INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Study on Prosodic Entrainment in Relation to Therapist Empathy in Counseling Conversation

Dehua Tao[1], *Tan Lee*[1], *Harold Chui*[2], *Sarah Luk*[2]

[1] Department of Electronic Engineering [2] Department of Educational Psychology
The Chinese University of Hong Kong

dhtao@link.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk, {haroldchui, sarah_luk}@cuhk.edu.hk

Abstract

Counseling is carried out as spoken conversation between a therapist and a client. The empathy level expressed by the therapist is considered an important index of the quality of counseling and often assessed by an observer or the client. This research investigates the entrainment of speech prosody in relation to subjectively rated empathy. Experimental results show that the entrainment of intensity is more influential to empathy observation than that of pitch or speech rate in client-therapist interaction. The observer and the client have different perceptions of therapist empathy with the same entrained phenomena in pitch and intensity. The client's intention to make adjustment on pitch variation and intensity of speech is considered an indicator of the client's perception of counseling quality.

Index Terms: counseling conversation, therapist empathy, prosodic entrainment, synchrony

1. Introduction

Counseling typically takes the form of spoken conversation between a therapist and a client. It aims to help the client to express thoughts and feelings, lower psychological distress, and make changes in life. In the field of psychotherapy, empathy is a natural human ability described as "the therapist's sensitive ability and willingness to understand the client's thoughts, feelings, and struggles from the client's point of view" [1]. Therapist empathy is considered an essential indicator of counseling outcomes [2, 3, 4]. The empathy level expressed by the therapist over the course of a counseling conversation is often evaluated either by the observer (expressed empathy) or the client (received empathy).

Entrainment is an important aspect of human-human communication, defined as the similarity of communicative behaviors between speakers [5]. This phenomenon correlates with the communicative success of interactive dialogue by achieving mutual comprehension and understanding [6, 7, 8], and building rapport and reducing social distance [9, 10, 11]. The relationship between entrainment and empathy has been studied attentively [12, 13, 14]. In particular, prosodic entrainment between therapist and client was found to be related to the therapist empathy [15, 16, 17, 18]. In the present study, we analyze the prosodic entrainment in counseling conversations and its relation to the therapist empathy perceived by an observer or the client. As representative acoustic features of speech prosody, pitch, intensity and speech rate are analyzed.

Speech entrainment tends to occur as a dynamic phenomenon, i.e., the similarity of speech behaviors between interlocutors may increase or decrease over the course of a conversation [5, 9, 19]. Considering such dynamicity, we propose to divide a conversation into overlapped sections based on speaker turns. A speaker turn refers to the time period during which only one person speaks. A conversation section is defined as a sequence of N consecutive speaker turns. The synchrony of prosody and the degree of entrainment at the section level are measured. Here synchrony refers to the similarity in the variation of prosodic features between two speakers in terms of the direction and magnitude of change [5, 19]. For example, as one speaker raises their pitch, the interlocutor does the same.

It is noted that the incidence of entrainment in client-therapist interaction varies with different prosodic cues. Therapists tend to adjust their speech intensity in accordance with that of the clients more often than to adjust pitch or speech rate. Toward the same speech behaviors of the client and therapist in counseling, the observer and the client have different perceptions about how they correlate with therapist empathy. In addition, experimental results suggest that observing how the client responds to the therapist can be used as a complementary cue to help assess the client's perception of counseling quality.

The remainder of the paper is organized as follows: Section 2 describes the counseling conversations and empathy ratings used in this work. Section 3 introduces the extraction of prosodic features and the measurement of prosodic entrainment. Section 4 presents the correlations between entrainment and therapist empathy. Section 5 discusses the results, followed by conclusion in Section 6.

2. Counseling Speech Dataset

The counseling conversations analyzed in this study come from a speech dataset named CUEMPATHY [20]. The audio recordings were collected during the counseling practicum for therapist trainees at a university in Hong Kong. Clients were adults who came to seek reduced-fee counseling services over a range of concerns related to stress, emotion, relationship, self-esteem, personal growth, and career. The dataset consists of 156 counseling conversations from 39 pairs of therapists and clients (39 distinct therapists and 39 distinct clients), i.e., each pair contributes 4 conversations. All therapists and clients are native Cantonese speakers. In a typical conversation, the therapist and the client take turn to speak. Each conversation is about 50 minutes long. On average there are 316 speaker turns in each conversation. The study was approved by the institutional review board, and informed consent was obtained from both the therapists and clients in written form.

