Automated pitch and volume annotations for multimodal textual transcriptions

Anonymous ACL submission

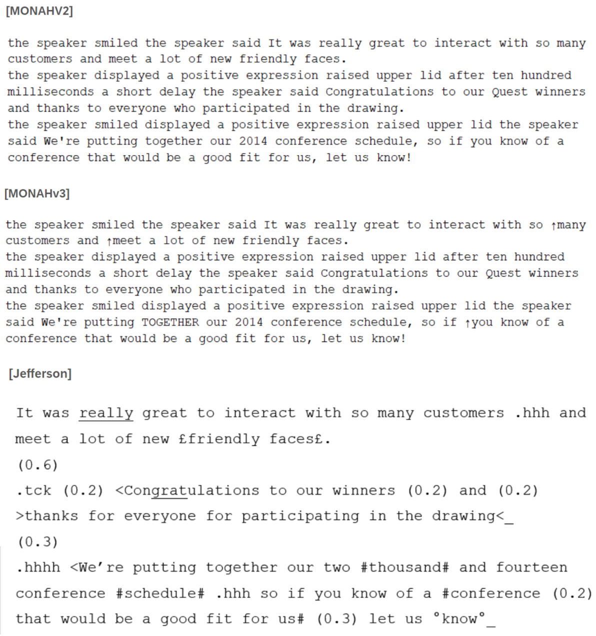
Abstract

Multimodal annotations add important cues to understand *how* a conversation proceeded. In this paper, we further extend the automated conversation annotation system MONAH with *pitch* and *volume* annotations to become the state-of-the-art automatic annotation system in terms of the number of aspects being annotated automatically. MONAHv3 provides an automated solution that is competitive against the widely used, manual Jefferson transcription system. In automatic evaluations, the additions significantly improves supervised learning in ten out of fifteen experiments. With human evaluations to guess the emotions, the additions significantly outperformed the Jefferson transcription system. In terms of usability, human evaluations also showed that the system is significantly more usable than the Jefferson system. Lastly, human evaluations also indicated that the additions significantly improved paralinguistics (describing tone and volume) annotations over MONAHv2, elevating MONAHv3 to be comparable with Jefferson in paralinguistics. MONAHv3 is already and remains more competitive in kinesics (describing actions).

1 Introduction

Multimodal annotations have emerged as a useful tool in various fields of study, like conversational analysis and human language learning and comprehension. For example, Boers et al. (2017); Patel and Furr (2011) both found that multimodal annotations – including both pictorial and verbal elements – significantly aid in reading fluency. Kim and Yacef (2023b) also identified other possible use-cases, such as (1) to be able to revisit a past conversation on paper, (2) to quickly sift through multiple conversations using search, (3) to improve the conversation accessibility to the hearing or visually impaired, and (4) to reduce individual annotator's biases.

Specifically, pitch and volume annotations help understand the prosodic context. Cole (2015) performed a comprehensive review on the role of prosody to the pragmatic context, expressing elements of meaning that lie "beyond words". This includes (1) focus, (2) identification of statement, question, or acknowledge, (3) signalling uncertainty or sarcasm, and (4) manage talker turn changes and convey cohesion.



```
[MONAHV2]

the speaker smiled the speaker said It was really great to interact with so many
customers and meet a lot of new friendly faces.
the speaker displayed a positive expression raised upper lid after ten hundred
milliseconds a short delay the speaker said Congratulations to our Quest winners
and thanks to everyone who participated in the drawing.
the speaker smiled displayed a positive expression raised upper lid the speaker
said We're putting together our 2014 conference schedule, so if you know of a
conference that would be a good fit for us, let us know!

[MONAHv3]

the speaker smiled the speaker said It was really great to interact with so ↑many
customers and ↑meet a lot of new friendly faces.
the speaker displayed a positive expression raised upper lid after ten hundred
milliseconds a short delay the speaker said Congratulations to our Quest winners
and thanks to everyone who participated in the drawing.
the speaker smiled displayed a positive expression raised upper lid the speaker
said We're putting TOGETHER our 2014 conference schedule, so if ↑you know of a
conference that would be a good fit for us, let us know!

[Jefferson]

It was really great to interact with so many customers .hhh and
meet a lot of new £friendly faces£.
(0.6)
.tck (0.2) <Congratulations to our winners (0.2) and (0.2)
>thanks for everyone for participating in the drawing<_
(0.3)
.hhhh <We're putting together our two #thousand# and fourteen
conference #schedule# .hhh so if you know of a #conference (0.2)
that would be a good fit for us# (0.3) let us °know°_
```

Figure 1: A side by side comparison of MONAHv2, the current addition of pitch and volume to MONAHv3, and Jefferson transcription.

