

Automated Empathy Detection for Oncology Encounters

Zhuohao Chen¹, James Gibson¹, Ming-Chang Chiu¹, Qiaohong Hu², Tara K. Knight³,
Daniella Meeker², James A. Tulsky⁴, Kathryn I. Pollak⁵, Shrikanth Narayanan¹

¹Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA

²Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

³Schaeffer Center for Health Policy & Economics, University of Southern California, Los Angeles, CA, USA

⁴Dana Farber Cancer Institute and Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

⁵Duke Cancer Institute, Durham, NC, USA

Email: ¹sail.usc.edu, ²dmeeker@usc.edu, ³knight@healthpolicy.usc.edu, ⁴JamesA_Tulsky@dfci.harvard.edu,
⁵kathryn.pollak@duke.edu

Abstract—Empathy involves understanding other people's situation, perspective, and feelings. In clinical interactions, it helps clinicians establish rapport with a patient and support patient-centered care and decision making. Understanding physician communication through observation of audio-recorded encounters is largely carried out with manual annotation and analysis. However, manual annotation has a prohibitively high cost. In this paper, a multimodal system is proposed for the first time to automatically detect empathic interactions in recordings of real-world face-to-face oncology encounters that might accelerate manual processes. An automatic speech and language processing pipeline is employed to segment and diarize the audio as well as for transcription of speech into text. Lexical and acoustic features are derived to help detect both empathic opportunities offered by the patient, and the expressed empathy by the oncologist. We make the empathy predictions using Support Vector Machines (SVMs) and evaluate the performance on different combinations of features in terms of average precision (AP).

Index Terms—oncology, empathic interactions, multimodal system, automatically detect

I. INTRODUCTION

In 2019, over 1.7 million Americans will be diagnosed with cancer, and it will be the cause of death for over half a million people [1]. Cancer diagnosis and treatment carries both a physical burden and, for many patients, severe psychological distress [2], [3]. Conversations between patients and oncologists acknowledge that emotion may reduce this distress [4]–[9]. Consensus exists that eliciting patient goals and responding to emotion are important in clinical practice [10]–[12], and patients whose physicians participate in communication training programs may benefit [12]–[14].

A comprehensive systematic review of 39 oncology studies concluded that empathy is associated with higher patient satisfaction, better psychosocial adjustment, lessened psychological distress and the need for information [15], [16]. Many definitions and metrics for empathy and closely related concepts in medical communication have been developed and deliberated. Metrics rely on provider reports, patient reports, and direct observation of conversations using structured coding systems [17]–[19]. There has been extensive development of coding systems for characterizing patient-provider communications.

Training programs that measure and provide feedback about empathic communication are effective in improving communications and outcomes [10], [20], [21]. Effective communication is a learnable clinical skill that can have meaningful consequences on patients' quality of life and decision-making.

A novel approach to training, SCOPE (Studying Communication in Oncologist Patient Encounters), provides personalized feedback to physicians based on manual coding on recorded real-world conversations between patients and providers [22]. Coded conversations are packaged so that clinicians can review feedback in online performance reports. In a multisite randomized controlled trial, physicians receiving personalized feedback were twice as likely to use empathic statements and to respond appropriately to empathic opportunities [21], and their patients reported greater trust in their physicians using the 5-point Trust in Physicians Scale [23]. While this approach to improving communication with personalized feedback increases empathic responses and patient trust at a fraction of the expense of conventional training, it has limited scalability. Efficiently scaling personalized training while addressing concerns of bias has potential to transform cancer care by improving communication and shared decision-making [24].

This interdisciplinary project is devoted to accelerating the feedback process with automation. Computational methods have promising results that are well-aligned with theories grounded in cognitive science and neuroscience. Work by the SAIL team [25]–[28] have focused on a wide range of provider-patient interactions in diagnostic and intervention settings, including psychotherapy [29]. Supervised learning algorithms have successfully been used to predict empathy categorizations based on the Motivational Interviewing Skill Code [30] and the Motivational Interviewing Treatment Integrity coding systems [31], [32]. Automated signal processing and machine learning tools that extract features from dialogues associated with empathy [33], were associated with manually assigned empathy scores [34], [35]. Although the basic science is in its early stage, computational empathy analysis has reinforced longstanding theories in cognitive science and

neuroscience. For example, empathic interactions have been recognized to have features of entrainment—a synchrony behavior associated with neurobiology as deeply rooted as mirror neurons [36].

However, annotation for empathic interactions is still a manual process which has a prohibitively high cost. The coders have to listen to the whole recording carefully to find when the empathic opportunities and responses occur before coding. Building a automatic system to detect the empathic interactions can help reduce the cost of this process. Recently, an empathic conversational system for human-human spoken dialogues was proposed in [37], that perceives and annotates empathy in customer-agent conversations from the call center corpus in Italian. However, these dialogues involves only one patient and one therapist who speak in separated audio channels. In this paper, we aim to detect empathic interactions in real-world face-to-face oncology conversational recordings, that included multiple speakers mixed in the same channel. We evaluated a multimodal system with a speech and language processing pipeline to 1) segment, diarize and transcribe the audio signals, and 2) extract different types of lexical and acoustic features to predict the empathic interactions. Our result shows the system can filter a subset of the recording which includes most of the empathic interactions.