Two measures of therapist empathy are analyzed in this work. They are the Therapist Empathy Scale (TES) [21] given by the observer, and the Barrett-Lennard Relationship Inventory (BLRI) [22] given by the client. Another client-rated measure, Session Evaluation Scale (SES) [23] is also included in our analysis. The SES reflects the overall effectiveness of counsel-

10.21437/Interspeech.2023-2498

ing from the client's perspective. Table 1 provides information about the three ratings. For more details, please refer to [20].

Table 1: *Description of three ratings for counseling sessions.*

Measure	Rater	No. of items	Point scale of each item	Range of total score
TES	Observer	9	$1 =$ *not at all* to $7 =$ *extremely*	9 to 63
BLRI	Client	16	$-3 =$ *strongly disagree* to $+3 =$ *strongly agree* (0 is not used)	-48 to +48
SES	Client	5	$1 =$ *strongly disagree* to $5 =$ *strongly agree*	5 to 25

Clients were asked to give both BLRI and SES ratings after each counseling session. For TES, eight observer raters with counseling training at master level or above were recruited. As a reliability check, about 40% (62 sessions) of the recorded sessions were rated by two observers. The intra-class coefficient (ICC) based on a mean-rating ($k = 2$), consistency, and two-way random effects model was 0.90. According to the Cicchetti's guidelines[1] [24], this value of ICC means excellent inter-rater reliability.

The audio recording of a counseling conversation is divided into speaker-turn-based audio segments based on manually annotated speaker turns and speech transcriptions (in traditional Chinese characters). Turn-level speech-text alignment is performed as described in [20]. Table 2 gives the summary of 155 counseling conversations used in the following experiments. One conversation was dropped because the BLRI and SES ratings were missing.

Table 2: *Summary of speech content in* 155 *counseling conversations. Average speech time per conversation (AvgST), average duration per turn (AvgDur), and average number of characters per turn (AvgChar) are given for each speaker.*

Speaker	AvgST (min)	AvgDur (sec)	AvgChar
Therapist (T)	15.04	6.38	26
Client (C)	33.42	16.40	64

3. Method

3.1. Prosodic features

Prosodic features are computed from the speech in each speaker turn. Pitch and intensity values are extracted on short-time frame basis using the openSMILE toolkit [25] on each turn. The median, mean, and standard deviation of frame-level pitch over the turn are computed as the turn-level pitch parameters. The median and mean indicate the overall pitch level of the turn. The standard deviation is used to measure the degree of pitch variation over the turn. The same set of statistical measures is computed for frame-level intensity. The speech rate is measured in terms of the number of characters (syllables) spoken per second. The turn-level speech rate is computed by dividing the number of syllables contained in the turn by the time duration of the turn. The turn-level feature parameters are normalized on speaker basis by subtracting the mean of raw turn-wise parameters from that speaker in the conversation.

3.2. Entrainment measurement

To capture entrainment dynamics during the course of a counseling conversation, the sequence of speaker turns in the conversation is divided into overlapped sections. Each section contains N consecutive turns. The step size between neighboring sections is M. It is assumed that the similarity between the speech behaviors of client and therapist is constant in a section, and it varies from one section to another. The local static entrainment is measured from adjacent turns in each section [5]. Since our focus is to observe how the therapist (T) responds to the client (C) during counseling, speaker order of the turn sequence in a section is denoted as $C, T, C, T, \ldots$. That is, the client's turn is followed by the therapist's turn. It is further assumed that (1) N and M are even numbers, and (2) each section begins with the client's turn (odd indices for the client and even for the therapist). Otherwise, the first and/or last turn of the conversation can be chopped to satisfy the assumptions. In the following paragraphs, we will explain how to analyze the synchrony of prosody and measure the degree of entrainment at section level.

