Can We Detect Speakers' Empathy?: A Real-Life Case Study

Firoj Alam, Morena Danieli, Giuseppe Riccardi
Department of Information Engineering and Computer Science,
University of Trento, Italy
Email: {firoj.alam, morena.danieli, giuseppe.riccardi}@unitn.it

***Abstract*—In the context of automatic behavioral analysis, we aim to classify empathy in human-human spoken conversations. Empathy underlies to the human ability to recognize, understand and to react to emotions, attitudes, and beliefs of others. While empathy and its different manifestations (e.g., sympathy, compassion) have been widely studied in psychology, very little has been done in the computational research literature. In this paper, we present a case study where we investigate the occurrences of empathy in call-centers human-human conversations. In order to propose an operational definition of empathy, we adopt the *modal* model of emotions, where the appraisal processes of the unfolding of emotional states are modeled sequentially. We have designed a binary classification system to detect the presence of empathic manifestations in spoken conversations. The automatic classification system has been evaluated using spoken conversations by exploiting and comparing performances of the lexical, acoustic and psycholinguistic features.**

***Keywords*—Empathy, Emotions, Spoken Conversation**

I. Introduction

Large corporations are hosting call centers to support their customers and at the same time monitor calls in real time as well as record them for later review [1]. One of the goals of reviewing the recorded data is to understand customers' affective behavior, which facilitates call center managers or decision makers. A common approach is that human experts listen to and analyze the recorded data, then generate a summary with behavioral descriptions. The main challenge for the human experts is that they need to deal with a large amount of recorded data, which is time-consuming and labor intensive. Moreover, they can only analyze a small random portion of the data, which is much less than 1% [1]. Hence, automated systems are required to process a large amount of data in a timely manner and also to reduce the cost of manual analysis. The analysis by human experts includes preparing a summary with behavioral descriptions of the agent's *empathic* ability and the customer's emotional manifestation. A summary may also include *what has been said*, that is, the content of the conversation, *how it has been said* such as cooperative and competitive turn-taking signals and factual metrics such as waiting time and duration of the overlapping speech. In this paper, we present our study towards classifying empathic behaviors of the call center agents.

Our empathic ability helps us to understand others' emotional states as well as their subsequent behaviors [2]. It may also serve as a critical component of the motivational system underlying altruistic behavior and cooperation. Finally, it can play an important role in socially regulated communication. The signals of human empathic abilities are an essential step to understand the social emotions [2]. In addition, empathy has been shown to play a crucial role in pathologies like autism and Asperger syndrome, and a central part for the success of psychological therap [3]. Currently, it is also an important issue towards the design of communicative effective virtual agents [4], human-robot interaction [5], artificial tutors [6] and call center applications [7]. The automatic recognition of complex social affective states, such as empathy, can also provide a useful insight in the newly emerging field - social cognitive informatics [8].

The design of automatic computational systems for empathy recognition poses several open challenges. First of all, the concept of empathy is still vague, even though there is a wide use of the notion of empathy in the psychological research. Secondly, among psychologists, there is still some uncertainty about the type of signals that supports empathic responses. The third issue is related to the complexity of data collections and annotation schemes for modeling empathic behavior. Last but not least, when the observed data are constituted by spoken interactions, it is necessary to understand the discriminative features of *verbal* and *vocal non-verbal*[1] components in conversations for designing automatic systems. To the best of our knowledge, automatic empathy classification has not been studied yet in the context of real life, i.e., not controlled, conditions.

Hence, the goal of this study was to classify agents' empathy in call center spoken conversations. We investigated the verbal and vocal non-verbal signals of agent's spoken content as it is evident that the conversations convey these cues, which can be exploited for classifying empathy [9]. In order to achieve our goal we defined an annotation scheme, annotated a dyadic customer-agent dialog corpus, and evaluated acoustic, linguistic, and psycholinguistic features for the automatic classification problem. This study would be useful to understand whether an agent can understand customer's emotional states and react appropriately. Understanding customer's emotions by the agent are important to soothe customer's disappointment or dissatisfaction. Subsequently, agents' can be trained to develop their empathic ability.

The paper is organized as follows. In Section II, we discuss the related work relevant to this study. Following that in Section III, we describe the corpus and annotation procedures. We provide the details of the classification experiments, results and

[1] Using the term *vocal non-verbal* we refer to the paralinguistic features in this paper

discussions in Section IV. Finally, we report the concluding remarks in Section V.

