Modeling Therapist Empathy through Prosody in Drug Addiction Counseling

Bo Xiao[1], *Daniel Bone*[1], *Maarten Van Segbroeck*[1], *Zac E. Imel*[2],
David C. Atkins[3], *Panayiotis G. Georgiou*[1], *Shrikanth S. Narayanan*[1]

[1]SAIL, Dept. Electrical Engineering, University of Southern California, U.S.A.
[2]Dept. Educational Psychology, University of Utah, U.S.A.
[3]Dept. Psychiatry & Behavioral Sciences, University of Washington, U.S.A.
[1]http://sail.usc.edu [2]zac.imel@utah.edu [3]datkins@u.washington.edu

Abstract

Empathy measures the capacity of the therapist to experience the same cognitive and emotional dispositions as the patient, and is a key quality factor in counseling. In this work we build computational models to infer the empathy of therapist using prosodic cues. We extract pitch, energy, jitter, shimmer and utterance duration from the speech signal, and normalize and quantize these features in order to estimate the distribution of certain prosodic patterns during each interaction. We find significant correlation between empathy and the distribution of prosodic patterns, and achieve 75% accuracy in classifying therapist empathy levels using this distribution. Experiment results suggest high pitch and energy of the therapist are negatively correlated with empathy. These observations agree with domain literature and human intuition.

Index Terms: Empathy; Prosody; Quantization; Motivational Interviewing; Behavioral Signal Processing

1. Introduction

Empathy is an evolutionarily acquired basic human ability, and is also evident across the phylogenetic tree (*e.g.,* rodents, apes) [1]. In recent years, empathy has been extensively studied in multiple disciplines including biology, neuroscience and psychology. Although the term *empathy* has been used for subtly different phenomena in various disciplines, presence of empathy generally encompasses (1) our internalization of another's thoughts and feelings (taking the perspective of others), and (2) our response with the sensitivity and care appropriate to the suffering of another (feeling for the other) [2]. Neuroscientific studies have found the physiological mechanisms on single-cell and neural-system levels that support the cognitive and social constructs of empathy [1, 3, 4]. These studies have established empathy as a core factor for human social behavior.

In psychotherapy research, and particularly drug abuse counseling, empathy by the therapist is considered essential to quality care. Higher ratings of therapist empathy are associated with treatment retention and positive clinical outcomes [5, 6, 7]. However, assessing therapist empathy is not straightforward since empathy is an internal state and highly dependent on the interaction. For example, there are four steps in each "empathy cycle" [8]: (1) client expression of experience (2) therapist empathic resonation (3) therapist expressing empathy (4) client perceiving empathy, and continue to (1). In psychotherapy studies, therapist empathy is often quantified by third-party observers (coders), based on cues expressed in multiple behavioral channels of both interlocutors. For example, Regenbogen *et al.* have examined the utility of three behavioral channels (facial expressions, prosody and speech content) towards emotional recognition and response via "neutralizing" one channel and testing the differential effect on empathic responses. The study showed that all three channels contributed to empathic responses [9]. This suggests that an observer may have employed information from the above channels to draw an conclusion of the therapist's empathy. Still, this process of empathy evaluation is challenging and non-scalable; computational methods may provide a useful alternative.

Within the emerging field of Behavioral Signal Processing (BSP) [10], we aim to build on previous works and provide computational models to infer therapist empathy based on observed signals. While our overall vision is to employ multimodal behavioral cues to predict empathy in real-time, the current work only employs acoustic cues to predict session level empathy ratings. In our previous work, we investigated the lexical information stream through competitive maximum likelihood language models for empathic and non-empathic utterances. We showed that our model's outcomes significantly correlated with manual annotation of therapist empathy [11]. We have also computed vocal similarity between the client and the therapist, finding it had significant correlation with therapist empathy and yielded better than chance classification of "high" and "low" empathy when combined with speaking time features [12]. In the literature, Kumano *et al.* estimated empathy with Naive Bayes models in natural group conversations, using a variety of manually annotated behavioral cues including facial expression, gaze, head gesture, voice activity, and response timing information [13]. Other related studies focused on empathy synthesis, *i.e.,* designing Embodied Computer Agent (ECA) that can simulate human empathic behavior [14, 15].