3.2.1. Synchrony of prosody

Section-level synchrony and anti-synchrony are measured by the Spearman correlation. Anti-synchrony refers to the phenomenon that contrasts synchrony. For each section, the Spearman correlation coefficient between two sets of turn-level prosodic parameters, representing the client and the therapist respectively, is computed. A large value of positive correlation indicates strong synchrony as reflected by the prosodic features in the section. A large value of negative correlation reveals there exists a strong anti-synchronous relation between the observed features. Other values of correlation manifest that neither synchrony nor anti-synchrony is exhibited in the interaction between client and therapist. The threshold for significant positive or negative correlations is set to be ±0.5 at a significant level of 0.05 in the experiments.

Borrowing the idea from [9], the ratios of synchronous/anti-synchronous states are calculated to quantify the entrainment of client-therapist interaction during the conversation. The ratio of synchronous state is obtained by (1) merging neighboring synchronous sections; (2) counting the number of turn pairs (i.e., C and T) in merged sections; (3) dividing the number of turn pairs in the synchronous state by the total number of turn pairs in the conversation. The ratio of anti-synchronous state is computed in a similar way. To analyze how empathy ratings are related to states of synchrony/anti-synchrony, the Pearson correlation between each of these two ratios and each type of empathy rating is calculated. Such analysis can reveal the level of synchrony with different prosodic features would have different effects on the empathy perceived by the observer or the client.

3.2.2. Degree of entrainment

The degree of entrainment in each section is measured as suggested in [16], i.e., computing the averaged absolute difference of turn-level prosodic parameters between the client and the therapist. Let $x(i)$ and $x(i+1)$ denote the feature parameters of two consecutive speaker turns that belong to the client and the therapist, respectively. The averaged absolute difference D_x of a section is defined by Eq. (1) below. The mean and standard deviation of section-wise differences are computed over a counseling conversation. The relationship between the mean or standard deviation and each type of empathy rating across counseling conversations is analyzed by the Pearson correlation.

[1] The Cicchetti's guidelines on inter-rater reliability: ICCs < 0.40 mean poor, $0.40 - 0.59$ mean fair, $0.60 - 0.74$ mean good, and > 0.75 mean excellent reliability.

$$D_x = \frac{1}{N/2} \sum_{i=1}^{N/2} |x(2i) - x(2i-1)| \tag{1}$$

4. Results

4.1. States of synchrony/anti-synchrony in counseling

With the assumption that the local entrainment in a section is constant, the choice of section size may have an impact on the result of entrainment analysis. Thus, different section sizes, i.e., $N \in \{20, 30, 40, 50\}$, with a fixed step size of $M = 10$, are attempted in the experiments. The correlation results obtained from different combinations of N and M are reported according to the significant level. For example, the correlation coefficient between the synchronous state in pitch median and the TES score at $N = 40$ is reported in Table 3 because its significant level is higher than at other values of N. This principle also applies to Section 4.2.

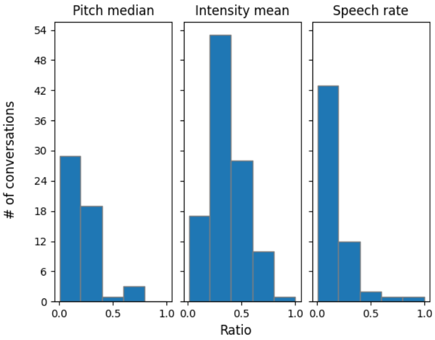

Figure 1: *Ratio of synchronous state in pitch median, intensity mean, and speech rate.*

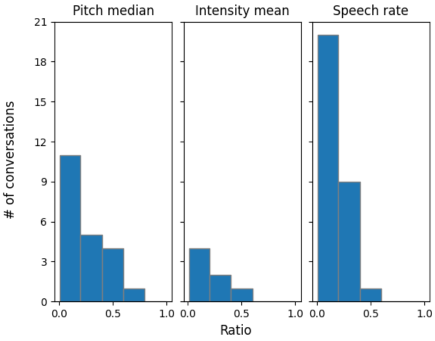

Figure 2: *Ratio of anti-synchronous state in pitch median, intensity mean, and speech rate.*

Taking the pitch median, intensity mean, and speech rate at $N = 40$ as examples, the distributions of ratios of synchronous and anti-synchronous states across all 155 counseling conversations are depicted in Figure 1 and 2 respectively. It is observed that the synchronous state is more common than the anti-synchronous state in client-therapist interaction during counseling. In addition, the occurrence rate of synchronous state is different among different prosodic cues. For example, synchrony may occur more frequently in intensity than pitch or speech rate.

