughout each conversation. Some conversations have more gaps than data, as the turns rarely overlap even with windowed smoothing. In agreement with the literature on prosodic dynamics in conversations (see section 1.1), the data indicates that psychotherapy conversations follow a dynamic accommodation pattern, where moments of high accommodation are followed by moments of low accommodation throughout the interaction, as opposed to it being a monotonic manifestation.

3.1. Description of TAMA results

From a visual inspection, it was evident that each different feature statistic (f_0 and intensity medians, standard deviations, and slopes) yielded very different numbers of correlation coefficients, with very different scores.

As per [10], only significant correlation coefficients were selected among the results of the TAMA analysis, with the significance threshold set at $p < 0.05$. An ANOVA was used to examine whether different feature smoothing methods on different features (e.g. f_0 median, median loudness, f_0 slope and so forth; henceforth referred to as 'group') would impact the amount of significant coefficients yielded by the Spearman rolling windows used in TAMA (see table 1). The proportion of significant coefficients was obtained for each session in each group, and the group differences in proportions were investigated; this proportion is going to be referred to as 'general'. Overall the model is significant ($F(5, 72) = 35.215, p < 0.001$), which indicates that there is a significant difference in how many significant coefficients are yielded by different feature group statistics. The difference between the amount of coefficients obtained by using data smoothed with f_0 median windowing and f_0 standard deviation windowing is not significant ($p > 0.05$), while all other groups approach $p = 0$. The model also shows that there is a higher proportion of significant correlation coefficients in relation to intensity features, and in the analysis of slope for both f_0 and intensity.

A similar analysis was performed to investigate whether certain groups yielded more positive correlations than others. The proportion of positive coefficients over total significant coefficients was obtained for each session within each group, and tested with an ANOVA. The overall model is significant again ($F(5, 71) = 11.317, p < 0.001$), indicating a difference in positive coefficient proportions in different groups. The difference in proportion of positive correlation coefficients obtained on data smoothed with f_0 median and standard deviation is not significant, reinforcing the finding obtained on the general proportion; however all other intergroup differences approach $p = 0$. The model also shows that, except for f_0 median and standard deviation, all other features yield significantly more negative than positive correlations.

Table 1: *Significant coefficients ANOVA results*

	coef	**P>|t|**
f_0 med	0.0398	0.168
f_0 slope	0.1789	0.000
f_0 std	-0.0006	0.988
Intensity med	0.4300	0.000
Intensity slope	0.3074	0.000
Intensity std	0.1921	0.000

3.1.1. Fake versus real conversations

Statistically significant coefficients were selected from the spurious conversations with the same methodology used for the real conversations (see section 3.1)

The fake conversation data was then compared to the real conversations, and the difference was tested with a t-test: the two groups are statistically different, with a p value approaching zero ($p < 0.001$), suggesting that the dyadic dynamics observed in the data are not spurious.

3.2. Prosodic dynamics and Empathy/Anxiety scores

Each of the group proportions was correlated with the patient's empathy scores, the patient's anxiety scores, and the therapist's empathy scores, to test whether certain features are more indicative of the co-creation of rapport.

Correlations performed with f_0 features showed stronger correlations with empathy and anxiety scores. There was a significant positive correlation between the proportion of significant coefficients in data windowed with f_0 standard deviation and client's perception of therapist's empathy ($rho = 0.5, p < 0.05$), and a strong negative correlation with client's anxiety scores ($rho = -0.7, p = 0.01$).

Correlations obtained with f_0 slope and client's empathy score and intensity slope and anxiety score approached significance but did not reach it: there was a positive relationship between the proportion of coefficients obtained for intensity slope and anxiety scores, and a negative relationship between proportion of coefficients obtained for f_0 slope and client's empathy score.

4. Discussion

This study consists of a preliminary analysis performed on a database obtained through a collaboration with the University of Padua, using a variety of speech and behavioural signal analysis tools as suggested by the literature.