In this work, we build computational models to analyze the relation of *prosodic* cues and therapist empathy (as perceived by human experts) in drug addiction counseling. Prosody refers to the non-verbal part of speech, such as intonation, volume, and other voice quality factors, which account for "how one says" rather than "what one says". The close relation of prosody and empathy has been mentioned above in [9]. Moreover, neurology studies have showed not only that the production and perception of prosody share the same brain area, but also that this area is related to affective empathy [16]. A recent psychology study found empirically that prosodic continuity (defined as continued intonation/rhythm of the client's preceding turn, and produced with a lower and/or quieter voice and with narrower pitch span) by the therapist points to higher empathy; whereas prosodic disjuncture (therapist evaluated or challenged the client's emotional descriptions and voice was higher and/or louder and the pitch span wider than in the client's previous turn) points to the opposite [17]. Correlation between the therapist's and the client's mean pitch values is higher in high empathy sessions [18].

Thus, past works have proven prosodic cues as indicators of empathy, but have yet to include a robust analysis of prosodic

This work is supported by NIH, NSF and DoD. Special thanks to Mary Francis for her devotion and help in all SAIL research efforts.

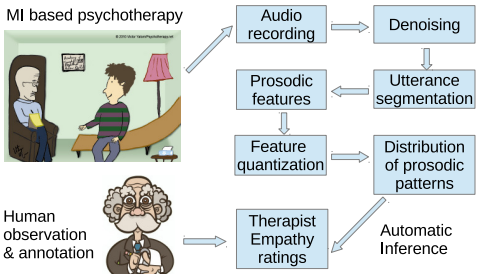

Figure 1: *System level overview of empathy quantification and evaluation process*

feature towards automatic prediction of empathy. Toward this end, in this work we consider five dimensions of prosodic features: pitch, vocal energy, jitter, shimmer, and utterance duration (a result of conversational factors and speaking rate). Pitch and vocal energy are integral to intonation. Jitter and shimmer — measures of short-term variation in pitch period duration and amplitude, respectively — are acoustic correlates of atypical voice quality attributes including breathiness, hoarseness, and roughness [19]. In addition to empathy, these prosodic features can capture important behavioral cues in various domains [20, 21].

We describe the drug abuse counseling dataset and the annotation of therapist empathy in Sec. 2. We explain the prosodic features as well as the extraction and normalization in Sec. 3. For robustness and generalization, we quantize each prosody feature into three intervals, and analyze the values on the unit of speaking utterances. This allows us to characterize the pattern of an utterance with a single or multiple prosodic features, and compute the distribution of various types of utterances in a session, as described in Sec. 4. We examine the relation between these distributions and therapist empathy, and attempt to capture salient prosodic patterns; we then carry out the prediction of "high" or "low" empathy of the therapist using the captured patterns in experiments in Sec. 5. We discuss the results in Sec. 6 and conclude the paper in Sec. 7. A system level overview is illustrated in Fig. 1.

2. Dataset

For the experiments in this work, we use the data from a counselor training study that follows the Motivational Interviewing (MI) counseling approach [22]. MI is a style of counseling focused on helping people to resolve ambivalence and emphasizing the intrinsic motivation of changing addictive behaviors. Therapist empathy is hypothesized to be one of the key drivers of change in patients receiving MI [23]. In the above study, 144 therapists serving in the community participated at the beginning, and 123 of them completed the entire process. Three researchers acted as *Standardized Patients* (SP), *i.e.,* taking the role of clients, in about half of all the counseling sessions recorded. The rest of the sessions involved real clients. Each interaction session is roughly 20 minutes long, recorded with a single channel far field microphone. At collection time the intended consumers were human annotators, and as such the audio quality is challenging for machine processing.

Three human coders reviewed the recordings and assessed the performance of the therapist using a specially designed coding system, the *Motivational Interviewing Treatment Integrity* (MITI) [23]. The therapist in each session received an overall rating of empathy on a Likert scale (discrete) from 1 to 7. Inter-coder reliability, assessed via *Intra-Class Correlation* (ICC), had a mean of 0.67±0.16, while ICC for the same coder over time had a mean of 0.79±0.13. Correlation of the empathy scores given at the first and second time is 0.87, based on all 182 sessions that were coded twice. No session was triple-coded.