II. DATA

A. Corpus

We use the data from the COPE study [21], [22] (a follow-up study to the original SCOPE trial) in which the audio of 435 oncology encounters, about 164.8 hours in total, were recorded at a 16kHz sampling rate. Among these recordings, 52 conversations are transcribed. All sessions included a patient (PAT), and one or more healthcare providers (HCPs). In some sessions a third party was present, typically friends or family members (FFs) of the patient. Besides audio recordings, the number of total speakers for each session was also provided. A single session had 3.66 speakers on average.

B. Annotation Information

These oncology sessions were coded by two trained research staff listening to the recordings following the measures described in [38]. Whenever the coder perceives the patient expressing a negative emotion, he records the start time and end time of this empathic interaction which includes both the patient’s empathic opportunity and the extent of oncologists’ empathic response. There are 270 empathic interactions in all 435 sessions. Among these 60 sessions are annotated by both coders, each coder listens to 30 encounters first coded by the other, and then decides whether he agrees with the recorded interactions and points out the empathic interactions not recorded. The coding agreement between coders is $\kappa=0.71$.

III. COMPUTATIONAL SYSTEM DESCRIPTION

In the COPE coding process, coders were instructed to identify empathic opportunities based on patient expressions of

emotion, both direct and indirect Physicians’ responses were categorized primarily based on lexical content. The empathic interactions were coded based on both lexical and acoustic characteristics. In Fig. 1, we constructed a multimodal system for empathy prediction. We extracted lexical and acoustic features with the help of a speech pipeline (shown inside the dashed border in the figure). Empathy prediction was made using the acoustic and lexical information.

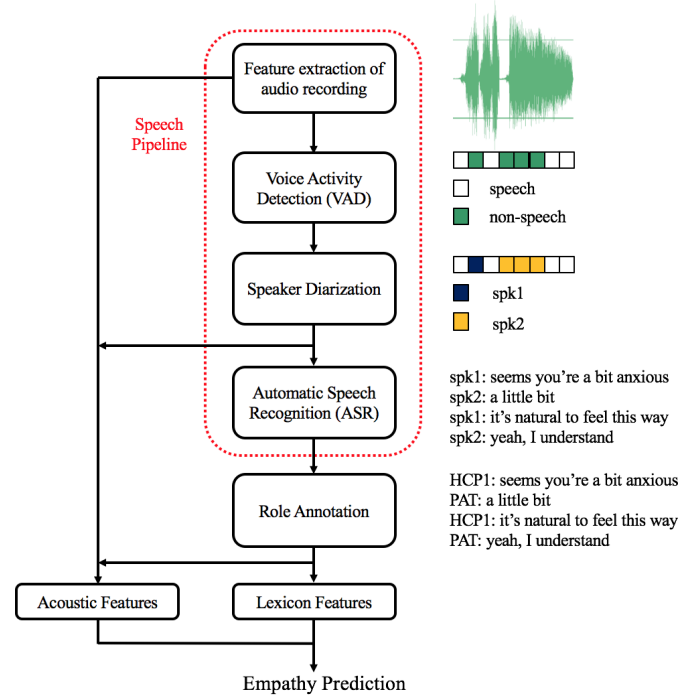


Fig. 1: The Multimodal System

A. Speech Processing Pipeline

The speech pipeline includes a number of components, further described below.

1) *Voice Activity Detection*: The Voice Activity Detection (VAD) model separates the audio into speech and non-speech part using MFCC features. We pre-trained a 2-layer feed-forward neural network using the DARPA RATs data of noisy speech [39].

2) *Speaker Diarization*: The speaker diarization module, followed by VAD, determined who is speaking when in an audio stream. It is challenging due to overlapping speech, rapid speaker changes and non-stationary noise. In this pipeline, we directly used the pre-trained Kaldi model available online [40] which is an x-vector diarization model with Probabilistic Linear Discriminant Analysis (PLDA) framework [41]. The Fig. 1 shows that the diarization module clusters speech segments for different speakers, but without role information.

3) *Automatic Speech Recognition*: The Automatic Speech Recognition (ASR) module transcribes the audio signal features into text. We trained the ASR model with a large combined dataset including Fisher [42], Librispeech [43], Tedlium [44], ICSI [45], AMI [46], WSJ [47] and Hub4 [48]

using the ASPIRE recipe [49]. The language model (LM) is a 3-gram model trained with the SRILM toolkit [50]. The model we applied is based on interpolating the LMs trained on Fisher (a corpus of telephone conversations) and a general psychotherapy corpus [51] respectively. The mixing weight for Fisher is 0.2. To assess how the speech pipeline performed, we applied the gentle forced aligner [52] on the 52 oncology encounter recordings with transcripts to achieve time-stamps for words, and then filled them into the segments to achieve the reference text. The Word Error Rate (WER) of the ASR model was 45.37%. One of the major sources for the error is that there were always multiple speakers in one session, making the diarization tougher. Nonetheless, the decoded transcripts still retained lexical information useful for empathy detection.