Given the importance of pitch and volume annotations, we extended the works of Kim et al. (2021) (MONAHv2) by adding annotations on speech pitch and volume to release MONAHv3 [1]. A comparison of the output is in Fig. 1. Table 1 summarises the various aspects of automated annotation produced by each system. This paper creates MONAHv3 by adding annotations at the sub-word level, whilst previously MONAHv2 only had talk-turn level annotations. These contributions

[1] https://github.com/provideAfterReview

Group	Aspect	Moore (2015)	Kim et al. (2021)	Umair et al. (2022)
verbatim(v)	Speech-to-text conversion	Yes	Yes	Yes
verbatim(v)	Speaker identification	None	Yes	Yes
prosody(p)	Phonetic representation (e.g., uhhh)	None	None	None
prosody(p)	Silence	Yes	Yes	Yes
prosody(p)	Audible breath	None	None	None
prosody(p)	Laughter	None	Yes	Yes
prosody(p)	Speech tempo	None	Yes	Yes
actions(a)	Facial expression	None	Yes	None
actions(a)	Body forward leaning	None	Yes	None
actions(a)	Head nodding	None	Yes	None
prosody(p)	Speech pitch	None	Contribution	None
prosody(p)	Speech volume	None	Contribution	None

Table 1: This paper extends upon the works of Kim et al. (2021) by adding annotations on speech pitch and volume (marked as "contribution").

improve the state-of-the-art in terms of the number of nonverbal aspects automatically annotated.

2 Background

The first attempts at automated multimodal annotations started with (Moore, 2015), where the authors attempted to automate annotations related to silences. Although it was the one of the first attempts at automated annotations, its annotation followed an established system of Jefferson transcription (Jefferson, 2004), which is still widely used by the linguistics community to annotate conversations manually.

Following the advances in deep learning, that brought about better tools for facial landmark recognition and speech processing. Umair et al. (2022) improved upon Moore (2015) and added speaker identification, laughter and speech tempo annotations. Both systems followed the established system of Jefferson transcription.

Kim et al. (2021) introduced MONAHv2, which further adds facial expression, body forward learning and head nodding annotations. Kim and Yacef (2023a) conducted a user survey to understand the relative strengths of MONAHv2 and the manually transcribed Jefferson transcript, and found that (1) video recordings outperformed all forms of transcripts, (2) MONAHv2 was more usable than Jefferson, (3) Jefferson was stronger in paralinguistics (pitch/volume).

3 Data

3.1 Dataset

We test our new multimodal annotations on the MOSI and MOSEI (Zadeh and Pu, 2018; Zadeh et al., 2016). MOSEI contains 23,453 annotated video segments spoken by 1000 distinct people across 250 topics. MOSI contains 3,702 annotated video segments spoken by 89 distinct people (Zadeh et al., 2016). For the automatic evaluation, we used the same partitions as provided by Zadeh and Pu (2018) and Zadeh et al. (2016). Both datasets are in English.

We also took steps to deidentify individual people by using the Named-Entity Recognition system (Peters et al., 2017) to replace people names with the "person-hashid" token at the word-level so that we preserve the number of words in the transcipt and annotations can be preserved at the word-level.

3.2 Dependent variables

For our purposes of supervised learning, each segment is annotated for its sentiment from -3 (highly negative, to +3 (highly positive) and multi-label emotion annotation across 6 classes – happiness, sadness, anger, fear, disgust, and surprise.

4 Multimodal features extraction

4.1 Description of pitch and volume annotations

In this paper, we contribute towards word-level pitch and volume annotations. The annotations are inserted in the same way as the Jefferson transcription system (Jefferson, 2004). We first discuss pitch annotations before volume annotations.

There are two types of pitch annotations, the up (down) arrow signifies the high (low) pitch respectively. The arrow can occur before or after the word, depending on when the change in pitch occurred. For example, if the speaker asks “would you like coffee?" with a higher pitch at the second syllabus of “coffee", the annotation would be “would you like coffee↑?" Pitch annotations are important for emotion recognition and sentiment analysis. Mairesse et al. (2012) found that analysis on pitch alone, without any text information significantly outperformed the baseline.

