MONAH: Multi-Modal Narratives for Humans to analyze conversations

Joshua Y. Kim
University of Sydney
New South Wales, Australia
josh.kim@sydney.edu.au

Greyson Y. Kim
Success Beyond Pain
Western Australia, Australia
greyson.kim@gmail.com

Chunfeng Liu
Hello Sunday Morning
New South Wales, Australia
ian@hellosundaymorning.org

Rafael A. Calvo
Imperial College London
London, United Kingdom
r.calvo@imperial.ac.uk

Silas C.R. Taylor
University of New South Wales
New South Wales, Australia
silas.taylor@unsw.edu.au

Kalina Yacef*
University of Sydney
New South Wales, Australia
kalina.yacef@sydney.edu.au

Abstract

In conversational analyses, humans manually weave multimodal information into the transcripts, which is significantly time-consuming. We introduce a system that automatically expands the verbatim transcripts of video-recorded conversations using multimodal data streams. This system uses a set of preprocessing rules to weave multimodal annotations into the verbatim transcripts and promote interpretability. Our feature engineering contributions are two-fold: firstly, we identify the range of multimodal features relevant to detect rapport-building; secondly, we expand the range of multimodal annotations and show that the expansion leads to statistically significant improvements in detecting rapport-building.

1 Introduction

Dyadic human-human dialogs are rich in multimodal information. Both the visual and the audio characteristics of how the words are said reveal the emotions and attitudes of the speaker. Given the richness of multimodal information, analyzing conversations requires both domain knowledge and time. The discipline of conversational analysis is a mature field. In this discipline, conversations could be manually transcribed using a technical system developed by Jefferson (2004), containing information about intonation, lengths of pauses, and gaps. Hence, it captures both *what* was said and *how* it was said[1]. However, such manual annotations take a great deal of time. Individuals must watch the conversations attentively, often replaying the conversations to ensure completeness.

Automated Jefferson (2004) transcripts could be generated from video-recordings (Moore, 2015). However, the potential issue with Jeffersonian annotations is that there are often within-word annotations and symbols which makes it hard to benefit from pre-trained word embeddings. Inspired by the Jeffersonian annotations, we expand the verbatim transcripts with multimodal annotations such that downstream classification models can easily benefit from pre-trained word embeddings.

Our paper focuses on the classification task of predicting rapport building in conversations. Rapport has been defined as a state experienced in interaction with another with interest, positivity, and balance (Cappella, 1990). If we can model rapport building in the medical school setting, the volunteer actors can let the system give feedback for the unofficial practice sessions, and therefore students get more practice with feedback. Also, the lecturer could study the conversations of the top performers and choose interesting segments to discuss. As student doctors get better in rapport building, when they graduate and practice as doctors, treatments are more effective and long-term (Egbert et al., 1964; DiMatteo, 1979; Travaline et al., 2005).

Outside of the healthcare domain, understanding and extracting the features required to detect rapport-building could help researchers build better conversational systems. Our first contribution is the identification of multimodal features that have been found to be associated with rapport building and using them to predict rapport building automatically. Our second contribution is to include them into a text-based multimodal narrative system (Kim et al., 2019b). Why go through text? It is because this is how human experts have been manually analyzing conversations in the linguistics community. Our text-based approach has the merit of emulating the way human analysts analyze conversations, and hence supporting better interpretability. We demonstrate that the additions bring statistically significant improvements. This feature-

* Corresponding author

[1] Please visit www.universitytranscriptions.co.uk/jefferson-transcription-example/ for an audio example.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 466–479
April 19 - 23, 2021.

engineering system[2] could potentially be used to accomplish a highly attention-demanding task for an analyst. With an automated text-based approach, we aim to contribute towards the research gap of automatic visualizations that support multimodal analysis (Kim et al., 2019a). The created multimodal transcript itself is a conversational analysis product, which can be printed out on paper.

In this paper, we first introduced the problem domain (section 3). Secondly, we motivated the new features (detailed in Fig. 1) to be extracted (section 4). Then, we extracted the features from videos and encoded them as text together with verbatim transcripts (section 4). To evaluate whether the text narratives were useful, we ran experiments that predict rapport-building using texts containing different amounts of multimodal annotations (section 5). Finally, we discuss the results and visualize the outputs of the system (section 6).

2 Related Works

The automated analysis of conversations has been the subject of considerable interest in recent years. Within the domain of doctor-patient communication, Sen et al. (2017) calculated session-level input features, including affective features (Gilbert, 2014). Analyses using session-level features have a drawback of not being able to identify specific defining multimodal interactions in the conversation (Zhao et al., 2016; Heylen et al., 2007). Therefore, we build upon the works of Sen et al. (2017) – in addition to the use of session-level features, we propose using a finer level of talk-turn multimodal text representation as inputs into a hierarchical attention network (HAN) (Yang et al., 2016).

We also build upon our previous work (Kim et al., 2019b) by broadening the range of multimodal features considered. As for the different methods of multimodal information fusion, Poria et al. (2017) completed an extensive review of the different state-of-the-art multimodal fusion techniques. Recent multimodal fusion research (such as ICON (Hazarika et al., 2018a), CMN (Hazarika et al., 2018b), MFN (Zadeh et al., 2018), DialogueRNN (Majumder et al., 2019), M3ER (Mittal et al., 2020)) has focussed on end-to-end approaches. Unlike the typical end-to-end approach of representing and fusing multimodal features using numeric vectors, our contribution is an entirely text-based multimodal narrative, thereby improving downstream analysis's interpretability. The approach of this system not only annotates the presence of nonverbal events (Eyben et al., 2011), but also the degree of the nonverbal event intensity at both the session-level and talkturn-level.

3 Data

This study uses data from the EQClinic platform (Liu et al., 2016). Students in an Australian medical school were required to complete at least one medical consultation on the online video conferencing platform EQClinic with a simulated patient who is a human actor trained to act as a patient. Each simulated patient was provided with a patient scenario, which mentioned the main symptoms experienced. The study was approved by the Human Research Ethics Committee of the University of New South Wales (project number HC16048).