II. Related Work

In this study, we first investigated the available psychometric scales, questionnaires or annotation schemes for emotions in general, and for empathy in particular, in order to use them in our study. We found that among psychologists there are some fundamental concerns about the adequacy of the various scales. For example, no significant correlation was found between the scores on empathy scales and the measurement of empathic accuracy [10]. The issue may be related to the fact that questionnaires assume that people have metacognition about their empathic abilities, but that is not always true for all of us. Nowadays, psychologists tend no longer to conceive empathy exclusively either in affective or cognitive terms but as encompassing both. The de-facto standardized tests, such as [11], seem to be effective mostly for clinical applications within well-established experimental settings. However, they can hardly be adapted to judge the empathic abilities of virtual agents and to evaluate our empathic behavior in everyday situations.

In the field of affective computing, researchers have been trying to design emotional intelligent systems that automatically recognize, model and synthesize full spectrum of short and long term states and traits [12], [13] by analyzing paralinguistic phenomena [14], facial expressions, gestures [15] and bio-signals [16]. However, there has been very little work for automatic empathy recognition compared to basic emotional categories or dimensional approach to emotions. A few studies are as follows. Kumano et al. [17] studied four-party meeting conversations to estimate empathy, antipathy and unconcerned emotional interactions utilizing facial expression, gaze, and speech-silence features. Leite et al. [18] attempted to design a social robot with a capability to understand user's affective states and display empathic behaviors. In healthcare domain therapists' conversations has been analyzed to classify empathic and non-empathic utterances [19].

III. Towards An Operational Model

A. Corpus

The corpus includes $1,894$ randomly selected customer-agent conversations, which were collected over the course of six-months, amounting to 210 hours of audio. These conversations were recorded on two separate channels of 16 bits per sample and $8kHz$ sampling rate. The average length of the conversations was 406 seconds.

B. Annotation Scheme for Spoken Conversations

For the annotation of empathy, we followed the psychological definition of Hoffman [20], which states empathy as "*an emotional state triggered by another's emotional state or situation, in which one feels what the other feels or would normally be expected to feel in his situation*". In order to design the operational model of empathy annotation, we adopted the *modal* model of emotion by Gross [21].

In the psychological literature, it has been shown that temporal unfolding of emotional states can be conceptualized and experimentally tested [22]. Gross has provided evidence that concepts such as *emergence* — derivation from the expectations of relationships — and *unfolding* — sequences that persist over time — may help in explaining emotional states. The modal model of emotions [21], [23] emphasizes the attentional and appraisal acts underlying the emotion-arousing process. In Figure 1, we provide the original schema of Gross model. The individuals' core *Attention-Appraisal* processes (included in the box) are affected by the *Situation* that is defined objectively in terms of physical or virtual spaces and objects. The *Situation* compels the *Attention* of the individual; it triggers an *Appraisal* process and gives rise to coordinated and malleable *Responses*. It is important to note that this model is dynamic and the situation may be modified (directed arc from the *Response* to the *Situation*) by the actual value of the *Response* generated by the *Attention-Appraisal* process. In this model, emotional states are seen as a way of experiencing the world: they are distinct functional states [24], and the appraisal acts describe the content of those functional states within a context.

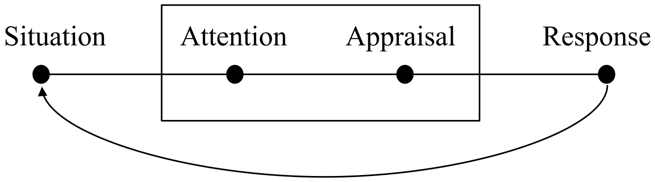

Fig. 1. The modal model of emotion [21].

Therefore, we believe that Gross' model provides a useful framework for describing the dynamics of emotional states within an *affective scene*[2] [25], because not only it does focus on appraisal, but also considers how responses are feeding back to the initial communicative situation.

In order to make it applicable in a real-life domain like the call center conversations, we operationally defined empathy as *"a situation where an agent anticipates or views solutions and clarifications, based on the understanding of a customer's problem or issue, that may help in relieving or preventing the customer's unpleasant feelings"*. For designing the annotation scheme, we have performed an extensive analysis of one hundred conversations (more than 11 hours), and selected dialog turns where the speech signal showed the emergence of both basic emotions, such as anger, and complex emotional states such as frustration and empathy. Our qualitative analysis supported the hypothesis that the relevant speech segments were often characterized by perceivable variations in the speech signal.