For volume, there is only one type of annotation, upper-cased words signify the louder words. Unlike pitch annotations, the upper-casing occurs at the word level, depending on the average loudness computed across the duration of the word. For example, if the speaker asks “how dare you?" with a louder “dare", the annotation would be “how DARE you?" Volume annotations are important for emotion recognition and sentiment analysis (Tzirakis et al., 2017).

4.2 Description of pre-existing annotations

We used the Google Speech-to-text service to obtain the transcript for each recording. Unfortunately, phonetic representation (e.g. “uhhhhh") is not possible as only the correctly spelt form (“uh") is returned, along with word-level timestamps. With the word-level timestamps, MONAHv2 uses a set of predefined rules to insert phrases that describe the talkturn. We used the open-source system to generate the features. As for Jefferson transcript, we have employed Jefferson Transcription Specialists to transcribe the video snippets.

4.3 Generation of pitch and volume annotations

Stage 1: Obtaining pitch and volume raw data In the first stage, we first split the audio file into multiple audio files, each consisting of one talkturn. The start and end time of each talk turn is supplied in the MOSEI dataset. We then send all talkturn audio files to Google Speech-to-text API to obtain the word-level timestamps. With the start time of each word, we can derive the duration for each word by using either the start time of the next word or the end time of the talkturn (if the word is the last word of the talkturn). Duration at the word level is fine for volume as we annotate at the word level. However, for pitch annotations, it is crucial to know whether the first or last syllable has significantly higher or lower pitch. We have used the Carnegie Mellon Pronouncing Dictionary from NLTK (Bird et al., 2009) to map words into the number of syllables. We then divided the word-level duration by the number of syllabus to obtain the syllabus-level duration. Finally, with the volume and pitch data extracted using OpenSMILE (Eyben et al., 2010), we calculate the average volume (pitch) using the word- (syllable-) level duration.

Stage 2: Compute z-score The z-score transformation helps the algorithm identify pitch and volume variations within the same video. We calculate the z-score for each word (or syllable) using the following formula, where x is the volume (pitch) of the word (syllable), μ is the average volume (pitch) of all words (syllables) in the same video, and σ is the standard deviation of volume (pitch) of all words (syllables) in the same video.

$$z = \frac{x - \mu_{Video}}{\sigma_{Video}} \qquad (1)$$

Stage 3: Generate text narrative from z-score Having computed the z-scores for pitch and volume, we insert unicode up(down) arrows to represent high(low) pitch if the absolute value of the z-score is above the threshold (say, 2.0). Similarly, for volume annotations, we captitalize the word if the z-score is above the threshold. Fig 4 (Appendix A.1) shows the sensitivity analysis of changing the thresholds.

5 Experimental settings

5.1 Research questions

We answer the following research questions using both automatic and human evaluation. Considering volume annotations and pitch annotations individually and collectively:

Q1: Do the annotations improve supervised learning?

Q2: Do the annotations improve human performance in guessing the emotions?

Q3: Do the annotations change the perceived usability?

Q4: Do the annotations change the perceived thoroughness of the three aspects of nonverbal annotation? The three aspects of nonverbal annotations are (1) Chronemics: the use of pacing of speech and length of silence, (2) Kinesic: body movements or postures, and (3) Paralinguistic: volume, pitch, and quality of voice.

For both automatic and human evaluations, two baselines are used: (B1) the unannotated verbatim transcripts, and (B2) the transcripts of the previous version MONAHv2 since this paper extends it.

5.2 Automatic evaluation

5.2.1 Applied models

In automatic evaluation, we test whether the annotation additions help improve supervised learning in three tasks. Sentiment prediction in MOSI (Zadeh et al., 2016) and MOSEI (Zadeh and Pu, 2018), and multilabel emotion prediction in MOSEI. The following models are selected to test the annotation additions. In appendix A.3, we found that the tokenizer behavior of BERT and DistilBERT were the same, but differ from DistilRoberta, when given a range of capitalization and arrow annotation variants.

1. Cased BERT: Bidirectional Encoder Representations from Transformer (Devlin et al., 2018) is a widely used baseline in the NLP community.

2. Cased DistilBERT (Sanh et al., 2019): built via the knowledge distillation technique (Buciluă et al., 2006; Hinton et al., 2015) on BERT.