The primary outcome measurement was the response to the rapport-building question on the Student-Patient Observed Communication Assessment (SOCA) form, an adapted version of the Calgary-Cambridge Guide (Kurtz and Silverman, 1996). Simulated patients used the SOCA form to rate the students' performances after each video consultation. Our dataset comprises of 873 sessions, all from distinct students. Since we have two recordings per session (one of the student, the second of the simulated patient), the number of recordings analyzed is 1,746. The average length per recording is 928 seconds (sd=253 seconds), amounting to a total of about 450 hours of recordings analyzed. The dataset's size is small relative to the number of multimodal features extracted; therefore, there is a risk of overfitting.

We used the YouTube platform to obtain the transcript per speaker from the recordings. We chose YouTube because we (Kim et al., 2019c) found that it was the most accurate transcription service (word error rate: 0.28) compared to Google Cloud (0.34), Microsoft Azure (0.40), Trint (0.44), IBM Watson (0.50), when given dyadic video-conferences of an Australian medical school. Jeong-Hwa and Cha (2020) found that among the four categories of YouTube errors (omission, addition, substitution, and word order), substitution recorded the highest amount of errors. Specifically, they found that phrase repetitions could be mis-transcribed into non-repetitions. From our experience, (a) repair-initiation techniques such as sound stretches (e.g. "ummmm") (Hosoda, 2006), were either omitted or

[2]Open-sourced at https://github.com/SpectData/MONAH

substituted with "um"; (b) overlapping speech was not a problem because our speakers were physically separated and recorded into separate files.

We brought together the two speakers' transcripts into a session-level transcript through word-level timings and grouped together words spoken by one speaker until the sequence is interrupted by the other speaker. When the interruption occurs, we deem that the talk-turn of the current speaker has ended, and a new talk-turn by the interrupting speaker has begun. The average number of talk-turns per session is 296 (sd=126), and the average word count per talk-turn is 7.62 (sd=12.2).

At this point, we note that acted dialogues differ from naturally occurring dialogues in a few ways. Firstly, naturally occurring dialogues tend to be more vague (phrases like "sort of", "kinda", "or something") due to the shared understanding between the speakers (Quaglio, 2008). Secondly, taboo words or expletives that convey emotions (like "shit", "pisssed off", "crap") is likely to be less common in an acted medical setting than naturally occurring conversations. Some conversations transform into genuine dialogues where the speakers "shared parts of themselves they did not reveal to everyone and, most importantly, this disclosure was met with acceptance" (Montague, 2012). This definition of genuine conversation is similarly aligned to our definition of rapport-building in section 4.1.

Figure 1 shows a summary of the features extracted. We annotated verbatim transcripts with two different levels of multimodal inputs – annotations at the session-level are labeled *coarse*, whilst annotations at the talk-turn-level are labeled *fine*. To facilitate comparisons, all input families belonging to the *coarse* (*fine*) level would be annotated with uppercase (lowercase) letters, respectively. In this paper, we refer to the previously existing set of features (with white background) as the "prime" ($'$) configuration. Families are also abbreviated by their first letter. For example, the *coarse* P' family would consist of only speech rate and delay, whilst the *coarse* P family would consist of P' plus tone. As another example, the *coarse* D' family is the same as the D family because there are no newly added features (in blue). We introduce the framework of our multimodal feature extraction pipeline in Figure 2.

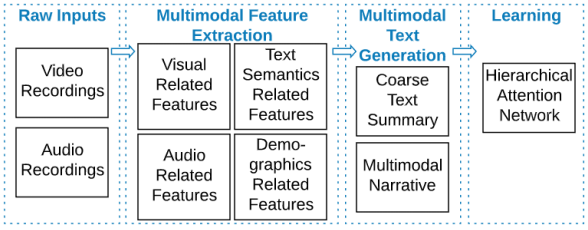

Figure 2: MONAH (Multi-Modal Narratives for Humans) Framework.

4 Multimodal features extractions

As an overview, we extracted the timestamped verbatim transcripts and used a range of pre-trained models to extract temporal, modality-specific features. We relied on pre-trained models for feature extraction and did not attempt to improve on them – demonstrating the value of using multidisciplinary pre-trained models from natural language processing, computer vision, and speech processing for conversational analysis.

Effectively, we extracted structured data from unstructured video data (section 4.2). With the structured data and verbatim transcript, we weaved a multimodal narrative using a set of predefined templates (sections 4.3 and 4.4). With the multimodal narrative, we employed deep learning techniques and pre-trained word embeddings to predict the dependent variable (section 5).

4.1 Dependent variable - rapport building

The dependent variable is defined as the success in rapport building. Rapport building is one of

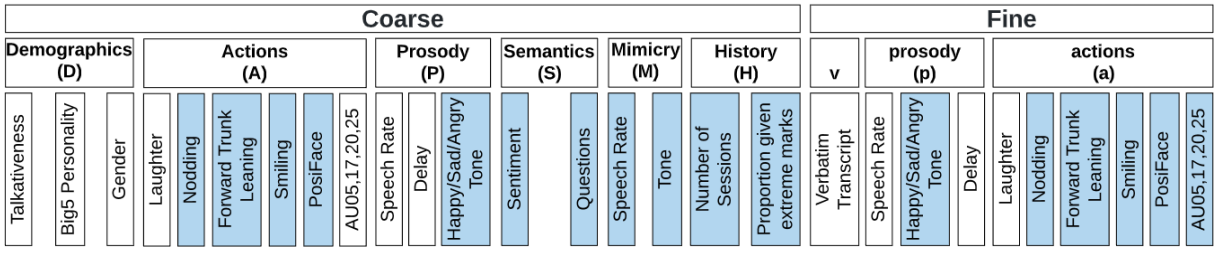

Figure 1: High-level features introduction. We build on our previous work (Kim et al., 2019b) – the new features introduced in this work are coloured in blue, whilst the existing set of features are in white.

the four items scored in the SOCA. The original 4-point Likert scale is Fail, Pass-, Pass, Pass+, we converted this scale into a binary variable where it is true if the rapport-building score is "Pass+" as we are concerned here with identifying good rapport building. "Pass+" means that the actor felt rapport such that all information could be comfortably shared. 38 percent of the population has achieved "Pass+". All actors followed the same pre-interview brief. Because only the actor scored the student performance and there is no overlap, the limitation is that we do not have measures of agreement.