In this work we employ 117 sessions that involve a SP and from 91 different therapists, with empathy ratings on the two extremes (take the average rating if the session is coded twice). From the 117 sessions, 71 have high-empathy scores with range 5∼7 and mean 6.05±0.65, while 46 sessions have low-empathy scores with range 1∼3.5 and mean 2.17±0.57. Since only overall ratings of empathy are available rather than localized labels for empathic events, we choose sessions on the extremes where empathic/non-empathic behaviors are more frequent and prominent, and thus binarize our data. The above sessions are manually diarized into therapist's speech and client's speech separately.

3. Prosodic feature extraction

3.1. Audio preprocessing

We first apply speech enhancement to reduce noise in the audio recordings due to the challenging audio quality. We adopt the approach of minimum Mean-Square-Error estimation of spectral amplitude [24] for denoising, implemented in the *Voicebox* speech processing toolbox [25]. The effectiveness of noise reduction was empirically confirmed on a few sessions.

The sessions were manually annotated for speakers; however, the segmentation boundaries were not precisely aligned with speech onsets or offsets, and pauses within the same speaker were not marked out. Therefore, we exploit our previously designed Voice Activity Detection (VAD) system to finely segment the audio into speech utterances [26]. The VAD system is based on a number of robust speech features with Neural Network learning. In this work, we train the model on 10 sessions of Motivational Interviewing which were manually segmented and are disjoint to the data we use for prosody analysis. During decoding the VAD outputs a probability measure for the presence of speech over time with a value that varies between 0 (non-speech) and 1 (speech). We empirically set a high threshold equal to 0.8.

We break a speech segment belonging to a single speaker if a pause inside the segment is longer than 0.2 seconds, otherwise we consider it as a single continuous segment. We also set a threshold for minimum duration of speech segment as 0.5 seconds; therefore detected speech of less than that was assigned as non-speech. No lower bound is set for the gap between speakers due to probable interruptions. However, we ignore speech regions that are labeled as overlapped speech, since they cannot represent the prosodic properties of a single speaker.

We denote the resultant sequence of speech utterances in a session as U_n, $n = 1, 2, \cdots, N$, where N is the total number of utterances. Let $r_n \in \{\texttt{Therapist(T)}, \texttt{Patient(P)}\}$ be the speaker of U_n. Let d_n be the time duration of U_n.

3.2. Pitch and jitter

We compute pitch using the method in [27] that is inspired by the subharmonic summation proposed in [28]. We suppress doubling and halving errors through dynamic programming. Pitch values are confined to the frequency range 50-800 Hz and are computed on a 30 ms window with a 10 ms shift. In order to reduce interference, we compute pitch values separately for the two interlocutors. We further prune the pitch against doubling/halving errors and other noises, respectively for the therapist and patient by the following two steps: (1) Find the central pitch p_0 for the speaker as the mode of the pitch values $p(t)$. (2) Discard the pitch value if $p(t) > 1.5p_0$ or $p(t) < p_0/1.5$ (symmetric in log domain). We observed that in average the pruning removed 6% pitch values in time.

Let $\overline{p_T}$ be the mean pitch after pruning for the therapist in

a session. For each utterance U_n, $r_n = \mathtt{T}$ we obtain the mean-normalized log pitch feature as in (1):

$$p_n = \frac{1}{K}\sum_{t_n=1}^{K} \mathtt{log}\frac{p(t_n)}{\overline{p_\mathtt{T}}}, \tag{1}$$

where t_n is the acoustic frame index within the time span of U_n.

We denote $g(t_n)$ the reciprocal of $p(t_n)$, *i.e.,* the fundamental period of the glottal pulse. Based on extracted pitch values, we approximate relative jitter values $\tilde{j}_n$, *i.e.,* normalized by the average fundamental period, for U_n as in (2)$\sim$(3):

$$\tilde{j}_n = \frac{1}{K-1}\sum_{t_n=2}^{K}\left|\frac{g(t_n)-g(t_n-1)}{\overline{g_\mathtt{T}}}\right| \tag{2}$$

$$= \frac{\overline{p_\mathtt{T}}}{K-1}\sum_{t_n=2}^{K}\left|\frac{1}{p(t_n)}-\frac{1}{p(t_n-1)}\right| \tag{3}$$

Moreover, we compute the averaged relative jitter $\overline{j_\mathtt{T}}$ for the therapist in the entire session (accumulating all therapist utterances) by applying (3), as the individual baseline for jitter. Finally, we define the normalized jitter feature $j_n = \tilde{j}_n - \overline{j_\mathtt{T}}$ for U_n. We obtain the pitch and jitter features for patient utterances in the same way.