As expected, such variations sometimes co-occurred with emotionally connoted words, but also with functional parts of speech, such as adverbs and interjections, which could play the

[2] *"The **affective scene** is an emotional episode where one individual is affected by an emotion-arousing process that (a) generates an emotional state variation, and (b) triggers a behavioral and linguistic response. The affective scene extends from the event triggering the unfolding of emotions on both individuals, throughout the closure event when individuals disengage themselves from their communicative context."* It is defined based on the emotion sequence between interlocutors. For example, the sequence of emotional states between an agent and a customer could be Frustration (C) → Empathy(A) → Satisfaction(C). A - Agent, C-Customer.

role of lexical supports for the variations in emotional states. On the basis of the above observations, we have designed an annotation scheme for empathy by taking into account the perception of the variations in the speech signal as well as variations in the linguistic content of the utterances [26].

The annotation scheme includes the following recommendations for the annotators:

1) Annotating the onset of the signal variations that supports the perception of the manifestation of emotions.
2) Identifying the speech segments preceding and following the onset position.
3) Annotating the context (left of the onset) and target (right of the onset) segments with a label of an emotional state (e.g., frustration, empathy etc.).

The context of the onset is defined to be neutral with respect to the target emotional state label. We have introduced *neutrality* as a relative concept to support annotators in their perception process of empathy while identifying the support of the situational context.

In the annotation process, given the limited resources, our goal was to maximize the number of annotated conversations. For this reason, we annotated only the first occurrence of a segment pair (e.g., neutral-empathy) within each conversation. Once candidate segment pair was selected, the annotators could listen to the speech segments as many times as needed to judge if the selected segment pair could be labeled. After that, the annotator tagged the right of the onset of the segment pair with an emotional label, and left of the onset was labeled as neutral. During the annotation process, annotator also needed to focus on the boundaries of the speech segment.

For our experiment, the annotation task was performed by two expert annotators who worked on non-transcribed spoken conversations by following the annotation scheme reported above. The annotators used the EXMARaLDA Partitur Editor [27] to perform their task. They annotated *Empathy* on the agent channel and *Frustration* and *Anger* on the customer channel. The annotators labeled *Neutral* on the segment that appeared on any emotional segment to define the context, as mentioned earlier. The average Kappa statistics [28] for the empathy annotated segments is 0.74.

IV. Agent's Empathy Classification

The importance of the automatic classification of empathy has been highlighted in [1], [29] where behavioral analysis experiment has been conducted by human experts in workplaces such as the call centers to evaluate the agent's empathic behavior during the phone conversations.

To conduct the experiments, we used a subset of the corpus, which contained a total of 905 conversations. We have chosen this subset because we have full manual transcriptions for this set, and also we have performed complete acoustic and lexical performance analyses. For the experiments, we designed binary classifiers. In order to define class labels, the conversations containing at least one empathic segment were considered as positive and rest of the conversations were considered as negative. We labeled 302 empathic conversations (33.30%) containing empathic segment(s) as positive examples and 603 non-empathic conversations (66.60%) as negative examples.

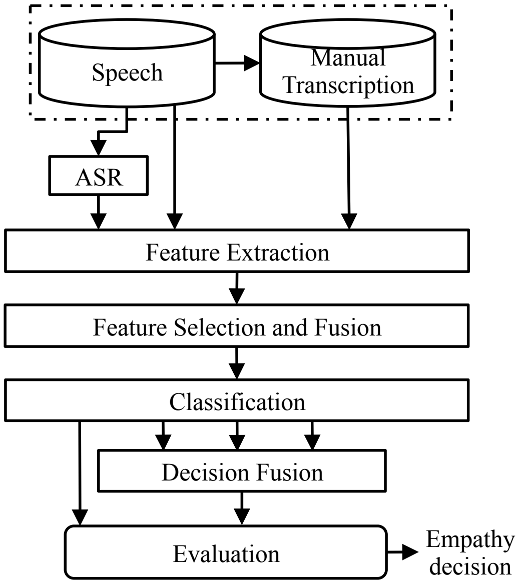

Fig. 2. Conversation Level Agent's Empathy Classification System

In Figure 2, we present a computational architecture of the automatic empathy classification system, which takes the agent's speech channel as input and generates a binary decision regarding the presence (absence) of empathy in the agent's behavior. The system evaluates the cues present throughout the spoken conversation and then commits to the binary decision. To evaluate the relative impact of lexical features, we considered the case of clean transcriptions of the conversation (right branch in Figure 2) as well as the case of noisy transcriptions (left branch in Figure 2) provided by an automatic speech recognizer (ASR). We extracted, combined, and selected acoustic features directly from the speech signal and designed the classification systems. We implemented both feature and decision fusion algorithms (bottom part of Figure 2) to investigate the performance of different configurations of the system.