Automatic sound source separation was not entirely successful in separating the noisy one channel audio, but the diarization was successfully performed through existing annotations, with some manual adjustments due to time stamp imprecisions. This means that any database of audio/video data with time stamped annotations can be successfully processed through semi-automatic methods even when the audio quality does not lend itself to automatic separation. The ability to successfully perform sound source separation on existing databases without the need for time-consuming manual annotations is promising for collaborations across different domains.

Various approaches to the analysis of the interaction patterns between speakers showed that existing tools and methodologies could be inadequate for the investigation of interactive speech behaviours in psychotherapeutic sessions. In fact, the long speaking turns, the long silences, and the relatively infrequent turn switches pose a considerable obstacle to the implementation of common methods in the literature.

Long turns and silences are a characteristic of psychotherapy, therefore existing methodologies had to be tweaked to allow for gap handling within the data: particularly, it is important to note that the gaps are not missing data to be imputed or dropped, as they are a natural by-product of the conversational dynamic. Even so, moving windows can be inadequate to capture the session-long dynamics: long speaking turns make it extremely difficult to capture speech values from both speakers within a single window, therefore constraining the analysis to turn-adjacent windows, where both speakers have values.

Despite the issues discussed, the analysis captured some of the interaction dynamics, and suggest that psychotherapeutic interactions follow non linear dynamics much like other conversations do [10], with moments of high and low accommodation occurring throughout the session.

When analysing the difference in proportion of significant correlation coefficients among different groups, f_0 median and standard deviation exhibited the least amount of coefficients, suggesting that different feature selections and different methods for data windowing can drastically change the results of the analysis. There was also a significant difference in proportion of positive coefficients between groups, with f_0 median and standard deviation having the highest proportion of positive coefficients (i.e. they yielded the least amount of significant coefficients overall, but the highest ratio of positive to negative coefficients).

However, the proportion of significant correlation coefficients obtained through TAMA analysis of data windowed using f_0 statistics had a stronger relationship with empathy and anxiety score than the equivalent proportion on intensity measures. This suggests that accommodation dynamics as captured by a rolling window analysis of f_0 related measures can be indicative of interpersonal rapport in psychotherapy.

5. Conclusion and future work

The main goal of this paper was to determine whether commonly used methods for the analysis of speech acoustic accommodation could be successfully employed for the analysis of audio collected during naturalistic psychotherapeutic interactions recorded in the therapist's own office. It is important to determine the viability of using standard speech analysis methodologies on audio collected during therapy because several databases of this kind exist, and could be an interesting avenue of investigation.

The study aimed to determine whether TAMA, a methodology commonly found in the literature for the assessment of dyadic interactions through speech analysis, could be easily adapted to the study of psychotherapeutic interactions. Further, it is a preliminary investigation of dynamic coordination in therapeutic interactions.

Several issues arose during the study: firstly, sound source separation had to be performed through annotations, rather than through automatic diarization software, due to the noisy quality of the audio; secondly, the large number and length of gaps had to be accounted for in the analysis; lastly, appropriate window lengths had to be determined to guarantee that both speakers would be captured in the same correlation window.

The analysis captured some dynamic interaction patterns, suggesting that psychotherapy interactions follow interpersonal adaptation patterns that are similar to more standard conversations. Moreover, the amount of significant correlations yielded by the TAMA analysis showed a relationship with empathy and anxiety scores of the patient, indicating that there is reason to believe that prosodic interpersonal dynamics have an effect on the co-creation of rapport and the mitigation of client's anxiety. However, a TAMA based analysis seems inadequate in accurately capturing interpersonal dynamics in psychotherapy, due to the long speaking turns and relatively few turn changes, which result in large gaps in the analysis. For this reason, future work should focus on alternative methods to investigate interpersonal dynamics in psychotherapy, such as turn based analysis, and rhythm related features such as speech rate, silence, and pauses.

6. Acknowledgements

The authors would like to thank Regan Ellis from University of West Virginia for her invaluable help with the scripts, and Johann Kleinbub from University of Padua for sharing the data. This research is funded by the Fiosraigh Dean of Graduate Research School Award, from Technological University Dublin.