B. Role Annotation

An empathic interaction contains empathic opportunities and responses between the PAT and HCP. To discern opportunity and response, it is important to know whether an utterance is made by a PAT or a HCP. We collected the number of words for PAT, HCP and FF by the transcribed sessions of 200K words, the ratio of their utterances is 41%:54%:5%. The general idea behind role annotation is to train 3-gram LMs of different roles, and for each speaker we find the LM which minimizes the perplexity of his corpus. The utterances of FFs are sparse which are always confused with the ones of PATs, so we attributed FF utterances to PAT. We trained 3-gram background LMs L_{Pb} , L_{Hb} for PAT and HCP with the utterances of the therapist and patient of previously mentioned general psychotherapy corpus respectively. The 3-gram in-domain LM L_{Pi} of PAT is trained using the PAT and FF's corpus of the transcribed COPE encounters, while the in-domain LM L_{Hi} of HCP is trained by the HCP's corpus of those transcripts. The LMs \tilde{L}_P and \tilde{L}_H for PAT and HCP are expressed as

$$L_P = \lambda_1 L_{Pb} \oplus (1 - \lambda_1) L_{Pi} \quad (1)$$

$$L_H = \lambda_1 L_{Hb} \oplus (1 - \lambda_1) L_{Hi} \quad (2)$$

$$\tilde{L}_P = \lambda_2 L_P \oplus (1 - \lambda_2) L_H \quad (3)$$

$$\tilde{L}_H = (1 - \lambda_2) L_P \oplus \lambda_2 L_H \quad (4)$$

where (1) and (2) interpolate the in-domain and background LMs with $\lambda_1 = 0.5$, while the formulas (3) and (4) ensure the two LMs are used with the same vocabulary with $\lambda_2 = 0.01$.

The 5-fold cross validation annotation result of the 52 transcribed sessions are shown in Table I. We find most HCPs and PATs are correctly assigned. For FFs, the majority class is PAT which is consistent with our proposition, though 39% of them are identified as HCP.

TABLE I: Role Annotation of Transcribed Sessions

True Role	Predicted Role	
	HCP	PAT
HCP	52	0
PAT	1	83
FF	13	20

IV. EMPATHY DETECTION

One of the challenging problems in comparison to prior work is that the manual coding process assigned time stamps for potentially empathic interactions, and the duration of empathic interactions varies from 3 seconds to 93 seconds which is hard to keep track of. So the first step of empathy detection is generating proper training and testing sample segments. Next we extract lexical and acoustic features for both roles in each segment and made multimodal prediction to find empathic interactions by different combinations of features. If there is only one role in one segment, we set a zero vector for the absent role. The structure of the detection schema is shown in Fig. 2.

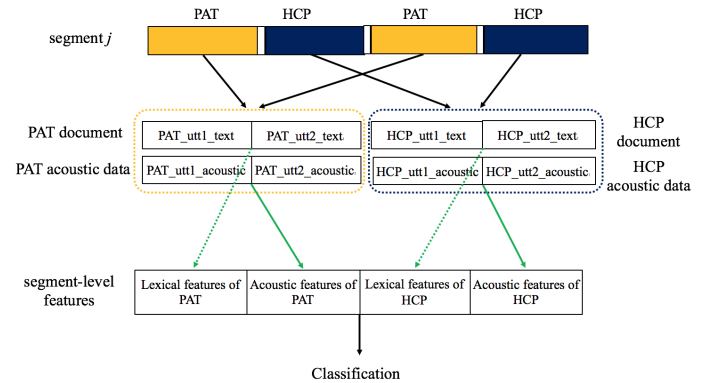


Fig. 2: Structure of the Detection Schema

A. Generating and Labeling Samples

The speech pipeline discussed in the previous section provides us the decoded utterances with time stamps and role annotations. Because decoded utterances are usually fragmented, we group the decoded utterances into segments of approximately 25 seconds, which is the average duration of empathic interactions. As shown in Fig. 3, we group the neighboring utterances into a single segment until the duration of the segment is closest to 25 seconds. For each session, the first group of utterances starts from the beginning of the recording. These extracted segments are samples in our task.

Having segmented samples generated, we label them as positive or negative according to its overlap with empathic interactions. The Fig. 4 denotes how we assign the labels to segments. We label a segment as positive if it has more than 1 second overlapping time with any empathic interaction. The example in the figure shows that sometimes an empathic interaction can produce more than one positive sample. We

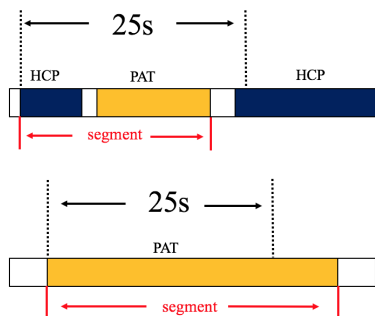


Fig. 3: Segment Generating

call these positive samples the "children" of this empathic interaction. Using this labeling schema we collect 470 positive samples and 21871 negative samples. From these segments we generate features based on lexical, linguistic, and acoustic properties.