The ASR system that we used to transcribe the conversations was designed using a portion of the data containing approximately 100 hours of conversations. The system has been designed using Mel-frequency cepstral coefficients (MFCCs) based features with a splice of three frames on each side of the current frame. Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature-space transformations were used to reduce the feature space. We trained the acoustic model using speaker adaptive training (SAT) and also used Maximum Mutual Information (MMI). Word Error Rate (WER) of the system is 31.78% on the test set split [30].

A. Feature Extraction

1) Acoustic Features: The use of large-scale acoustic features was inspired by previous studies in emotion and personality recognition tasks, in which low-level features were extracted and then projected onto statistical functionals [31], [32]. For this study, we extracted features using openSMILE [33]. Before extracting features, we automatically pre-processed speech signals of the conversations to remove silence at the beginning and end of the recordings. We also removed silences longer than one second.

We extracted and categorized features into four different groups, voice-quality, cepstral, spectral, and prosody together with the list of statistical functionals. It was recently shown that grouping the acoustic features followed by feature selection improves the performance of the classification [34]. We thus grouped a large set of acoustic features. In addition to the feature set defined in [35], which has 130 low-level features including first-order derivatives, we also used formants features, constituting 150 low-level features in total. We extracted low-level acoustic features at approximately 100 frames per second. For the voice-quality features the frame size was 60 milliseconds with a gaussian window function and $\sigma = 0.4$. A frame size of 25 milliseconds with a hamming window function was used for the other low-level features.

The low-level acoustic features include zero crossing rate, mel-frequency cepstral coefficients (MFCC1-12), root mean square (RMS) frame energy, fundamental frequency (F0), pitch (F0 with cepstral and autocorrelation), harmonics-to-noise ratio (HNR), voice quality (probability of voicing), mel-spectrum band $1-26$, spectral features with different bands ($0-250Hz$, $0-650Hz$, $250-650Hz$, $1-4kHz$), spectral roll-off points, centroid, flux, max and min. The 39 statistical functional includes range, max, min, linear and quadratic regression coefficients, arithmetic and quadratic mean, geometric and quadratic mean of non-zero values, quartiles and interquartile range, percentile (95%, 98%), standard deviation, variance, kurtosis, skewness, centroid, zero crossing rate and different peaks.

2) Lexical Features: We extracted lexical features from both manual and automatic transcriptions. To utilize the contextual benefits, we extracted trigram features, which eventually results in a very large dictionary. Therefore, we filtered out lower frequency features by preserving $10K$ most frequent n-grams. We then transformed lexical features into bag-of-ngrams (vector space model) with logarithmic term frequency (tf) multiplied with inverse document frequency (idf) – tf-idf, as presented in the equation 1. Here, we considered the conversation as a document.

$$tf \times idf = log(1 + f_{ij}) \times log\left(\frac{\text{number of conversations}}{\text{number of conversations that include word i}}\right) \quad (1)$$

where f_{ij} is the frequency for word i in conversation j. It assigns a weight to each term of the conversation. Its value is highest when the word, i, appears many time in a few conversation, which leads to higher discriminating power for the classification. It is lower when the word appears fewer times in a conversation and appears in many conversations, which represents less discriminating power.

3) Psycholinguistic Features: Psycholinguistic features were extracted using Linguistic Inquiry Word Count (LIWC). The LIWC is designed by Pennebaker et al. It includes psycholinguistic word categories associated with a set of lexica [36]. The word categories include family, cognitive mechanism, affect, occupation, body, article, and function words. It is a knowledge-based system comprised with dictionaries for different languages including Italian. The Italian version of the dictionary contains 85 word categories [37]. We also extracted 5 general descriptors and 12 punctuation categories constituting a total of 102 features. We then removed LIWC features not observed in our training dataset and obtained a final set of 89 features. The LIWC feature processing differs according to types of features such as counts and relative frequencies [36].