4.2 Description of features

Table 1 gives an overview of all features for each speaker. We define six families of *coarse*-level inputs -– *demographics*, *actions*, *prosody*, *semantics*, *mimicry*, and *history*. We computed the features per speaker. From all families, there are a total of 77 features per session.

We first discuss the family of *demographics*. *Talkativeness* is chosen because the patient's talkativeness would initiate the doctor's active listening while aiding identification of patient's concerns – processes that could establish rapport. In Hall et al. (2009), it appears that patients appreciate a certain degree of doctor's dominance in the conversation, which itself is also correlated with higher rapport. *Big 5 Personality* consists of Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience (McCrae and Costa, 1987). This personality structure is widely used in research and practice to quantify aspects of a person's natural tendency in thought, feeling, and action, with good validity and reliability indicators (McCrae, 2017). It is chosen because traits of agreeableness and openness on the part of both doctor and patient predict higher rapport. Among doctors, higher openness and agreeableness predict higher empathy towards patients (Costa et al., 2014). Among patients, higher agreeableness predicted higher trust towards doctors (Cousin and Mast, 2013), and higher openness predicted higher doctor affectionate communication (Hesse and Rauscher, 2019). *Big 5 Personality* is extracted through feeding transcripts to the IBM Watson Personality Insights API (version 2017-10-13), costing a maximum of 0.02 USD per call. *Gender* is chosen because personality differences between genders were observed cross-culturally. Among twenty-three thousand participants across cultures

Family	Child	Template
Demographics	Talkativeness	Total word count, total distinct word count, and proportion of word count
	Big 5 Personality	Percentile scores for each of the big 5 personality
	Gender	Male or Female
Actions	Laughter	Total laughter count
	Head Nodding*	Count of nods
	Forward Trunk Leaning*	Count of leaning in
	Smiling*	Count of smiles
	PosiFace*	Counts of times of positive and negative facial expressions
	AU	Summary statistics of the selected AU (05,17,20,25) intensities
Prosody	Delay	Summary statistics of time gaps between talk-turns
	Speech rate	Average speech rate
	Tone*	Happy, sad, angry tone
Semantics	Sentiment*	Composite, positive, neutral, and negative sentiment
	Questions*	Proportion of talk-turns that are open/closed questions
Mimicry	Speech Rate*	Dynamic time wrapping distance for speech rate
	Tone*	Dynamic time wrapping distance for tone
History	Num. Sessions*	Number of past sessions the assessor has scored before this
	Proportion given extreme marks*	Proportion of past sessions that the assessor has given an extreme score

Table 1: Session-level input features for each participant. * indicates new features outside of Kim et al. (2019b).

for both college-age and adult samples, females reported higher agreeableness, warmth, and openness to feelings than males (Costa Jr et al., 2001), traits that could be linked to rapport building.

Secondly, for the family of *actions*, *laughter* is chosen because humor (which was defined in part by the presence of laughter) on the part of both doctor and patient was found to be twice as frequent in high-satisfaction than low-satisfaction visits (Sala et al., 2002). Laughter events were detected using the Ryokai et al. (2018) algorithm. *Facial expressions* that resemble smiling is another behavioral indicator of humor appreciation, and approval of one another (Tickle-Degnen and Rosenthal, 1990). *Head nodding* is a type of backchannel response (i.e., response tokens) that has been shown to reflect rapport between doctor and patient, especially when the primary activity is face to face communications (Manusov, 2014). *Forward trunk leaning* is chosen because it has long been found to reflect an expression of interest and caring, which are foundational to rapport building (Scheflen, 1964). Additionally, facial positivity (*posiface*) is included as it is useful in rapport building detection in small groups (Müller et al., 2018). Lastly, *action units* (AU) that describe specific facial expressions, in particular AU 05 (upper lid raiser), 17 (chin raiser), 20 (lip stretcher), 25 (lips part), are also included as they were useful in automated dyadic conversational analyses to detect depression in our previous work (Kim et al., 2019b). All features introduced in this paragraph were calculated using the AU and landmark positioning features extracted using OpenFace (Baltrušaitis et al., 2016).

Thirdly, for the family of *prosody*, *delay* is chosen because it has been shown to be an indicator of doctor-to-patient influence – patients of low rapport with their doctors were found to speak less in response to doctor's comments (Sexton et al., 1996). *Speech rate* is chosen because doctor's fluent speech rate and patient's confident communication have been positively correlated with the patient's perception of rapport (Hall et al., 2009). Delay and speech rate are calculated using the time-stamped transcripts. *Tone* is chosen because a warm and respectful tone on the part of both doctor and patient is positively correlated with the patient's perception of rapport (Hall et al., 2009). Tone is calculated using the Vokaturi algorithm (version 3.3) (Vokaturi, 2019).

Fourthly, for the family of *semantics*, *sentiment* is chosen because the provision of positive regard from a practitioner to a patient is an important factor to foster therapeutic alliance; additionally, this process may be further enhanced if the patient also demonstrates positive behaviors towards the practitioners (Farber and Doolin, 2011). Sentiment is extracted using the VADER algorithm (Gilbert, 2014), in line with Sen et al. (2017). *Questions* is chosen because higher engagement by the doctor (e.g., asking questions) with the patient and the patient asking fewer questions have been shown to positively correlate with the patient's perception of rapport (Hall et al., 2009). Questions are detected using Stanford CoreNLP Parser (Manning et al., 2014) and the Penn Treebank (Bies et al., 1995) tag sets.