3.3. Vocal energy and shimmer

We compute short time vocal energy over a 300 ms window with 10 ms shift as the mean-squared value of speech signal. We denote the log scale of the energy as $e(t)$. Due to the variations of microphone gain and speaker-to-microphone distance, it is necessary to normalize the energy for each interlocutor. Let the mean and variance of the therapist's energy be μ_T and σ_T^2. We define the vocal energy feature e_n for U_n, $r_n = \mathtt{T}$ as in (4):

$$e_n = \frac{1}{K}\sum_{t_n=1}^{K}\frac{e(t_n)-\mu_T}{\sigma_T}, \tag{4}$$

where t_n is the acoustic frame index within the time span of U_n.

We compute the averaged difference of $e(t_n)$ as shimmer value $\tilde{s}_n$ for U_n, as in (5):

$$\tilde{s}_n = \frac{1}{K-1}\sum_{t_n=2}^{K}\left|\frac{e(t_n)-e(t_n-1)}{\sigma_T}\right| \tag{5}$$

Moreover, we compute the averaged shimmer $\overline{s_\mathtt{T}}$ as an individual baseline for the therapist by applying (5) over the accumulated speech signal of the therapist in the entire session. We finally define the normalized shimmer feature as $s_n = \tilde{s}_n - \overline{s_\mathtt{T}}$ for U_n.

We obtain the vocal energy and shimmer features for the patient in a similar way. In summary, $(d_n, p_n, j_n, e_n, s_n)$ is the five-dimensional prosodic feature for U_n.

4. Modeling prosodic features

4.1. Feature quantization

We quantize each prosodic feature into Q equally populated intervals, for the therapist and the patient separately. We find boundaries of the intervals on aggregated training samples of utterances from multiple sessions involving different therapists and patients. Such aggregate quantization is applicable due to the normalization and subtraction of individual baselines. Note that the disparities of feature distributions still exist in different sessions, hence the equally populated quantization does not imply that the quantized features are uniformly distributed in each session. Unseen utterances (test set) can be quantized with the same boundaries obtained on the training set.

Taking $Q = 3$ for the therapist utterances for example, we quantize each feature by its 33 and 67 percentile into discrete values. These discrete bins conceptually represent low, medium and high values for each feature dimension. Similarly we carry out the quantization for patient utterances.

4.2. Distribution of prosodic patterns

We denote the quantized feature values as $(\hat{d}_n, \hat{p}_n, \hat{j}_n, \hat{e}_n, \hat{s}_n)$ for utterance U_n. We compute the joint distributions of $P_U(r_n, F_n)$ and $P_U(r_n, F_n, r_{n+1}, F_{n+1})$, where r_n is binary in Therapist or Patient, *i.e.,* $r_n \in \{\mathtt{T}, \mathtt{P}\}$, and F_n can be any combination drawn from the five quantized prosodic features. Because of speech segmentation and quantization of the feature set, there are integer counts of utterances in each pattern and finite types of prosodic patterns. We count the occurrences of each discrete pattern of (r_n, F_n) and $(r_n, F_n, r_{n+1}, F_{n+1})$, and divide by the total number of segments. The above probabilistic model is akin to a maximum likelihood "bag-of-words" model.

Specifically, we consider the following feature combinations in $P_U(r_n, F_n)$: (1) $F_n = f_n^1$ where f_n^1 is one of the five prosodic features. (2) $F_n = (f_n^1, f_n^2)$ where (f_n^1, f_n^2) is any combination of two features. (3) $F_n = (f_n^1, f_n^2, f_n^3)$ where (f_n^1, f_n^2, f_n^3) is any combination of three features. For $P_U(r_n, F_n, r_{n+1}, F_{n+1})$, we set $F_n = f_n^1$, $F_{n+1} = f_{n+1}^1$, *i.e.,* a single feature out of the five prosodic features. For the robustness of probability estimation, we do not incorporate more complex prosodic patterns (*e.g.,* combination of more features) due to the limit of samples (speech segments) in each session.