B. Feature Selection and Combination

We extracted a large number of features for both acoustic and lexical sets. In order to reduce the computational cost and avoid overfitting we have chosen Relief [38] as a feature selection technique. In a previous study [13], we comparatively evaluated this technique against other algorithms such as information gain, and it performed best in terms of classification performance and computational cost. In order to select the best set of features, we ranked the features according to the Relief's score and generated feature learning curve by incrementally adding batches of ranked features. Before applying feature selection, we discretized the feature values into 10 equal frequency bins, where each bin contains an approximately equal number of values. For the feature fusion, we merged acoustic and lexical features into a single vector to represent each instance.

C. Classification and Evaluation

In this study, we designed binary classification models using Support Vector Machines (SVM) [39]. We chose the linear kernel in order to alleviate the problem of higher dimensions of lexical and combination of $acoustic + lexical$ features. We used a gasussian kernel with different groups of acoustic features and psycholinguistic features as it performed better with the small-sized feature set. We optimized the penalty parameter C of the error term by tuning it in the range $C \in [10^{-5}, ..., 10]$ and the gaussian kernel parameter G in the same range as well, using cross-validation.

At the feature fusion level, we applied feature selection on the combined acoustic and lexical features. For the decision fusion, we combined decisions from the best classifiers of three different feature sets by applying *majority voting*. In the experiment with acoustic features, we first applied feature selection for each group, then merged the feature vectors into one single vector. We then re-applied the feature selection process to the merged feature vector to obtain an optimal subset from all groups.

We measured the performance of the system using the Un-weighted Average (UA), which has been widely used in the evaluation of paralinguistic tasks [14]. UA is the average recall of positive and negative classes and is computed as $UA = \frac{1}{2}\left(\frac{tp}{tp+fn} + \frac{tn}{tn+fp}\right)$, where tp, tn, fp, fn are the number of true positives, true negatives, false positives and false negatives, respectively. Due to the limited size of the conversational dataset and the skewed distribution of the agents we opted to use the Leave-One-Speaker-Group-Out (LOSGO) cross-validation method. In LOSGO, for each fold, we included a) agent's spoken conversation-side features, b) a random selection of conversations, and c) a class label distribution close to the corpus empirical distribution.

D. Results and Discussion

In Table I, we report the performances of the classification system for a single feature type, feature combination, and

TABLE I. EMPATHY CLASSIFICATION RESULTS AT THE CONVERSATION LEVEL USING ACOUSTIC, LEXICAL, AND PSYCHOLINGUISTIC (LIWC) FEATURES TOGETHER WITH FEATURE AND DECISION LEVEL FUSION. AC - ACOUSTIC FEATURES; LEX (M) - LEXICAL FEATURES FROM MANUAL TRANSCRIPTIONS; LEX (A) - LEXICAL FEATURES FROM AUTOMATIC TRANSCRIPTIONS; AC+LEX - LINEAR COMBINATION OF ACOUSTIC AND LEXICAL FEATURE; MAJ - MAJORITY VOTING; LIWC (M) - PSYCHOLINGUISTIC FEATURES EXTRACTED FROM MANUAL TRANSCRIPTIONS; LIWC (A) - PSYCHOLINGUISTIC FEATURES EXTRACTED FROM AUTOMATIC TRANSCRIPTIONS. DIM. - FEATURE DIMENSION.

Experiments	**Dim.**	**UA-Avg**	**UA-Std**
Random baseline		49.7	2.2
Ac	200	61.1	4.3
Lex (M)	5000	63.5	5.5
Lex (A)	3800	62.3	5.3
LIWC (M)	89	63.4	4.8
LIWC (A)	89	62.9	4.1
Ac+Lex (M)	6800	62.3	5.9
Ac+Lex (A)	6600	60.0	4.4
Maj: {Ac,Lex(M),LIWC(M)}		**65.1**	6.2
Maj: {Ac,Lex(A),LIWC(A)}		**63.9**	4.5

classifier combination. We report them in terms of average UA of the LOSGO cross-validation and its standard deviation. We computed the baseline by randomly selecting the class labels, such as empathy and non-empathy, based on the prior class distribution of the training set.