7. References

[1] L. W. Samstag, S. T. Batchelder, J. C. Muran, J. D. Safran, and A. Winston, “Early identification of treatment failures in short-term psychotherapy. An assessment of therapeutic alliance and interpersonal behavior.” *The Journal of psychotherapy practice and research*, vol. 7, no. 2, pp. 126–43, 1998.

[2] D. C. Mohr, “Negative Outcome in Psychotherapy: A Critical Review,” *Clinical Psychology: Science and Practice*, vol. 2, no. 1, pp. 1–27, mar 1995.

[3] B. Vaughan, C. De Pasquale, L. Wilson, C. Cullen, and B. Lawlor, “Investigating Prosodic Accommodation in Clinical Interviews with Depressed Patients,” in *Proceedings of 7th International MindCare Workshop on Pervasive Computing Paradigms for Mental Health*, in press ed., Boston, USA, 2018, pp. 150–159.

[4] J. B. Nienhuis, J. Owen, J. C. Valentine, S. Winkeljohn Black, T. C. Halford, S. E. Parazak, S. Budge, and M. Hilsenroth, “Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: A meta-analytic review,” *Psychotherapy Research*, pp. 1–13, jul 2016.

[5] E. Geerts, T. Van Os, J. Ormel, and N. Bouhuys, “Nonverbal behavioral similarity between patients with depression in remission and interviewers in relation to satisfaction and recurrence of depression,” *Depression and Anxiety*, vol. 23, no. 4, pp. 200–209, 2006.

[6] F. Ramseyer and W. Tschacher, “Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome.” *Journal of Consulting and Clinical Psychology*, vol. 79, no. 3, pp. 284–295, 2011.

[7] E. Weiste and A. Peräkylä, “Prosody and empathic communication in psychotherapy interaction,” *Psychotherapy Research*, vol. 24, no. 6, pp. 687–701, nov 2014.

[8] Z. E. Imel, J. S. Barco, H. J. Brown, B. R. Baucom, J. S. Baer, J. C. Kircher, and D. C. Atkins, “The association of therapist empathy and synchrony in vocally encoded arousal.” *Journal of Counseling Psychology*, vol. 61, no. 1, pp. 146–153, feb 2014.

[9] L. Tickle-Degnen and R. Rosenthal, “The Nature of Rapport and Its Nonverbal Correlates,” *Psychological Inquiry*, vol. 1, no. 4, pp. pp. 285–293, 1990.

[10] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, “Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction,” *Speech Communication*, vol. 58, pp. 11–34, mar 2014.

[11] J. Edlund, M. Heldner, and J. Hirschberg, “Pause and gap length in face-to-face interaction,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2779–2782, 2009.

[12] C. De Looze and S. Rauzy, “Measuring speakers’ similarity in speech by means of prosodic cues: Methods and potential,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 1393–1396.

[13] S. Mukherjee, A. D’Ausilio, N. Nguyen, L. Fadiga, and L. Badino, “The Relationship Between F0 Synchrony and Speech Convergence in Dyadic Interaction,” in *Interspeech 2017*, no. August. ISCA: ISCA, aug 2017, pp. 2341–2345.

[14] R. V. Palumbo, M. E. Marraccini, L. L. Weyandt, O. Wilder-Smith, H. A. McGee, S. Liu, and M. S. Goodwin, “Interpersonal Autonomic Physiology: A Systematic Review of the Literature,” *Personality and Social Psychology Review*, vol. 1, pp. 1–43, feb 2016.

[15] S. Kousidis, D. Dorran, C. McDonell, and E. Coyle, “Time Series Analysis of Acoustic Feature Convergence in Human Dialogues,” in *Specom 2009*, St. Petersburg, Russian Federation, 2009, pp. 1–6.

[16] J. R. Kleinbub, “THE RHYTHM OF THERAPY: PSYCHOPHYSIOLOGICAL SYNCHRONIZATION IN CLINICAL DYADS,” Doctoral, University of Padova, 2016.

[17] F. Ramseyer and W. Tschacher, “Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome.” *Journal of Consulting and Clinical Psychology*, vol. 79, no. 3, pp. 284–295, 2011.