Next, *mimicry* is chosen because doctor-patient synchrony is an established proxy for rapport. In a review paper, rapport is theorized to be grounded in the coupling of practitioner's and patient's brains (Koole and Tschacher, 2016). Such a coupling process would eventuate in various forms of mimicry in the dyad, for instance, vocally (e.g., matching speech rate and tone), physiologically (e.g., turn-taking, breathing), physically (e.g., matching body language) (Wu et al., 2020). In this study, we aim to use vocal mimicry to capture this underlying phenomenon. Session level mimicry scores are approximated through Dynamic Time Wrapping distances (Giorgino and others, 2009), in line with Müller et al. (2018).

Lastly, *history* is chosen because the scores given by the assessors could be subjective evaluations where the evaluations are unduly influenced by the assessor's leniency bias (Moers, 2005). We attempted to mitigate the leniency bias by introducing history features that indicate the assessor's leniency and its consistency.

4.3 Generation of *coarse* multimodal narrative

In this section, we discuss the *coarse* multimodal narrative. We summarized the automatic generation of the text representation in Table 2.

We calculated the z-score for all the above templates (except Template 3 which is categorical) using the following z-score formula. The average (μ), and standard deviation (σ) are computed using observations from the training observations. Using the z-score, we bucketed them into "very low" (z<-2), "low" (z<-1), "high" (z>1) and "very high"

Family	Child	ID	Template
Demo graphics	Talkativeness	1	**doctor** number of words **high**, **doctor** number of distinct words high
	Big 5 Personality	2	**doctor** openness **high**
	Gender	3	The **patient** is **female**
Actions	Laughter	4	**doctor** laughter counts **high**
	Head Nodding	5	**doctor** head nod counts **high**
	Forward Trunk Leaning	6	**doctor** forward trunk leaning **high**
	Smiling	7	**doctor** smiling counts **high**
	PosiFace	8	**doctor** positive face expression counts **high**
	AU	9	**doctor** minimum lip depressor **very low**, maximum lip depressor **low**, average lip depressor **low**, variance lip depressor **low**
Prosody	Delay	10	minimum delay **very low**, maximum delay **low**, average delay **low**, variance delay **low**
	Speech rate	11	speech rate **high**
	Tone	12	angry tone **high**
Semantics	Sentiment	13	positive sentiment **high**
	Questions	14	open questions **high**
Mimicry	Speech Rate	15	speech rate mimicry **high**
	Tone	16	tone mimicry **high**
History	Num. Sessions	17	patient number of sessions before this **very high**
	Proportion given extreme marks	18	patient question four proportion given maximum marks **high**

Table 2: Templates for the session-level *coarse* summary.

(z>2). The reason for z-transformation is to create a human-readable text through bucketing continuous variables into easy-to-understand buckets ("high" vs. "low").

$$z = \frac{x - \mu_{\text{Train}}}{\sigma_{\text{Train}}} \quad (1)$$

4.4 Generation of *fine* multimodal narrative

In addition to the verbatim transcript, we introduced two new families of information – *prosody*, and *actions*. Table 3 gives an overview of the templates, and the bold-face indicates a variable. The motivations of the features have been discussed; we discuss the rules of insertion in the next few paragraphs.

Template 19 is the verbatim transcript returned from the ASR system. Before each talk-turn, we identified the speaker (doctor/patient) and added multimodal information using templates 20-29. Speech rate and tone were standardized across all training observations. We appended template 20, 21 where possible values are dependent on the z-score – "quickly" ($1 <$ z-score < 2) and "very quickly" (z-score ≥ 2). For delay, we used time intervals of 100 milliseconds, and between 200 and 1200 milliseconds – in line with Roberts and Francis (2013). We appended template 22 at the front of the talk-turn if a delay of at least 200 milliseconds is present between talk-turns. In addition, we appended template 23 where possible values are dependent on the standardized duration of delay – "short" ($<$ 1 z-score), "long" ($<$ 2 z-score) and "significantly long" ($\geq$ 2 z-score). Template 23 captures longer than usual delay, considering the unique turn-taking dynamics of each conversation. The standardized duration of delay is calculated using talk-turn delays from the respective session. Lastly, as for the actions family, templates 24 – 28 were added if any of the actions are detected during the talk-turn. For template 29, it was only added if the AU is detected throughout the entire duration of the talk-turn.

5 Experimental settings

There are two main types of inputs – (1) numeric inputs at the session-level, and (2) *coarse* and/or *fine*

Family	Child	ID	Template
Verbatim	Transcript	19	Transcript returned from the ASR system
Prosody	Speech rate	20	the doctor **quickly** said
	Tone	21	the doctor said **angrily**
	Delay	22	after **two** hundred milliseconds
		23	a **long** delay
Actions	Laughter	24	the doctor **laughed**
	Nodding	25	the doctor **nodded**
	Forward trunk learning	26	the doctor **leaned forward**
	Smiling	27	the doctor **smiled**
	PosiFace	28	the doctor **displayed positive facial expression**
	AU05, 17, 20, 25	29	the doctor **exhibited lip depressor**

Table 3: Templates for the talkturn-level *fine* summary.

multimodal narrative text inputs. As an overview, for (1), we trained the decision tree classifier using session-level numeric inputs. As for (2), we trained the HAN (Yang et al., 2016). We aim to facilitate how humans analyze conversations – HAN can work with text and has easy interpretation with single-headed attention, making it a suitable candidate. Relative to BERT (Devlin et al., 2018), the HAN is faster to train and easier to interpret.

5.1 Research questions

The proposed features have been motivated by scientific studies in Section 4. A natural next question is, "what are the impacts of these proposed features on model performance?" We break this broad question into three questions.

Firstly, (**Q1**) do the newly added features improve performance over the existing set of features for the classification tree and/or HAN?

Secondly, modelling using unstructured text input data (as opposed to using numeric inputs) has the risk of introducing too much variability in the inputs. Therefore, we investigate (**Q2**) – given the *coarse*-only inputs, do the performance between the HAN and classification tree differ significantly?