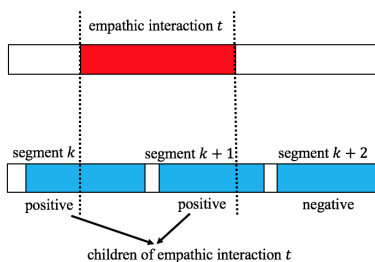


Fig. 4: Labeling Segments

B. Lexical Features

As presented in Fig. 2, we collected the decoded transcripts for PAT and HCP in each segment. To exploit the lexical characteristics of segments, three types of features are extracted.

1) *Doc2vec*: A sentence embedding approach, doc2vec, learning to represent variable-length pieces of texts with the fixed-length features is adopted [53]. It clusters sentences with similar meanings which can be used to learn the general behavior and context represented in the text transcripts. In this paper, we pre-trained the doc2vec with the general psychotherapy corpus and MI corpus and output 100-dimension embedding for each document.

2) *Linguistic Inquiry Word Count*: The Linguistic Inquiry Word Count (LIWC) [54] is a dictionary based text analysis tool which has been widely used to classify texts along linguistic and psychological dimensions and to predict human behaviors. In this paper, we extracted 66 LIWC features for each role of the segment. They include 2 general descriptor categories (total word count, percentage of words captured by the dictionary), 22 standard linguistic dimensions, 32 word categories tapping psychological constructs, 7 personal concern categories and 3 paralinguistic dimensions.

3) *Empath*: Empath [55] is a tool similar to LIWC with much larger lexicon mined on the modern text on web. We took all the pre-built concepts provided by Empath to obtain

194-dimension feature vectors for PAT and HCP documents in each segment.

C. Acoustic Features

We separated acoustic features into two categories, cepstra (representing segmental speech properties) and prosody (for capturing suprasegmental speech properties). The first group consists of 12 mel-frequency cepstral coefficients (MFCC 1–MFCC 12) which are used in the speech pipeline. The second includes pitch, energy, jitter and shimmer which are relevant to the prosody. The energy feature is presented by MFCC0. Both MFCCs and pitch are extracted using Kaldi toolkit [56], [57]. Jitter and Shimmer are computed by openSMILE [58] toolkit. For all the features the frame size is set to 25ms at a shift of 10ms and we applied z-normalize to them in the speaker level. For each role, we concatenate the normalized features and extract their descriptive statistics of max, min, mean, median, standard deviation, skewness and kurtosis.

TABLE II: Features Description

Feature	Description	Dims for Each Role
Doc2vec	document embedding	100
LIWC	features of psychological states	66
Empath	pre-built features generate from common topics on web	194
cepstrum	max, min, mean, median, standard deviation, skewness and kurtosis of MFCC1-12	84
Prosodic	max, min, mean, median, standard deviation, skewness and kurtosis of log-pitch, energy, jitter and shimmer	28

D. Prediction Scheme

The summary of features is shown in Table II. We concatenated different combinations of doc2vec, LIWC, empath, cepstral and prosodic features for both PAT and HCP. Then we trained the classification model using Support Vector Machines (SVMs). In testing we returned the posterior probabilities [59] and sorted out the instances in terms of their probabilities of being a positive empathy sample.

We randomly split the data into training and testing sets by sessions with ratio of roughly 3:1, while excluding the transcribed sessions from the testing one and making sure no speaker exists in both sets. The data is shown in Table III

TABLE III: Frequency of Samples

Dataset	Positive	Negative	Total Samples	empathic interactions
Train	341	16247	16588	194
Test	129	5624	5753	76

V. EXPERIMENT

A. Evaluation Metrics

The goal of our experiment is to filter out irrelevant content (negative samples) from the recordings while preserving as

many empathic interactions as possible. For the evaluation metric, we adopted the average precision (AP), the area under the precision-recall curve, which is claimed to be a better measure of success of prediction than receiver operating characteristic (ROC) curve when the classes are imbalanced [60].

From Section 3.1 we know that some of the positive samples might come from a single empathic interaction. Considering this, we defined the Empathy Detection Rate (EDR) - the % of the samples needed with the highest probability being positive to recall a certain ratio of empathy interactions. For example, given the output of a model, EDR 50% at 5% means that 50% of the empathy interactions are recalled by 5% of samples with the highest probability being a positive sample. An empathy interaction is detected if any one of its children is among these selected segments. We report EDR of recalling 20%, 50%, and 80% empathic interactions.