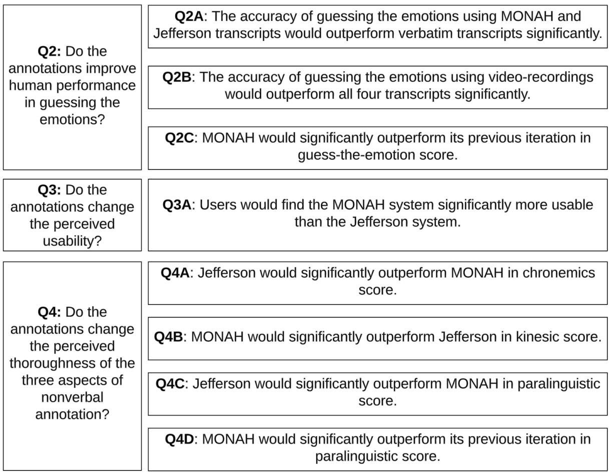
3. Cased DistilRoberta. A distilled version of RoBERTa (Liu et al., 2019).

In our ablation tests, we have performed hyper-parameter tuning over the following range (see Appendix Table 6) for each input configuration. We used a random search to tune the hyper-parameters, and each configuration was given 15 trials. We picked the hyper-parameters with the best development set performance and reported their test set performance in this paper. As for the input thresholds for pitch-only (P), volume-only (V) and pitch plus volume (PV), we increased the threshold in increments of 0.5 from 0 to 2.0 inclusive.

5.2.2 Evaluation metrics

We compare the automatic evaluations using MAE for the sentiment task in both MOSEI and MOSI. As for the multi-label emotions task, we use the average class-wise weighted accuracy. We bootstrapped 1000 samples to compute statistical significance.

5.3 Human evaluation



Q2: Do the annotations improve human performance in guessing the emotions?	**Q2A**: The accuracy of guessing the emotions using MONAH and Jefferson transcripts would outperform verbatim transcripts significantly.
	Q2B: The accuracy of guessing the emotions using video-recordings would outperform all four transcripts significantly.
	Q2C: MONAH would significantly outperform its previous iteration in guess-the-emotion score.
Q3: Do the annotations change the perceived usability?	**Q3A**: Users would find the MONAH system significantly more usable than the Jefferson system.
Q4: Do the annotations change the perceived thoroughness of the three aspects of nonverbal annotation?	**Q4A**: Jefferson would significantly outperform MONAH in chronemics score.
	Q4B: MONAH would significantly outperform Jefferson in kinesic score.
	Q4C: Jefferson would significantly outperform MONAH in paralinguistic score.
	Q4D: MONAH would significantly outperform its previous iteration in paralinguistic score.

Figure 2: The list of the eight hypotheses tested under the three areas.

To demonstrate MONAHv3's competitiveness against the Jefferson transcripts, we adapted a shorter version of Kim and Yacef (2023a), testing eight hypotheses in three areas (**Q2** - **Q4**). The hypotheses are summarized in Fig. 2. Hypotheses group **Q2** is associated with whether the additional multimodal annotations improves guess-the-emotion siginicantly over not having the annotations (**Q2A**). Since multimodal annotations are a video-to-text compression, video recordings would have significantly higher guess-the-emotion accuracy (**Q2B**). Since we added the pitch and volume annotations, MONAHv3 should outperform the MONAHv2 (**Q2C**). Hypotheses group **Q3** is associated with only whether users would find the MONAHv3 system significantly more usable than the Jefferson system with the addition of pitch symbols and capitalization (**Q3A**). Hypotheses group **Q4** is associated with the three aspects of multimodal annotation (chronemics, kinesic, paralinguistic). Since Jefferson transcripts annotate within-talkturn delays, Jefferson would significantly outperform MONAH (v2 and 3) in chronemics score (**Q4A**). As for kinesic score, MONAH (v2 and 3) should outperform Jefferson transcripts as MONAH (v2 and 3) annotates more actions than Jefferson (**Q4B**). As for paralinguistics, Jefferson has a lot more types of symbolic annotations, like phonetic representations and symbols that indicate an elongation of a syllabus, therefore Jefferson should outperform MONAHv3 (**Q4C**). However, since MONAHv3 made a significant improvement in its pitch and volume annotations, it should outperform MON-

AHv2 (**Q4D**). We selected the threshold 2.0 for both pitch and volume and generated the narratives for the user study. As for Jefferson transcripts, we used the gold standard, human manually generated transcripts from a professional transcriber.