Lastly, adding more granular talkturn-level inputs to the *coarse* session-level inputs has the benefit of deeper analyses, because it allows the analyst to analyze important talkturns of the conversation. On top of this benefit, (**Q3**) do we also have significant performance improvement between *coarse*-only vs. both *coarse* and *fine* inputs?

For all models, the area under the receiver-operator curve (AUC) was used as the evaluation metric. The AUC measures the goodness of ranking (Hanley and McNeil, 1982) and therefore does not require an arbitrary threshold to turn the probabilities into classes. The partitioning of the dataset to the five-folds is constant for decision tree and HAN to facilitate comparison. The five folds are created through stratified sampling of the dependent variable.

5.2 Classification tree set-up

To answer (**Q1**) and (**Q2**), we tested for all 72 configurations of prime ($2^3 = 8$) plus full ($2^6 = 64$) family inputs for the decision tree. We performed the same z-transformation pre-processing (as in section 4.3) on the decision tree input variables and limited random search to twenty trials.

The algorithm used is from the *rpart* package with R. As part of hyperparameter tuning, we tuned the cp (log-uniform between 10^{-7} to 10^{-9}), maximum depth (uniform between 1 to 20), and minimum split (uniform between 20 to 80) through five-fold cross-validation and random search.

5.3 HAN set-up

To answer (**Q1**) and (**Q2**), we chose the input configurations that performed that best for the classification tree, and used the same input configurations in HAN to compare the difference. Therefore, this test is biased in favour of the classification tree. To answer (**Q3**), we added the *fine* narratives to each *coarse*-only configuration, and compared the difference.

The model architecture is the HAN architecture by Yang et al. (2016), with about 5 million parameters. We used the pre-trained Glove word embeddings (Pennington et al., 2014) of 300-dimensions to represent each word. Words not found in the Glove vocabulary are replaced with the "unk" to-

Coarse Inputs	Tree	Coarse-only (HAN)	Significance of Difference (Coarse-only vs. Tree)	Coarse + Fine (HAN)	Significance of Difference (Coarse + Fine vs. Coarse-only)
$D'A'P'$ (Existing Features)	0.577 (0.011)	0.637 (0.018)	^^^ [0.038, 0.082]	0.629 (0.041)	[-0.054, 0.038]
H	0.613 ** (0.036)	0.642 (0.038)	[-0.025, 0.083]	0.652 (0.048)	[-0.053, 0.073]
DH	0.670 *** (0.049)	0.670 ** (0.034)	[-0.062, 0.062]	0.654 (0.030)	[-0.063, 0.031]
PAH	0.684 *** (0.022)	0.645 (0.043)	[-0.089, 0.011]	0.661 (0.029)	[-0.038, 0.070]
$APMH$	0.664 *** (0.037)	0.643 (0.036)	[-0.074, 0.032]	0.657 (0.037)	[-0.039, 0.067]
$APSMH$	0.649 *** (0.021)	0.644 (0.049)	[-0.060, 0.050]	0.653 (0.051)	[-0.064, 0.082]
$DAPSMH$	0.630 *** (0.032)	0.661 * (0.030)	[-0.014, 0.076]	0.650 (0.028)	[-0.053, 0.031]

Table 4: Summary of the model performances. We report the average five-fold cross-validation AUC and its standard deviation in brackets. Row-wise: We begin with the $D'A'P'$, which is the full existing feature set from Kim et al. (2019b), and progressively compare it against the new sets of features to answer **Q1**. Column-wise: We compare the difference in AUC between the classification tree and *coarse*-only HAN to answer **Q2**. We compare the difference in AUC between the *coarse*-only HAN and *coarse* + *fine* HAN to answer **Q3**. Asterisks (*) indicate significance relative to the $D'A'P'$ row. Carets (^) indicate significance relative to column-wise comparisons, we also provide the confidence intervals in square brackets [] for the difference in performance. The number of symbols indicate the level of statistical significance, e.g., ***: 0.01, **: 0.05, *: 0.10.

ken. The hyperparameter tuning procedure is reported in Appendix A, and the best hyperparameter configurations are reported in Appendix B. There are twenty hyperparameter search trials for each input configuration[3].

6 Experimental results

The results are summarized in Table 4. The key findings are: (**Q1**) with the extended inputs, we observed statistically significant improvements in both the HAN and tree over the existing full set of features (one-tailed *t*-test); (**Q2**) given the *coarse*-only inputs, the performances between the HAN and classification tree did not differ significantly (two-tailed *t*-test), therefore it is plausible that feature engineering into text features do not risk performance; (**Q3**) although adding the *fine* narratives allow deeper analyses by the analyst, it does not lead to significant differences over the *coarse*-only inputs (two-tailed *t*-test).

(**Q1**) When compared to the full set of existing features, the classification tree achieved statistically significant improvements (at $\alpha = 0.05$) in all six out of six *coarse* input families. For HAN, it achieved statistically significant improvements in one (at $\alpha = 0.05$) or two (at $\alpha = 0.10$) out of six *coarse* input families. This demonstrates the value of the newly introduced *coarse* features[4].

(**Q2**) Across the seven *coarse* input configurations, there are no significant differences in the performance from the classification tree when compared to the HAN in six out of seven input configurations. The only exception is in the baseline $D'A'P'$ configuration where the HAN is significantly better. However, the lack of statistically significant differences does not mean that the performances are the *same*. In line with Quertemont (2011) recommendation, we provided the confidence interval around the difference in performance for discussion. Of all confidence intervals that included zero in the fourth column of Table 4, the

[3]We conducted additional tuning experiments for the tree in Appendix C to observe potential improvements in performance.

[4]We performed additional tests in Appendix D to observe the impact of the additions to the *fine* narratives, and found small improvements (but statistically insignificant) in all three out of three input families (va, vp, vpa).

confidence intervals do not suggest that that the effect sizes are negligible (for example, less than 0.01). In summary, we cannot conclude that the performance of HAN differs significantly from tree nor are they the same.