B. Training Routine

To reduce the class imbalance, we under-sampled the negative instances by 5 in the training set resulting in the ratio between negative:positive being 9.5:1. The SVM module we applied was imported from the scikit-learn package [61]. The `sklearn.svm.SVC` is implemented with Gaussian kernel and we set the parameter `probability=True` to enable probability estimates. We tried different configurations of the penalty parameter $C \in [10^{-2}, 10^{-1}, 1]$, kernel coefficient $\gamma \in [10^{-4}, 10^{-3}, 10^{-2}]$ and the class weight `negative:positive = 1 : W`, where $W \in [1, 2, \dots, 10]$. The optimal parameters were selected by 5-fold cross validation.

C. Results and Analysis

The performances of different feature combinations in terms of average precision is shown in Table IV. We set the baseline performance by random prediction, which is equal to the proportion of the positive samples. By comparing Prosody VS Prosodic + Cepstrum, and Doc2vec + LIWC + Empath + Prosody VS Doc2vec + LIWC + Empath + Prosody + Cepstrum we conclude the cepstral features do not help for empathy prediction. We also find the lexical features have a much better performance than acoustic features, while AP of Empath features is the highest among the single-type lexical features. The best result is achieved by combining Doc2vec, LIWC, Empath and Prosody, which obtains the AP of 7.61% and is much better than the random baseline. The result of AP is generally low because of the sparsity of the empathic interactions in oncology sessions.

A more detailed result is shown in Fig. 5. From this we find that the feature set consisting of Doc2vec, LIWC, Empath and Prosody outperform all the other combinations at most recall levels. When recall < 0.3 , the single feature-type prediction of Empath could detect the positive samples more efficiently. The combined lexical features of Doc2vec + LIWC + Empath have comparable performance compared to Doc2vec + LIWC + Empath + Prosody at low recall levels, but degrade quickly as the recall increases. One possible reason is that the

positive segments with empathic lexical information are easier to detect, while some of the positive samples, which do not have distinct words conveying PAT’s empathic opportunities or HCP’s empathy towards PAT, are more likely to be captured by prosodic features.

TABLE IV: Average Precision for Different Combined Features

Experiment (Feature Combination)	AP (%)
Random Baseline	2.10
Prosody	3.97
Cepstrum	3.20
Prosody + Cepstrum	3.24
Doc2vec	5.56
LIWC	5.09
Empath	6.99
Doc2vec + LIWC + Empath	7.38
Doc2vec + LIWC + Empath + Prosody	7.61
Doc2vec + LIWC + Empath + Prosody + Cepstrum	7.11

The performances of Empathy Detection Rate are shown in Table V. Besides the percentage of samples (POS) needed to recall empathic interactions at different levels, we also present the percentage of length of audio (POA) by summing up the length of the selected samples, and divide it by the total duration of all the audio sessions. The results echo the precision-recall curves in Fig. 5. The outcome of the multimodal prediction from Doc2vec + LIWC + Empath + Prosody shows us that we can detect half of the empathic interactions by listening to only 6.61% of the recording, and recall 80% empathic interactions from 23.48% of the recordings.

TABLE V: Results of Empathy Detection Rate

Level of Recall	20%		50%		80%	
	POS (%)	POA (%)	POS (%)	POA (%)	POS (%)	POA (%)
Metrics						
Prosody	9.66	9.17	23.50	22.16	48.15	45.52
Cepstrum	11.98	11.36	30.73	28.75	51.80	48.47
Prosody + Cepstrum	8.31	7.96	31.76	30.90	60.09	57.28
Doc2vec	2.43	2.33	15.19	14.71	54.49	51.25
LIWC	3.53	3.31	10.41	10.01	52.62	47.28
Empath	2.07	1.94	9.49	8.90	37.96	36.10
Doc2vec + LIWC + Empath	2.66	2.48	7.65	7.16	30.37	28.83
Doc2vec + LIWC + Empath + Prosody	2.62	2.42	7.07	6.61	24.80	23.48
Doc2vec + LIWC + Empath + Prosody + Cepstrum	3.06	2.83	7.51	7.04	31.03	30.04

POS: percentage of samples needed to recall empathic interactions at different levels

POA: percentage of length of audio needed to recall empathic interactions at different levels

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a system for detecting empathic interactions in real-world oncology conversational recordings. The audio was segmented, diarized and transcribed by a speech processing pipeline from which different types of acoustic and lexical features were extracted. After investigating the effectiveness of acoustic and lexical features and their combinations, we found the combination of Doc2vec (sentence embedding), LIWC (psycholinguistic features), Empath and prosodic features achieved the best performance with a recall rate of 80% of the empathic interactions from about 23% of