Table 3 shows the results of Pearson correlation analyses between ratios of synchronous/anti-synchronous states and three ratings (i.e., TES, BLRI, and SES) for each of the prosodic parameters. The ratio of synchronous state in pitch level is positively correlated with the observer-rated TES score while negatively correlated with the client-rated BLRI and SES scores. This indicates that when the client raises (or lowers) the pitch level, the observer tends to perceive it as an empathic behavior if the therapist also raises (or lowers) the pitch level. However, the client may feel that the therapist is not expressing empathy. Neither synchrony nor anti-synchrony in pitch variation shows a correlation with empathy ratings. In terms of the intensity level, the ratio of synchronous state is negatively correlated with all three ratings. Thus, as the client speaks progressively louder (or softer), it would be a positive sign of empathy if the therapist decreases (or increases) the volume. For intensity variation, the synchrony in client-therapist interaction may be a negative sign for observer-rated and client-rated empathy. In addition, it would be a non-empathic behavior from the observer's perspective (i.e., TES) if the therapist adjusts their speech rate in accordance with that of the client.

Table 3: *Correlations between ratios of synchrony/anti-synchrony and three ratings. Significant level:* * *for* $p < .1$, ** *for* $p < .05$, *and* *** *for* $p < .01$. *"N.S." refers to "Not Significant".*

Feature	Synchrony			Anti-synchrony		
	TES	BLRI	SES	TES	BLRI	SES
Pitch Median	0.232^{***}	-0.240^{***}	-0.138^{*}	N.S.	N.S.	N.S.
Pitch Mean	N.S.	-0.198^{**}	N.S.	N.S.	N.S.	N.S.
Pitch Std	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.
Intensity Median	N.S.	-0.159^{**}	-0.133^{*}	0.172^{**}	N.S.	N.S.
Intensity Mean	-0.197^{**}	N.S.	N.S.	0.166^{**}	0.133^{*}	N.S.
Intensity Std	-0.185^{**}	N.S.	N.S.	N.S.	0.326^{***}	0.251^{***}
Speech Rate	N.S.	N.S.	N.S.	-0.137^{*}	N.S.	N.S.

4.2. Degree of entrainment in relation to empathy

In order to investigate if section-wise averaged absolute differences increase or decrease continuously over the course of a counseling conversation, the Pearson correlation coefficient between section-wise differences and section time is computed. The significant level is set at 0.05. Figure 3 illustrates the percentages of conversations that show significant positive (i.e., convergence) or negative (i.e., divergence) correlations for different prosodic features at $N = 40$. The section-wise differences are found to be convergent or divergent in only 37.4% (58 conversations), 45.2% (70 conversations), and 36.1% (56 conversations) of total 155 conversations for pitch mean, intensity mean, and speech rate, respectively. This confirms that the prosodic entrainment often dynamically evolves in most counseling conversations.

The mean and standard deviation of section-wise differences are calculated to represent the degree and the fluctuation of entrainment over the course of a conversation, respectively. The relationship between each of the two statistics and each of the three empathy ratings is analyzed through the Pearson correlation. The significant correlation coefficients are reported in Table 4. Both the mean and standard deviation of section-wise

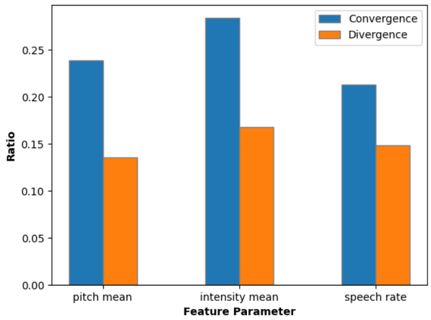

Figure 3: *Ratios of conversations with section-wise differences showing the trend of convergence or divergence for different prosodic features.*

differences in pitch parameters positively correlate with the SES score given by the client. This suggests that the lower degree and greater fluctuation in pitch entrainment in client-therapist interaction may be more helpful in improving the counseling quality from the client's perspective. However, the standard deviation of section-wise differences in pitch level is found to be negatively correlated with the TES score given by the observer. Thus it may be considered a positive sign of observer-rated empathy that the degree of entrainment in pitch level remains stationary throughout a conversation. In addition, it is noted that higher entrainment in intensity level between the client and the therapist is associated with higher observer-rated empathy but lower client-rated empathy. The degree of entrainment in intensity variation is negatively correlated with the client ratings. Similar results are found in the standard deviation of intensity differences. For speech rate, the greater fluctuation of entrainment is found to be associated with higher empathy from the observer's perspective.