We consider the joint rather than conditional probability with respect to the speaker, according to the previous finding that therapist empathy is correlated with the ratio of therapist's speech [12]. The total dimension of different probability entries is given in (6) ($\mathtt{C}_m^n$ represents combinatorial function), which equals 930 in case of $Q = 3$. Note that these probability entries can also be viewed as the frequencies of occurrence for different prosodic patterns; we examine the relation of therapist empathy and these probabilities in the experiments.

$$2(Q\mathtt{C}_5^1 + Q^2\mathtt{C}_5^2 + Q^3\mathtt{C}_5^3) + (2Q)^2\mathtt{C}_5^1 \tag{6}$$

5. Experiment and results

5.1. Correlation of therapist empathy and prosody

For the analysis of correlation between therapist empathy and prosody, we extract prosodic features in each session and derive the quantization of $Q = 3$ as well as sessions-wise distribution P_U over the entire dataset. We will discuss the choice of Q in Sec. 6.

The coded therapist empathy rating E, as introduced in Sec. 2, is in the range of 1 to 7. We compute the Pearson's correlation ρ between E and elements of P_U, and test the significance using Student's t-distribution. In Table 1 we report some of the most prominent prosodic patterns associated positively and negatively with E. We can see that high pitch and energy are negatively associated with therapist empathy; this is consistent with the empirical findings from psychology literature *e.g.,* [17]. We discuss the results further in Sec. 6.

5.2. Classification of therapist empathy level

We carry out leave-one-therapist-out cross-validation in prediction of the binary levels of therapist empathy $\hat{E}$ ($\hat{E} = 1$ if $E \geq 4.5$, otherwise $\hat{E} = 0$) using P_U. This means we do the following operations in each round. For training (1) determine

Table 1: *Prominent prosodic patterns for correlations ρ between E and P_U: T — Therapist, P — Patient, L — Low, M — Medium, H — High*

r_n	f_n^1	f_n^2	f_n^3	ρ	p-value
T	$\hat{d}_n$ =M	$\hat{p}_n$ =H	$\hat{e}_n$ =H	-0.47	8×10^{-8}
T	$\hat{d}_n$ =M	$\hat{p}_n$ =H	—	-0.42	2×10^{-6}
T	$\hat{d}_n$ =M	$\hat{e}_n$ =H	$\hat{s}_n$ =M	-0.41	4×10^{-6}
T	$\hat{d}_n$ =M	$\hat{p}_n$ =H	$\hat{j}_n$ =M	-0.41	5×10^{-6}
...				...	

r_n	f_n^1	r_{n+1}	f_{n+1}^1	ρ	p-value
T	$\hat{e}_n$=M	T	$\hat{e}_{n+1}$=M	-0.40	7×10^{-6}
T	$\hat{j}_n$=M	T	$\hat{j}_{n+1}$=H	-0.34	2×10^{-4}
P	$\hat{d}_n$=H	T	$\hat{d}_{n+1}$=L	0.34	2×10^{-4}
P	$\hat{p}_n$=M	P	$\hat{p}_{n+1}$=L	0.34	2×10^{-4}
...				...	
In total 51 features				$\lvert\rho\rvert > 0.3$	$p < 10^{-3}$

the quantization boundaries of the prosodic features; (2) quantize using these thresholds; (3) compute P_U separately for each session; (4) train the classifier of $\hat{E}$. For testing employ the test data and (1) quantize using thresholds derived at training and compute P_U; (2) predict $\hat{E}$. We use linear Support Vector Machine (SVM) as the classifier.