In Table I, we present that the system trained on lexical features extracted from manual transcriptions outperformed any other system trained on single feature type. The features from the ASR transcriptions outperformed all automatically extracted features, including the acoustic-only system, Ac. The results of the acoustic feature are better than random baseline, which was statistically highly significant with $p-value < 0.001$. The value 0.001 refers to the significance level with statistically highly significant. It provides a useful label prediction when no transcriptions are available. We obtained better results with *majority voting*. The statistical significance test between Lex, and $Maj(A)$ revealed that the improved performance of $Maj(A)$ was statistically significant with $p-value < 0.05$. We performed significance test using paired t-test over the set, where each set contains 10 LOSGO cross-validated estimates. Compared to the baseline, the best model for automatic classification provides a relative improvement over the baseline of 31%. In addition, all systems' results are higher and statistically highly significant with $p - value < 0.001$ compared to the baseline results. Linear combination of lexical with acoustic features in the $Ac + Lex(M)$ and $Ac + Lex(A)$ systems did not provide statistically significant change in performance. Despite its success in other paralinguistic tasks [32], the linear combination of the feature space does not necessarily provide improved performance even when combined with feature selection.

The results of the psycholinguistic feature set indicate its usefulness with which we obtained a comparable performance compare to other feature sets. Some of the distinguishing features of this feature set are perceptual e.g., feel and cognitive e.g., certainty, which are ranked using relief feature selection technique. From the investigation of acoustic features, our findings suggest low-level spectral, F0-envelope and MFCC features contribute most to the classification decision, whereas the higher-level statistical functionals are peak and regression (linear and quadratic) coefficients.

V. CONCLUSIONS

Being empathic is critical for humans and their prosocial behavior as well as to facilitate human-machine interactions. In this paper, we propose an automatic empathy classification system based on an operational model. It has been designed by following Gross' modal model of emotions and by analyzing real-life call center's spoken conversations. We designed binary classifiers and investigated acoustic, lexical and psycholinguistic features, and their decision and feature level fusion. The results of the automatic classification system on call center conversations are very promising compared to the baseline. The findings also suggest that lexical and psycholinguistic features extracted from automatic transcription can be useful for the automatic classification task. Clearly, this study shades the light towards designing natural human-machine interaction system, speech, behavioral analytics systems and summarizing large-scale call center conversations in terms of emotional manifestations. In our subsequent study, we focus on designing a fully automated segment level classification system, which will lead us to the design of affective scene i.e., emotional sequence over a complete conversation.

VI. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610916- SENSEI.

REFERENCES

[1] E. Stepanov, B. Favre, F. Alam, S. Chowdhury, K. Singla, J. Trione, F. Béchet, and G. Riccardi, "Automatic summarization of call-center conversations," in *In Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

[2] F. De Vignemont and T. Singer, "The empathic brain: how, when and why?" *Trends in cognitive sciences*, vol. 10, no. 10, pp. 435–441, 2006.

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

[4] M. Ochs, D. Sadek, and C. Pelachaud, "A formal model of emotions for an empathic rational dialog agent," *Autonomous Agents and Multi-Agent Systems*, vol. 24, no. 3, pp. 410–440, 2012.

[5] I. Leite, A. Pereira, S. Mascarenhas, C. Martinho, R. Prada, and A. Paiva, "The influence of empathy in human–robot relations," *Int. Jr. of Human-Computer Studies*, vol. 71, no. 3, pp. 250–260, 2013.

[6] A. Deshmukh, G. Castellano, A. Kappas, W. Barendregt, F. Nabais, A. Paiva, T. Ribeiro, I. Leite, and R. Aylett, "Towards empathic artificial tutors," in *Proc. of the 8th Int. Conf. HRI2013*. IEEE Press, 2013, pp. 113–114.

[7] H. Boukricha, I. Wachsmuth, M. N. Carminati, and P. Knoeferle, "A computational model of empathy: Empirical evaluation," in *Proc. of Humaine Association Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 1–6.

[8] P. Baranyi, A. Csapo, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*. Springer, 2015.

[9] P. R. Gesn and W. Ickes, "The development of meaning contexts for empathic accuracy: Channel and sequence effects." *Journal of Personality and Social Psychology*, vol. 77, no. 4, p. 746, 1999.