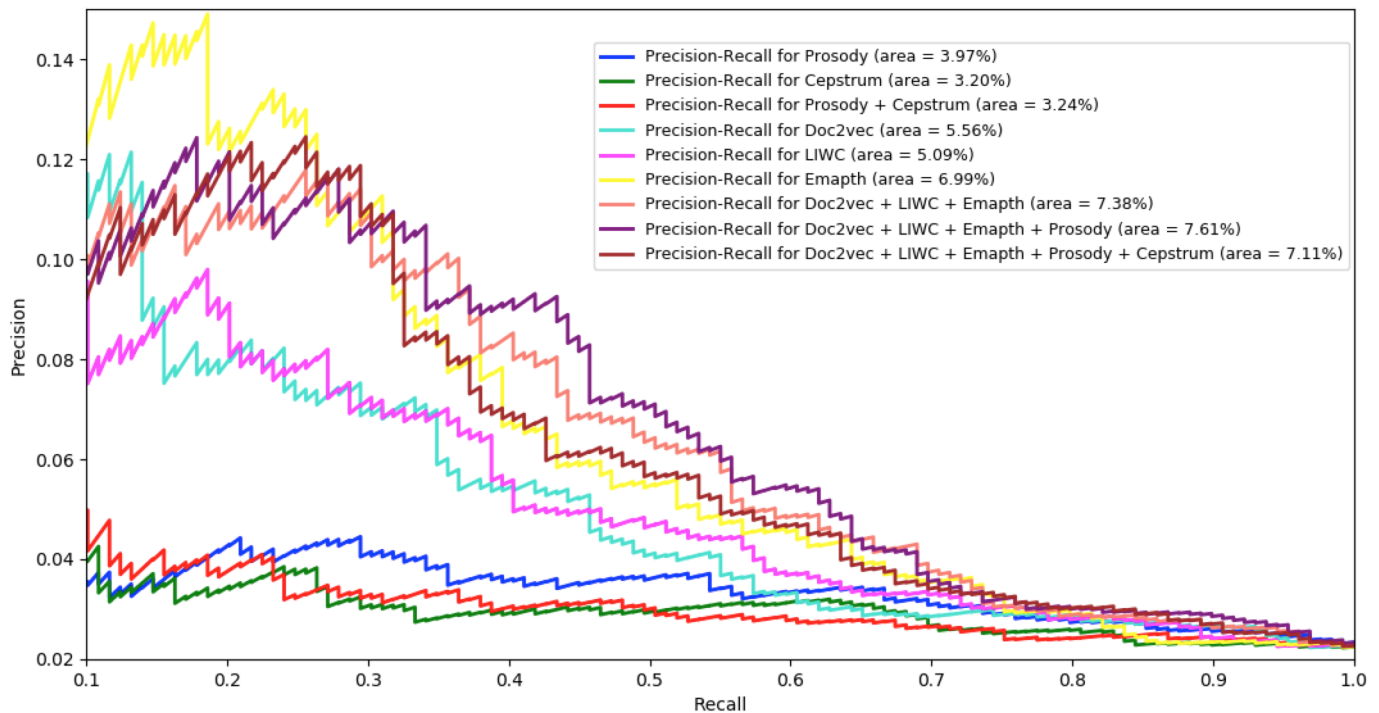


Fig. 5: Precision-Recall Curves for Different Feature Combinations

the recording. This result shows that implementing such a multimodal system is a feasible method that might accelerate and ultimately contribute to replacing the manual annotating processes used in SCOPE. It is possible that the effectiveness of SCOPE feedback can be achieved with very few examples; if this is the case, a low-recall approach may be sufficient. The current approach did not attempt to distinguish patient expressions from providers' empathic responsiveness. Future work may further improve upon this by detecting potentially empathic interactions applying lexical and acoustic methods tuned to identify empathic opportunities from patient utterances [62], [63], which are based on expressions of distress.