Table 4: *Correlations between mean and standard deviation of section-wise differences and three ratings. Significant level:* * *for* $p < .1$, ** *for* $p < .05$, *and* *** *for* $p < .01$. *"N.S." refers to "Not Significant".*

Feature	Mean			Standard Deviation		
	TES	BLRI	SES	TES	BLRI	SES
Pitch Median	N.S.	N.S.	0.174^{**}	-0.195^{**}	N.S.	0.156^{*}
Pitch Mean	N.S.	N.S.	0.164^{**}	-0.186^{**}	N.S.	0.215^{***}
Pitch Std	N.S.	N.S.	0.133^{*}	N.S.	N.S.	0.178^{**}
Intensity Median	-0.154^{*}	0.146^{*}	0.168^{**}	-0.235^{***}	N.S.	N.S.
Intensity Mean	-0.139^{*}	0.181^{**}	0.191^{**}	-0.234^{***}	0.203^{**}	0.211^{***}
Intensity Std	N.S.	0.228^{***}	0.262^{***}	N.S.	0.218^{***}	0.224^{***}
Speech Rate	N.S.	N.S.	N.S.	0.233^{***}	N.S.	N.S.

5. Discussion

Experimental results show that the observer and the client may have different perceptions of the same speech behaviors in relation to therapist empathy in counseling. For example, when the pitch levels of the client and the therapist change in a synchronous manner during a conversation, the observer tends to see this as a positive sign of empathy, while the client does not. The higher entrainment in intensity level between the client and the therapist is associated with higher observer-rated empathy but lower client-rated empathy. In addition, the correlation analyses among TES, BLRI, and SES scores show that there is no significant correlation between the observer-rated TES and the client-rated BLRI ($r = -0.07$) or SES ($r = -0.07$). However, a strong positive correlation exists between BLRI and SES ($r = 0.72$). The discrepancy between observer and client ratings reveals that different raters may focus on different aspects of counseling when making their rating decisions. This issue has also been mentioned in previous studies [4, 26].

To investigate whether the way the client reacts to the therapist in terms of prosodic features is related to empathy ratings, prosodic synchrony analyses are conducted in the context of the turn order of $T, C, T, C, ...$ in a section, i.e., the therapist's turn is followed by the client's turn. There is no significant correlation between synchrony or anti-synchrony and observer-rated empathy. This suggests that the observer tends to focus on how the therapist responds to the client when evaluating the empathy level. However, synchrony in pitch variation is found to be positively correlated with client-rated BLRI and SES scores. This correlation is not found in client-therapist interaction. In addition, synchrony in intensity level is also negatively correlated with BLRI scores, consistent with the result reported in Table 3. This indicates that observing how the client interacts with the therapist can also help assess the client's perceptions of counseling quality.

As mentioned in Section 4.1, the results reported in Table 3 and 4 are obtained from different combinations of $N \in \{20, 30, 40, 50\}$ and $M = 10$. It has been found that the section size affects the correlation analysis of prosodic synchrony. For some feature parameters, significant correlations between states of synchrony or anti-synchrony and empathy ratings are found only for a specific value of N. However, the section size has little effect on measuring the degree of prosodic entrainment because the absolute difference is averaged over C-T turn pairs in a section as shown in Eq. (1).

6. Conclusion

In this work, we investigate the prosodic entrainment between the client and the therapist and its relation to therapist empathy in the counseling conversation. Synchrony of prosody and the degree of entrainment are measured on turn pairs of client-therapist at the section level. It is observed that the occurrence rate of synchrony in client-therapist interaction is varied for different prosodic features. The synchronous state in intensity is more often exhibited than that in pitch or speech rate during counseling. Experimental results indicate that the observer and client may give opposite ratings when observing the same entrainment behaviors in the conversation, including synchrony in pitch level and high entrainment in intensity level. In addition, it is found that the way the client responds to the therapist in terms of pitch variation and intensity level also reflects the client's perception of counseling quality.