(**Q3**) The addition of *fine* narratives to the *coarse* narrative did not result in significantly stronger (nor weaker) performance in any of the seven input configurations. We posit that this negative finding is due to the difficulty in prioritizing the back-propagation updates to the parts of the network interacting with the *coarse* features, where there is likely a high signal-to-noise ratio. Despite the negative finding, we think it is important to explore fine features' addition onto coarse features because it produces a complete transcript for the human to understand how the conversation proceeded.

6.1 Qualitative Analysis

We visualized the talkturn-level and word-level attention weights from the model. Attention weights are normalized using z-transformation and bucketed into four buckets (< 0, < 1, < 2, ≥ 2) (Kim et al., 2019b). The analyst could analyze an important segment in detail (as in Fig. 3) or see an overview of the important segments in the conversation (see appendix E). In the example (Fig. 3), we observed that the multimodal annotations of leaning forward and positive expression were picked up as important words by the model.

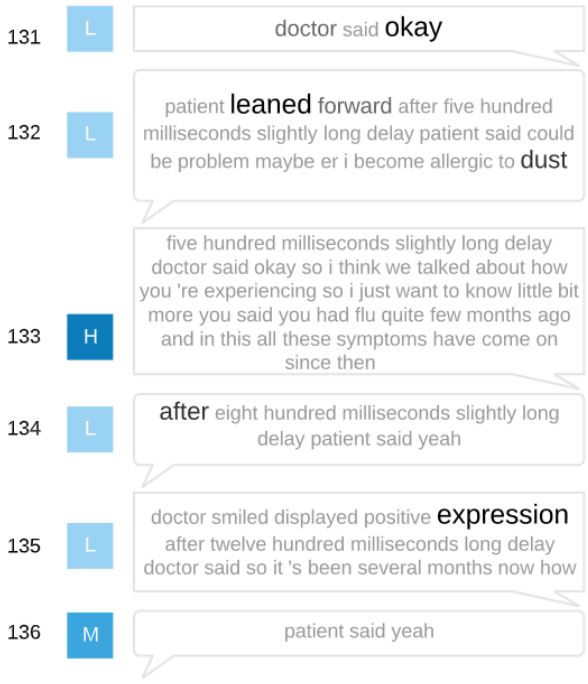

Figure 3: Conversation analysis for a true positive. The talkturn-level attentions are labelled Low (L), Medium (M) and High (H), while the words with higher attention have a larger and darker font. We also transcribed this segment using the Jefferson system in Appendix F.

7 Conclusion

In this paper, we build upon a fully text-based feature-engineering system. We motivated the added features with existing literature, and demonstrated the value of the added features through experiments on the EQClinic dataset. This approach emulates how humans have been analyzing conversations with the Jefferson (2004) transcription system, and hence is human-interpretable. It is highly modular, thereby allowing practitioners to inject modalities. In this paper, we have used a wide range of modalities, including *demographics*, *actions*, *prosody*, *mimicry*, *actions*, and *history*. The ablation tests showed that the added *coarse* features significantly improve the performance for both decision tree and HAN models.

Future research could (1) investigate whether this feature engineering system is generalizable to wider applications of conversational analysis; (2) conduct user studies to validate the usability and ease of interpretability of the visualization.

Acknowledgments

We acknowledge the Sydney Informatics Hub and the University of Sydney's high-performance computing cluster, Artemis, for providing the computing resources and Marriane Makahiya for supporting the data manipulation work. Video data collection was carried out as part of the OSPIA platform project, funded by the Department of Health Clinical Training Fund from the Australian Government.

References

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for Treebank II style Penn Treebank project. *University of Pennsylvania*, 97:100.

Joseph N Cappella. 1990. On defining conversational

coordination and rapport. *Psychological Inquiry*, 1(4):303–305.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Patricio Costa, Raquel Alves, Isabel Neto, Pedro Marvao, Miguel Portela, and Manuel Joao Costa. 2014. Associations between medical student empathy and personality: a multi-institutional study. *PloS one*, 9(3).

Paul T Costa Jr, Antonio Terracciano, and Robert R McCrae. 2001. Gender differences in personality traits across cultures: robust and surprising findings. *Journal of personality and social psychology*, 81(2):322.

Gaëtan Cousin and Marianne Schmid Mast. 2013. Agreeable patient meets affiliative physician: how physician behavior affects patient outcomes depends on patient personality. *Patient Education and Counseling*, 90(3):399–404.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

M Robin DiMatteo. 1979. A social-psychological analysis of physician-patient rapport: toward a science of the art of medicine. *Journal of Social Issues*, 35(1):12–33.

Lawrence D Egbert, George E Battit, Claude E Welch, and Marshall K Bartlett. 1964. Reduction of postoperative pain by encouragement and instruction of patients: a study of doctor-patient rapport. *New England Journal of Medicine*, 270(16):825–827.

Florian Eyben, Martin Wöllmer, Michel F Valstar, Hatice Gunes, Björn Schuller, and Maja Pantic. 2011. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *Face and Gesture 2011*, pages 322–329. IEEE.

Barry A Farber and Erin M Doolin. 2011. Positive regard. *Psychotherapy*, 48(1):58.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Toni Giorgino and others. 2009. Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software*, 31(7):1–24.

Judith A Hall, Debra L Roter, Danielle C Blanch, and Richard M Frankel. 2009. Observer-rated rapport in interactions between medical students and standardized patients. *Patient Education and Counseling*, 76(3):323–327.

James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

Colin Hesse and Emily A Rauscher. 2019. The relationships between doctor-patient affectionate communication and patient perceptions and outcomes. *Health communication*, 34(8):881–891.

Dirk Heylen, Elisabetta Bevacqua, Marion Tellier, and Catherine Pelachaud. 2007. Searching for prototypical facial feedback signals. In *International Workshop on Intelligent Virtual Agents*, pages 147–153. Springer.

Yuri Hosoda. 2006. Repair and relevance of differential language expertise in second language conversations. *Applied linguistics*, 27(1):25–50.