For comparison, we design a baseline method for classification using functionals of prosodic features $(d_n, p_n, j_n, e_n, s_n)$ in each session, separately for the therapist and the patient utterances. This is hypothesizing that the overall empathy is reflected in the ensemble statistics of individual prosodic features. Specifically, we employ the following functionals: (1) 1, 25, 50, 75, 99 percentile; (2) range of 1∼25, 25∼50, 50∼75, 75∼99 percentile; (3) mean, variance, skewness and kurtosis of the prosodic feature. This in total derives 14 (functional) $\times$ 5 (prosody) $\times$ 2 (speaker) $=$ 140 dimensional functional features for the SVM classifier. Note that the mean value of the prosodic features are not necessarily zero, since the normalization is applied to acoustic frames while the functional is computed over utterances. Numerically, it is equivalent to weighting shorter utterances higher, such that treating an utterance as a basic unit of expression.

We use a simple feature selection scheme to reduce complexity and avoid over-fitting in the classification, by thresholding on the p-value of one-factor ANOVA [29] test (*i.e.,* a test of different mean values in two groups) on the training samples for each feature. We set the threshold to 10^{-3} for P_U, while we loosen the threshold to 10^{-2} for the baseline functionals as we observe that their significances are in general lower.

In Table 2 we list the classification accuracies by the different approaches with the same data and cross-validation method. The P_U features yield the best performance that is higher than chance level (and statistically significant; binomial test $p < 10^{-3}$) and higher than the result in [12] (but not statistically significant). The performance of the baseline method is higher than chance level but not statistically significant. We further discuss the results in Sec. 6.

6. Discussion

An interesting scientific question is whether the prosodic patterns of the therapist can themselves, out of contextualization of the patient behavior, provide important information regarding the therapist empathy. To address this, we compute the conditional distribution $P_U(F_n|r_n = \texttt{T})$. In comparison to the upper half of Table 1, the prominent correlations ($|\rho| \geq 0.3$) between $P_U(F_n|\texttt{T})$ and empathy are listed in Table 3. We can see the

Table 2: *Therapist empathy $\hat{E}$ classification accuracies*

Approach	Accuracy
Chance level	0.61
Vocal similarity and speech ratio [12]	0.70
Distribution of prosodic patterns P_U	0.75
Functionals of prosodic features	0.67

Table 3: *Prominent prosodic patterns for correlations ρ between E and $P_U(F_n|\texttt{T})$: L — Low, M — Medium, H — High*

f_n^1	f_n^2	f_n^3	ρ	p-value
$\hat{d}_n$ =M	$\hat{p}_n$ =H	$\hat{e}_n$ =H	-0.33	2×10^{-4}
$\hat{d}_n$ =L	$\hat{e}_n$ =L	$\hat{s}_n$ =H	0.31	6×10^{-4}
$\hat{e}_n$ =L	—	—	0.30	1×10^{-3}

effect of high energy and high pitch is still negative, but the statistical significance is reduced; similarly for the other therapist prosodic patterns in Table 1. In addition, low energy patterns show positive correlation to empathy, which is consistent with the empirical finding [17].

In Sec. 5.2 we find that the functionals of prosodic features are less effective to infer empathy than the distribution of prosodic patterns. The most significant correlation between the functionals and E is -0.3 by the median of therapist energy. This trend of higher energy implying lower empathy is consistent with the results by P_U; however, it is less discriminative. The quantized prosodic patterns proposed in this work on the other hand, may only focus on part of the interaction. For example, the most significant pattern of (d_n =M, p_n=H, e_n=H) represents only 6% (range 1% to 15%) of therapist utterances in average. This suggests that it is important to study salient behavior patterns for high level summative behavioral characteristics like empathy. Such high level judgments are often a non-trivial integration of local evidences, where some cues may be more important than others. In addition, it may be beneficial to jointly model multiple aspects of behavior (*e.g.,* multiple features from prosody).

The other interest is on the order of quantization Q. We tested the choices of $Q = 2, 4, 5$ in addition to $Q = 3$. In general we observe a similar trend compared to the findings in Sec. 5, however, the significances and accuracies are in general lower than the case of $Q = 3$. We believe that having more quantization bins may cause sparsity, even though fewer bins may reduce the discriminative power of the feature set.

7. Conclusion

In this work we have extracted, quantized and modeled the distribution of prosodic cues in order to infer therapist empathy in motivational interviewing based psychotherapy. We found salient prosodic patterns that are significantly correlated with empathy, which was used to classify "high" and "low" empathy ratings achieving an accuracy of 75%. The results suggest that the use of high energy and pitch by the therapist is a negative sign of empathy. The quantization of prosodic features enabled the capture of salient patterns that led to more accurate inference of high level behavior like empathy, and outperformed the approach based on functionals of prosodic features.