[10] C. A. Lietz, K. E. Gerdes, F. Sun, J. M. Geiger, M. A. Wagaman, and E. A. Segal, "The empathy assessment index (eai): A confirmatory factor analysis of a multidimensional model of empathy," *JSSWR*, vol. 2, no. 2, pp. 104–124, 2011.

[11] S. Baron-Cohen, M. Lombardo, H. Tager-Flusberg, and D. Cohen, Eds., *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*, 3rd ed. Oxford University Press, 2013.

[12] G. Riccardi and D. Hakkani-Tür, "Grounding emotions in human-machine conversational systems," *Lecture Notes in Computer Science, Springer-Verlag*, pp. 144–154, 2005.

[13] F. Alam and G. Riccardi, "Comparative study of speaker personality traits recognition in conversational and broadcast news speech," in *Proc. of Interspeech*. ISCA, 2013, pp. 2851–2855.

[14] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *Proc. of Interspeech*, 2009, pp. 312–315.

[15] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[16] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2005, pp. 940–943.

[17] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, "Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings," in *Proc. of IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011, pp. 43–50.

[18] I. Leite, A. Pereira, S. Mascarenhas, G. Castellano, C. Martinho, R. Prada, and A. Paiva, "Closing the loop: from affect recognition to empathic interaction," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*. ACM, 2010, pp. 43–48.

[19] B. Xiao, Bo, S. Daniel, I. Maarten Van, A. Zac E., G. David C., N. Panayiotis G., and S. S., "Modeling therapist empathy through prosody in drug addiction counseling," in *Proc. of Interspeech*, 2014, pp. 213–217.

[20] M. L. Hoffman, "Empathy and prosocial behavior," *Handbook of Emotions*, vol. 3, pp. 440–455, 2008.

[21] J. J. Gross, "The emerging field of emotion regulation: An integrative review," *Review of General Psychology*, vol. 2, no. 3, p. 271, 1998.

[22] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Networks*, vol. 18, no. 4, pp. 317–352, 2005.

[23] J. J. Gross and R. A. Thompson, "Emotion regulation: Conceptual foundations," *Handbook of Emotion Regulation*, vol. 3, p. 24, 2007.

[24] J. J. Gross and L. F. Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion review*, vol. 3, no. 1, pp. 8–16, 2011.

[25] M. Danieli, G. Riccardi, and F. Alam, "Emotion unfolding and affective scenes: A case study in spoken conversations," in *Proc. of Emotion Representations and Modelling for Companion Systems (ERM4CT) 2015,*. ICMI, 2015.

[26] M. Danieli, G. Riccardi, and F. Alam, "Annotation of complex emotion in real-life dialogues," in *Proc. of 1st Italian Conf. on Computational Linguistics (CLiC-it) 2014*, R. Basili, A. Lenci, and B. Magnini, Eds., vol. 1, no. 122–127, 2014.

[27] T. Schmidt, "Transcribing and annotating spoken language with EXMARALDA," in *Proc. of LREC 2004 Workshop on XML-based Richly Annotated Corpora*, 2004, pp. 69–74.

[28] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[29] D. Morena, R. Giuseppe, E. Barker, J. Foster, A. Funk, R. Gaizauskas, M. Hepple, E. Kurtic, M. Poesio, L. Molinari, and V. Giliberti, *Preliminary Version of Use Case Design. SENSEI project deliverable D1.1*, D. Morena and R. Giuseppe, Eds. University of Trento, 2014. [Online]. Available: http://www.sensei-conversation.eu/

[30] S. A. Chowdhury, G. Riccardi, and F. Alam, "Unsupervised recognition and clustering of speech overlaps in spoken conversations," in *Proc. of Workshop on Speech, Language and Audio in Multimedia - SLAM2014*, 2014, pp. 62–66.

[31] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[32] F. Alam and G. Riccardi, "Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 955–959.

[33] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM international conference on Multimedia (ACMM)*. ACM, 2013, pp. 835–838.

[34] S. A. Chowdhury, M. Danieli, and G. Riccardi, "Annotating and categorizing competition in overlap speech," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[35] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. of Interspeech*, 2013.

[36] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[37] F. Alparone, S. Caso, A. Agosti, and A. Rellini, "The italian liwc2001 dictionary." LIWC.net, Austin, TX, Tech. Rep., 2004.

[38] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Proc. of Machine Learning: European Conference on Machine Learning (ECML)*. Springer, 1994, pp. 171–182.

[39] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Research, Tech. Rep., 1998.