7. Acknowledgements

This research is partially supported by the Sustainable Research Fund of the Chinese University of Hong Kong (CUHK) and an ECS grant from the Hong Kong Research Grants Council (Ref.: 24604317).

8. References

[1] C. R. Rogers, *A way of being*. Houghton Mifflin Harcourt, 1995.

[2] R. Elliott, A. C. Bohart, J. C. Watson, and D. Murphy, "Therapist empathy and client outcome: An updated meta-analysis," *Psychotherapy*, vol. 55, no. 4, p. 399, 2018.

[3] T. B. Moyers and W. R. Miller, "Is low therapist empathy toxic?" *Psychology of Addictive Behaviors*, vol. 27, no. 3, p. 878, 2013.

[4] R. Elliott, A. C. Bohart, J. C. Watson, and L. S. Greenberg, "Empathy," *Psychotherapy*, vol. 48, no. 1, p. 43, 2011.

[5] C. J. Wynn and S. A. Borrie, "Classifying conversational entrainment of speech behavior: An expanded framework and review," *Journal of Phonetics*, vol. 94, p. 101173, 2022.

[6] M. J. Pickering and S. Garrod, "Alignment as the basis for successful communication," *Research on Language and Computation*, vol. 4, pp. 203–228, 2006.

[7] S. Garrod and M. J. Pickering, "Why is conversation so easy?" *Trends in Cognitive Sciences*, vol. 8, no. 1, pp. 8–11, 2004.

[8] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[9] 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, 2014.

[10] L. K. Miles, L. K. Nind, and C. N. Macrae, "The rhythm of rapport: Interpersonal synchrony and social perception," *Journal of Experimental Social Psychology*, vol. 45, no. 3, pp. 585–589, 2009.

[11] J. L. Lakin and T. L. Chartrand, "Using nonconscious behavioral mimicry to create affiliation and rapport," *Psychological Science*, vol. 14, no. 4, pp. 334–339, 2003.

[12] T. G. Arizmendi, "Linking mechanisms: Emotional contagion, empathy, and imagery," *Psychoanalytic Psychology*, vol. 28, no. 3, p. 405, 2011.

[13] J. Decety and P. L. Jackson, "The functional architecture of human empathy," *Behavioral and Cognitive Neuroscience Reviews*, vol. 3, no. 2, pp. 71–100, 2004.

[14] S. D. Preston and F. B. De Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and Brain Sciences*, vol. 25, no. 1, pp. 1–20, 2002.

[15] D. Schoenherr, B. Strauss, U. Stangier, and U. Altmann, "The influence of vocal synchrony on outcome and attachment anxiety/avoidance in treatments of social anxiety disorder." *Psychotherapy*, vol. 58, no. 4, pp. 510–522, 2021.

[16] B. Xiao, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Proc. Interspeech*, 2015, pp. 2489–2493.

[17] 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, p. 146, 2014.

[18] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling," in *Proc. Interspeech*, 2013, pp. 2861–2865.

[19] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. Interspeech*, 2011, pp. 3081–3084.

[20] D. Tao, H. Chui, S. Luk, and T. Lee, "Cuempathy: A counseling speech dataset for psychotherapy research," in *Proc. ISCSLP*, 2022, pp. 354–358.

[21] S. E. Decker, C. Nich, K. M. Carroll, and S. Martino, "Development of the therapist empathy scale," *Behavioural and Cognitive Psychotherapy*, vol. 42, no. 3, pp. 339–354, 2014.

[22] G. T. Barrett-Lennard, *The relationship inventory: A complete resource and guide*. John Wiley & Sons, 2015.

[23] C. E. Hill and I. S. Kellems, "Development and use of the helping skills measure to assess client perceptions of the effects of training and of helping skills in sessions," *Journal of Counseling Psychology*, vol. 49, no. 2, p. 264, 2002.

[24] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychological Assessment*, vol. 6, no. 4, p. 284, 1994.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010, pp. 1459–1462.

[26] A. S. Gurman, "The patient's perception of the therapeutic relationship," *Effective Psychotherapy: A Handbook of Research*, pp. 503–543, 1977.