REFERENCES

- [1] "Cancer facts & figures 2019 — american cancer society," <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2019.html>.
- [2] J. Zabora, K. BrintzenhofeSzoc, B. Curbow, C. Hooker, and S. Piantadosi, "The prevalence of psychological distress by cancer site," *Psycho-oncology*, vol. 10, no. 1, pp. 19–28, 2001.
- [3] L. R. Derogatis, G. R. Morrow, J. Fetting, D. Penman, S. Piasetsky, A. M. Schmale, M. Henrichs, and C. L. Carnicke, "The prevalence of psychiatric disorders among cancer patients," *Jama*, vol. 249, no. 6, pp. 751–757, 1983.
- [4] M. Neumann, M. Wirtz, N. Ernstmann, O. Ommen, A. Längler, F. Edelhäuser, C. Scheffer, D. Tauschel, and H. Pfaff, "Identifying and predicting subgroups of information needs among cancer patients: an initial study using latent class analysis," *Supportive Care in Cancer*, vol. 19, no. 8, pp. 1197–1209, 2011.
- [5] C. S. Roberts, C. E. Cox, D. S. Reintgen, W. F. Baile, and M. Gibertini, "Influence of physician communication on newly diagnosed breast patients' psychologic adjustment and decision-making," *Cancer*, vol. 74, no. S1, pp. 336–341, 1994.
- [6] T. Takayama, Y. Yamazaki, and N. Katsumata, "Relationship between outpatients' perceptions of physicians' communication styles and patients' anxiety levels in a japanese oncology setting," *Social science & medicine*, vol. 53, no. 10, pp. 1335–1350, 2001.
- [7] R. Spencer, M. Nilsson, A. Wright, W. Pirl, and H. Prigerson, "Anxiety disorders in advanced cancer patients: correlates and predictors of end-of-life outcomes," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 116, no. 7, pp. 1810–1819, 2010.
- [8] D. R. Rutter, G. Ionomou, and L. Quine, "Doctor-patient communication and outcome in cancer patients: An intervention," *Psychology and Health*, vol. 12, no. 1, pp. 57–71, 1996.
- [9] R. Zachariae, C. G. Pedersen, A. B. Jensen, E. Ehrnrooth, P. B. Rossen, and H. von der Maase, "Association of perceived physician communication style with patient satisfaction, distress, cancer-related self-efficacy, and perceived control over the disease," *British journal of cancer*, vol. 88, no. 5, p. 658, 2003.
- [10] J. M. Clayton, P. N. Butow, A. Waters, R. C. Laidsaar-Powell, A. O'Brien, F. Boyle, A. L. Back, R. M. Arnold, J. A. Tulskey, and M. H. Tattersall, "Evaluation of a novel individualised communication-skills training intervention to improve doctors' confidence and skills in end-of-life communication," *Palliative medicine*, vol. 27, no. 3, pp. 236–243, 2013.
- [11] A. L. Back, R. M. Arnold, W. F. Baile, K. A. Fryer-Edwards, S. C. Alexander, G. E. Barley, T. A. Gooley, and J. A. Tulskey, "Efficacy of communication skills training for giving bad news and discussing transitions to palliative care," *Archives of internal medicine*, vol. 167, no. 5, pp. 453–460, 2007.
- [12] L. Butcher, "Dana-farber program seeks to improve conversations with patients with serious illness," 2016.
- [13] K. A. Lorenz, J. Lynn, S. M. Dy, L. R. Shugarman, A. Wilkinson, R. A. Mularski, S. C. Morton, R. G. Hughes, L. K. Hilton, M. Maglione *et al.*, "Evidence for improving palliative care at the end of life: a systematic review," *Annals of internal medicine*, vol. 148, no. 2, pp. 147–159, 2008.
- [14] R. M. Epstein, P. Franks, C. G. Shields, S. C. Meldrum, K. N. Miller, T. L. Campbell, and K. Fiscella, "Patient-centered communication and diagnostic testing," *The Annals of Family Medicine*, vol. 3, no. 5, pp. 415–421, 2005.
- [15] S. Lelorain, A. Brédart, S. Dolbeault, and S. Sultan, "A systematic review of the associations between empathy measures and patient