Gail Jefferson. 2004. Glossary of transcript symbols with an introduction. *Pragmatics and Beyond New Series*, 125:13–34.

Lee Jeong-Hwa and Kyung-Whan Cha. 2020. An analysis of the errors in the auto-generated captions of university commencement speeches on youtube. *Journal of Asia TEFL*, 17(1):143.

Joshua Kim, Rafael A Calvo, Kalina Yacef, and N J Enfield. 2019a. A Review on Dyadic Conversation Visualizations - Purposes, Data, Lens of Analysis. *arXiv preprint arXiv:1905.00653*.

Joshua Y Kim, Greyson Y Kim, and Kalina Yacef. 2019b. Detecting depression in dyadic conversations with multimodal narratives and visualizations. In *Australasian Joint Conference on Artificial Intelligence*, pages 303–314. Springer.

Joshua Y Kim, Chunfeng Liu, Rafael A Calvo, Kathryn McCabe, Silas CR Taylor, Björn W Schuller, and Kaihang Wu. 2019c. A comparison of online automatic speech recognition systems and the nonverbal

responses to unintelligible speech. *arXiv preprint arXiv:1904.12403*.

Sander L Koole and Wolfgang Tschacher. 2016. Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance. *Frontiers in psychology*, 7:862.

Suzanne M Kurtz and Jonathan D Silverman. 1996. The Calgary-Cambridge Referenced Observation Guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. *Medical education*, 30(2):83–89.

Chunfeng Liu, Renee L Lim, Kathryn L McCabe, Silas Taylor, and Rafael A Calvo. 2016. A web-based telehealth training platform incorporating automated nonverbal behavior feedback for teaching communication skills to medical students: A randomized crossover study. *Journal of Medical Internet Research*, 18(9).

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Valerie Lynn Manusov. 2014. *The sourcebook of nonverbal measures: Going beyond words*. Psychology Press.

Robert R McCrae. 2017. *The Five-Factor Model across cultures.* Praeger/ABC-CLIO.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *AAAI*, pages 1359–1367.

Frank Moers. 2005. Discretion and bias in performance evaluation: the impact of diversity and subjectivity. *Accounting, Organizations and Society*, 30(1):67–80.

Ryan R Montague. 2012. Genuine dialogue: Relational accounts of moments of meeting. *Western Journal of Communication*, 76(4):397–416.

Robert J Moore. 2015. Automated transcription and conversation analysis. *Research on Language and Social Interaction*, 48(3):253–270.

Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting low rapport during natural interactions in small groups from non-Verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*, pages 153–164. ACM.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Paulo Quaglio. 2008. Television dialogue and natural conversation: Linguistic similarities and functional differences. *Corpora and discourse: The challenges of different settings*, pages 189–210.

Etienne Quertemont. 2011. How to statistically show the absence of an effect. *Psychologica Belgica*, 51(2):109–127.

Felicia Roberts and Alexander L Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–EL477.

Kimiko Ryokai, Elena Durán López, Noura Howell, Jon Gillick, and David Bamman. 2018. Capturing, Representing, and Interacting with Laughter. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 358. ACM.

Fabio Sala, Edward Krupat, and Debra Roter. 2002. Satisfaction and the use of humor by physicians and patients. *Psychology and Health*, 17(3):269–280.

Albert E Scheflen. 1964. The significance of posture in communication systems. *Psychiatry*, 27(4):316–331.

Taylan Sen, Mohammad Rafayet Ali, Mohammed Ehsan Hoque, Ronald Epstein, and Paul Duberstein. 2017. Modeling doctor-patient communication with affective text analysis. *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, 2018-Janua:170–177.

Harold C Sexton, Kristin Hembre, and Guri Kvarme. 1996. The interaction of the alliance and therapy microprocess: A sequential analysis. *Journal of Consulting and Clinical Psychology*, 64(3):471.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293.

John M Travaline, Robert Ruchinskas, and Gilbert E D'Alonzo Jr. 2005. Patient-physician communication: why and how. *Journal of the American Osteopathic Association*, 105(1):13.

Vokaturi. 2019. Vokaturi Overview.

Kaihang Wu, Chunfeng Liu, and Rafael A Calvo. 2020. Automatic Nonverbal Mimicry Detection and Analysis in Medical Video Consultations. *International Journal of Human–Computer Interaction*, pages 1–14.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*, pages 218–233. Springer.

Appendices

A Tuning procedure

We tuned the SGD optimizer with a learning rate between 0.003 to 0.010, batch size to be between 4 to 20, L2 regularization between 10^{-6} and 10^{-3}, and trained for up to 350 epochs without early stopping. We tuned the number of gated recurrent units (GRU) (Cho et al., 2014) between 40 to 49 in both the word-level and talk-turn-level layers, with both the GRU dropout and recurrent dropout (Gal and Ghahramani, 2016) to be between 0.05 to 0.50. The method of choosing hyperparameters is through uniform sampling between the above-mentioned bounds, except for the learning rate where log-uniform sampling is used. Training is performed on a RTX2070 GPU or V100 GPU.

B Hyperparameter configurations for best-performing models

Table 5 (HAN) and Table 6 (Tree) report the hyperparameter configurations for each of the best-performing model reported in Table 4.

C Performance of additional tuning

We conducted additional experiments on the tree configurations to (1) compare the improvements in performance when tuning the HAN and tree, and (2) evaluated the increase in performance if the tree is allowed twenty more hyperparameters random search trials (Fig. 4).

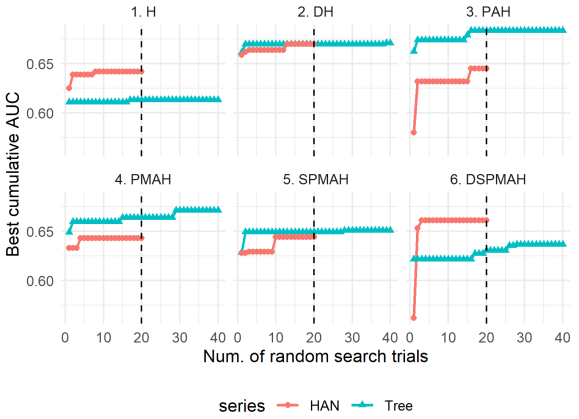

Figure 4: Best cumulative AUC performance given N random search trials.