In the future, we aim to validate empirical settings applied in this work on larger-scale data, and in the end automate the parameter adaptation for robust analysis in practical use. For the inference of empathy, it would be useful to jointly model the lexical and prosodic information, in order to have a complete account of both "what they say" and "how they say it".

8. References

[1] S. D. Preston and F. De Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and Brain Sciences*, vol. 25, no. 01, pp. 1–20, 2002.

[2] C. D. Batson, "These things called empathy: eight related but distinct phenomena," *The social neuroscience of empathy*, pp. 3–15, 2009.

[3] M. Iacoboni, "Imitation, empathy, and mirror neurons," *Annual review of psychology*, vol. 60, pp. 653–670, 2009.

[4] N. Eisenberg and N. D. Eggum, "Empathic responding: Sympathy and personal distress," *The social neuroscience of empathy*, pp. 71–83, 2009.

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

[6] W. R. Miller and G. S. Rose, "Toward a theory of motivational interviewing," *American Psychologist*, vol. 64, no. 6, p. 527, 2009.

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

[8] G. T. Barrett-Lennard, "The empathy cycle: Refinement of a nuclear concept," *Journal of Counseling Psychology*, vol. 28, no. 2, p. 91, 1981.

[9] C. Regenbogen, D. A. Schneider, A. Finkelmeyer, N. Kohn, B. Derntl, T. Kellermann, R. E. Gur, F. Schneider, and U. Habel, "The differential contribution of facial expressions, prosody, and speech content to empathy," *Cognition & emotion*, vol. 26, no. 6, pp. 995–1014, 2012.

[10] S. Narayanan and P. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceeding of IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[11] B. Xiao, D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "Analyzing the language of therapist empathy inmotivational interview based psychotherapy," in *APSIPA ASC*, Dec. 2012.

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

[13] S. Kumano, K. Otsuka, M. Matsuda, and J. Yamato, "Analyzing perceived empathy/antipathy based on reaction time in behavioral coordination," in *Automatic Face and Gesture Recognition*. IEEE, 2013, pp. 1–8.

[14] S. W. McQuiggan and J. C. Lester, "Modeling and evaluating empathy in embodied companion agents," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 348–360, 2007.

[15] H. Boukricha, I. Wachsmuth, M. N. Carminati, and P. Knoeferle, "A computational model of empathy: Empirical evaluation," in *Proc. ACII*. IEEE, 2013, pp. 1–6.

[16] L. Aziz-Zadeh, T. Sheng, and A. Gheytanchi, "Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability," *PLoS One*, vol. 5, no. 1, p. e8759, 2010.

[17] E. Weiste and A. Peräkylä, "Prosody and empathic communication in psychotherapy interaction," *Psychotherapy Research*, pp. 1–15, 2014.

[18] 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.

[19] A. McAllister, J. Sundberg, and S. R. Hibi, "Acoustic measurements and perceptual evaluation of hoarseness in children's voices," *Logopedics Phoniatrics Vocology*, vol. 23, 1998.

[20] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, 2014, (in Press).

[21] B. Z. Pollermann, "A place for prosody in a unified model of cognition and emotion," in *Proc. Speech Prosody*, 2002.

[22] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, "Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors," *Journal of substance abuse treatment*, vol. 37, no. 2, p. 191, 2009.

[23] T. Moyers, T. Martin, J. Manuel, and W. Miller, "The motivational interviewing treatment integrity (miti) code: Version 2.0," *University of New Mexico, Center on Alcoholism, Substance Abuse and Addictions. Albuquerque, NM*, 2008.

[24] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[25] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, 1997.

[26] M. Van Segbroeck, A. Tsiartas, and S. S. Narayanan, "A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice," in *Proc. InterSpeech*, Aug. 2013.

[27] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 213–216.

[28] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The journal of the acoustical society of America*, vol. 83, no. 1, pp. 257–264, 1988.

[29] R. V. Hogg and E. A. Tanis, *Probability and Statistical Inference*. Pearson Prentice Hall, 2009, pp. 379–389.