- outcomes in cancer care,” *Psycho-Oncology*, vol. 21, no. 12, pp. 1255–1264, 2012.
- [16] D. S. Morse, E. A. Edwardsen, and H. S. Gordon, “Missed opportunities for empathy in lung cancer communication,” *Archives of internal medicine*, vol. 168, no. 17, pp. 1853–1858, 2008.
- [17] G. W. E. D. L. Wilt and M. R. A. M. O’Neil, “Empathy: A study of two types,” *Issues in Mental Health Nursing*, vol. 19, no. 5, pp. 453–461, 1998.
- [18] D. Kuyk and J. K. Olson, “Clarification of conceptualizations of empathy,” *Journal of Advanced nursing*, vol. 35, no. 3, pp. 317–325, 2001.
- [19] C. Rohani, M. S. Kesbakh, and J. Mohtashami, “Clinical empathy with cancer patients: a content analysis of oncology nurses’ perception,” *Patient preference and adherence*, vol. 12, p. 1089, 2018.
- [20] H. Riess, J. M. Kelley, R. W. Bailey, E. J. Dunn, and M. Phillips, “Empathy training for resident physicians: a randomized controlled trial of a neuroscience-informed curriculum,” *Journal of general internal medicine*, vol. 27, no. 10, pp. 1280–1286, 2012.
- [21] J. A. Tulsky, R. M. Arnold, S. C. Alexander, M. K. Olsen, A. S. Jeffreys, K. L. Rodriguez, C. S. Skinner, D. Farrell, A. P. Abernethy, and K. I. Pollak, “Enhancing communication between oncologists and patients with a computer-based training program: a randomized trial,” *Annals of internal medicine*, vol. 155, no. 9, pp. 593–601, 2011.
- [22] C. M. Koropchak, K. I. Pollak, R. M. Arnold, S. C. Alexander, C. S. Skinner, M. K. Olsen, A. S. Jeffreys, K. L. Rodriguez, A. P. Abernethy, and J. A. Tulsky, “Studying communication in oncologist-patient encounters: the scope trial,” *Palliative Medicine*, vol. 20, no. 8, pp. 813–819, 2006.
- [23] L. A. Anderson and R. F. Dedrick, “Development of the trust in physician scale: a measure to assess interpersonal trust in patient-physician relationships,” *Psychological reports*, vol. 67, no. 3_suppl, pp. 1091–1100, 1990.
- [24] A. K. Pham, M. T. Bauer, and S. Balan, “Closing the patient–oncologist communication gap: a review of historic and current efforts,” *Journal of Cancer Education*, vol. 29, no. 1, pp. 106–113, 2014.
- [25] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, “‘rate my therapist’: Automated detection of empathy in drug and alcohol counseling via speech and language processing,” *PLoS one*, vol. 10, no. 12, p. e0143055, 2015.
- [26] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, “Computational analysis and simulation of empathic behaviors: A survey of empathy modeling with behavioral signal processing framework,” *Current psychiatry reports*, vol. 18, no. 5, p. 49, 2016.
- [27] D. Can, R. A. Marín, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. S. Narayanan, “‘it sounds like...’: A natural language processing approach to detecting counselor reflections in motivational interviewing,” *Journal of counseling psychology*, vol. 63, no. 3, p. 343, 2016.
- [28] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, “Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [29] 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 *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [30] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, “Manual for the motivational interviewing skill code (misc),” *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.
- [31] T. Moyers, T. Martin, J. Manuel, W. Miller, and D. Ernst, “Revised global scales: Motivational interviewing treatment integrity 3.1. 1 (miti 3.1. 1),” *Unpublished manuscript, University of New Mexico, Albuquerque, NM*, 2010.
- [32] B. Xiao, D. Can, P. G. Georgiou, D. Atkins, and S. S. Narayanan, “Analyzing the language of therapist empathy in motivational interview based psychotherapy,” in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.
- [33] J. Jaffe and S. Feldstein, *Rhythms of dialogue*. Academic Press, 1970, vol. 8.
- [34] 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 *INTERSPEECH*, 2013, pp. 2861–2865.
- [35] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, “Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions,” *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.
- [36] V. Gallese, “The ‘shared manifold’ hypothesis. from mirror neurons to empathy,” *Journal of consciousness studies*, vol. 8, no. 5-6, pp. 33–50, 2001.
- [37] F. Alam, M. Danieli, and G. Riccardi, “Annotating and modeling empathy in spoken conversations,” *Computer Speech & Language*, vol. 50, pp. 40–61, 2018.
- [38] K. I. Pollak, R. M. Arnold, A. S. Jeffreys, S. C. Alexander, M. K. Olsen, A. P. Abernethy, C. Sugg Skinner, K. L. Rodriguez, and J. A. Tulsky, “Oncologist communication about emotion during visits with patients with advanced cancer,” *Journal of Clinical Oncology*, vol. 25, no. 36, pp. 5748–5752, 2007.
- [39] K. Snyder and S. Strassel, “The rats radio traffic collection system,” in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.
- [40] Callhome diarization xvector model. [Online]. Available: <https://kaldi-asr.org/models/m6>
- [41] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in *Interspeech*, 2017, pp. 999–1003.
- [42] C. Cieri, D. Miller, and K. Walker, “The fisher corpus: a resource for the next generations of speech-to-text,” in *LREC*, vol. 4, 2004, pp. 69–71.
- [43] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [44] A. Rousseau, P. Deléglise, and Y. Esteve, “Ted-lium: an automatic speech recognition dedicated corpus,” in *LREC*, 2012, pp. 125–129.
- [45] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, “The icisi meeting corpus,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03)*, vol. 1. IEEE, 2003, pp. 1–1.
- [46] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillelot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, “The ami meeting corpus: A pre-announcement,” in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.
- [47] G. R. Doddington, “Csr corpus development,” in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 363–366.
- [48] D. Graff, Z. Wu, R. MacIntyre, and M. Liberman, “The 1996 broadcast news speech and language-model corpus,” in *Proceedings of the DARPA Workshop on Spoken Language technology*, 1997, pp. 11–14.
- [49] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, “Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 539–546.
- [50] A. Stolcke, “Srlm—an extensible language modeling toolkit,” in *Seventh international conference on spoken language processing*, 2002.
- [51] A. S. Press, “Counseling and psychotherapy transcripts, client narratives, and reference works,” 2009.
- [52] “gentle forced aligner.” [Online]. Available: <https://lowerquality.com/gentle/>
- [53] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*, 2014, pp. 1188–1196.
- [54] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, “The development and psychometric properties of liwc2015,” *Tech. Rep.*, 2015.
- [55] E. Fast, B. Chen, and M. S. Bernstein, “Empath: Understanding topic signals in large-scale text,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 4647–4657.
- [56] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” *IEEE Signal Processing Society, Tech. Rep.*, 2011.
- [57] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

- [58] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.
- [59] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
- [60] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLoS one*, vol. 10, no. 3, p. e0118432, 2015.
- [61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [62] S. Sondhi, M. Khan, R. Vijay, and A. K. Salhan, "Vocal indicators of emotional stress," *International Journal of Computer Applications*, vol. 122, no. 15, 2015.
- [63] O. Simantiraki, G. Giannakakis, A. Pampouchidou, and M. Tsiknakis, "Stress detection from speech using spectral slope measurements," in *Pervasive Computing Paradigms for Mental Health*. Springer, 2016, pp. 41–50.