From the larger increases in HAN performances, it is plausible that HAN is more sensitive to the hyperparameter tuning than the tree.

D Additional tests for additions to the *fine* narratives

Table 7 reports the additional tests on the impact of the added *fine* features. We observe that whilst all three input configurations (va, vp, vpa) have small increases in performance, none of them are statistically significant.

E Conversation thumbnail visualization

By illustrating the talkturn-level attention weights as a heatmap thumbnail (Fig. 5), the analyst could quickly get a sense of the important segments of the conversation without reading the content and zoom-in if required.

F Jefferson example

As an optional reference, we engaged a professional transcriptionist to transcribe the conversation segment presented (Fig. 3) using the Jefferson system. The Jefferson example is presented in Fig. 6. The verbal content is slightly different due to (1) different methods to determine talkturns transitions and (2) automatic speech recognition accuracy.

Config.	Batch Size	Num. of GRU	Learning Rate	GRU dropout	GRU recurrent dropout	L2 regularization	Epoch
H	19	42	0.010	0.10	0.23	1×10^{-4}	223
DH	11	46	0.010	0.07	0.09	3×10^{-6}	74
PAH	14	47	0.005	0.16	0.50	2×10^{-5}	329
$APMH$	8	44	0.005	0.29	0.16	1×10^{-3}	275
$APSMH$	9	43	0.005	0.16	0.48	4×10^{-5}	305
$DAPSMH$	14	41	0.010	0.49	0.48	2×10^{-5}	138
$D'A'P'$	19	46	0.004	0.06	0.50	1×10^{-4}	260
v	16	40	0.009	0.15	0.09	2×10^{-5}	316
va	13	43	0.007	0.13	0.48	1×10^{-6}	347
vp	8	42	0.006	0.13	0.05	2×10^{-5}	310
vpa	9	48	0.010	0.45	0.46	1×10^{-5}	349
va'	12	40	0.006	0.11	0.30	1×10^{-4}	346
vp'	11	42	0.007	0.44	0.19	2×10^{-5}	341
$vp'a'$	10	45	0.008	0.31	0.41	4×10^{-6}	267
H-vpa	8	42	0.005	0.38	0.33	2×10^{-5}	346
DH-vpa	12	44	0.009	0.25	0.14	1×10^{-5}	316
PAH-vpa	11	47	0.005	0.08	0.49	5×10^{-5}	349
$APMH$-vpa	18	46	0.008	0.13	0.50	1×10^{-5}	339
$APSMH$-vpa	9	43	0.010	0.13	0.21	2×10^{-6}	240
$DAPSMH$-vpa	15	46	0.009	0.15	0.50	2×10^{-5}	340
$D'A'P'$ - $vp'a'$	13	46	0.008	0.26	0.16	1×10^{-5}	262

Table 5: Best HAN configurations for the development set.

Config.	Min. split	Max. depth	cp
H	27	17	3.13×10^{-6}
DH	72	18	1.14×10^{-6}
PAH	70	15	8.84×10^{-5}
$APMH$	72	18	1.14×10^{-6}
$APSMH$	68	14	5.26×10^{-5}
$DAPSMH$	37	10	2.94×10^{-5}
$D'A'P'$	68	21	3.74×10^{-5}

Table 6: Best Tree configurations for the development set.

Config.	Existing inputs	New inputs	Significance of Difference (existing vs. new)
v	0.617 (0.053)		N/A
vp'	0.630 (0.037)	0.636 (0.055)	[-0.062, 0.074]
va'	0.616 (0.055)	0.622 (0.033)	[-0.060, 0.072]
$vp'a'$	0.630 (0.038)	0.648 (0.027)	[-0.030, 0.066]

Table 7: Summary of the model performances for the fine narratives. We report the average five-fold cross-validation AUC and its standard deviation in brackets. Row-wise, we begin with the v configuration to show the impact of *fine* multi-modal annotations over the verbatim transcript. Then, we show the impact of the additions (**Q1**) over the existing fine annotations from Kim et al. (2019b) using column-wise comparisons. Asterisks (*) indicate significance relative to the v row. Carets (^) indicate significance relative to column-wise comparisons, we also provide the confidence intervals in square brackets [] for the difference in performance. The number of symbols indicate the level of statistical significance, e.g., ***: 0.01, **: 0.05, *: 0.10.

Figure 5: Heatmap thumbnail. Darker blue indicates higher talkturn attention weights.

DOCTOR:	Ok::ay (0.2) and so you would say that i-those
	things are all worse when you're exerting yourself?
	(0.5)
PATIENT:	.hhh ↑↑ye::h↑↑ <I think so::> >I think so< .h but
	also you know maybe because I'm doing household
	chores and ↑stuff that could be: (0.2) a problem
	maybe I-I've become allergic to ↑↑↑dust?
	(0.5)
DOCTOR:	Oka:y (0.5) so:: (0.5) I ↑think we've talked about
	how you're experiencing >so I just ↑wanna know a
	little bit more.=You said you've had the (0.2) the
	flu: quite few months ago↑: .h and of this- all
	these symptoms have come on since the↑::n.
	(0.2)
PATIENT:	[Yeah.]
DOCTOR:	[Um::] (0.7) so it's been several months no:w.
	[0.5]
PATIENT:	[Yeah_]
DOCTOR:	[How] long did you have the flu: (.) period for.

Figure 6: Jefferson transcription example. : (colon) - stretched sound; (0.2) - a pause of 0.2 seconds; .hhh - in breath, .h - short in breath; ↑ - Rise in intonation; underline - emphasis; <> - slowed speech rate, >< - quickened speech rate; [] - overlapping speech.