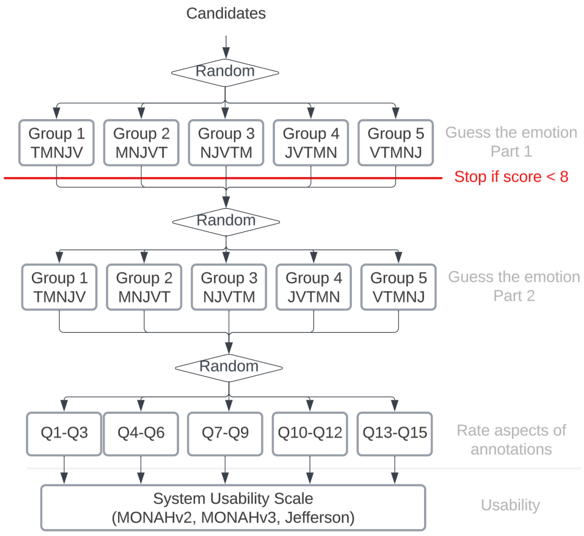


Figure 3: Experiment design. The first two phases are about guessing the emotions, where participants are randomly placed into one of the five groups so that the order in which the transcripts are presented are shuffled. T: Verbatim, M: MONAHv2, N: MONAHv3, J: Jefferson, V: Video.

Our experimental design comprises a sequence with four stages in the user study (Fig. 3). We provide a quick introduction to the Jefferson transcript in the beginning (see Appendix A.4) before starting the first stage. In the first stage, the participant attempts to guess the emotions of fifteen snippets. Three of each of the five types of question is shown to the participant. We controlled for question difficulty by fixing the set of fifteen questions across the five groups. In each group, participants were exposed to three questions of each transcript variant.

If the participant gets at least seven out of fifteen correct, the participant enters part two. In part two, the participant again attempts to guess the emotions of another fifteen snippets, the questions of part one do not overlap with part two. The setup of part two is similar to part one, where the set of fifteen questions is fixed across the five groups, each group was exposed to three questions of each transcript variant. The participant needs to get at least six out of fifteen correct to proceed on to the next stage, the aspects rating stage.

The aspects rating stage consists of Likert-scale questions that ask the users to rate the thoroughness of each of the three aspects of nonverbal annotation (Chronemics, Kinesics, Paralinguistics). The set of 15 questions is divided into five groups, so each group answers a set of three non-overlapping question. Each of these three questions has nine subparts each, because there were three transcripts (MONAH v2 and 3, plus Jefferson), and there were three aspects (Chronemics, Kinesics, Paralinguistics). The phrasing of the question is, “The amount of information is sufficient to interpret how the talkturn is being said, for (1) Use of pacing of speech and length of silence in conversation; (2) Body movements or postures; (3) Volume, pitch and quality of voice."

The last stage is the System Usability Scale stage which is a set of ten questions for each system (Brooke, 1996). First, the participant gets a refresher on MONAH v2 and 3, and on the Jefferson transcript. Then, for each of the three transcripts types, the participant answers the ten questions on the Likert scale. The list of questions is detailed in Appendix A.5.

We administrated the survey on Amazon Mechanical Turk, and the hosting of the survey was on Qualtrics. The base reward for the survey is 0.5 USD. A 2.50 USD bonus is awarded if the participant is able to get at least seven correct answers out of the 15 questions in the first guess-the-emotion section. The survey is terminated if the participant could not get at least seven correct answers in the first section. An additional 2 USD bonus is awarded if the participant is able to get at least six correct answers out of the fifteen questions in the second guess-the-emotion section. The ethics approval number is 123456 (masked for review anonymity). In total, 616 workers received the base reward, 62 workers received the 2.50 USD bonus, and 64 workers received the 4.50 USD bonus. The total cost of administering the survey is 1088 USD, including MTurk fees and taxes, and compensating 12 workers 5 USD each due to errors in the survey setup. Workers who have received any base, bonus or compensation are added to the MTurk exclusion list to prevent multiple attempts. On the Qualtrics end, we have also followed the recommended best practices[2] to prevent fraud responses.

When extracting the data for analysis, we discarded participants with score less than 7 for either

[2] https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/

	Baseline: Verbatim						Baseline: MONAH					
	MOSEI			MOSI			MOSEI			MOSI		
	DB	BT	DR	DB	BT	DR	DB	BT	DR	DB	BT	DR
Baseline	0.551	0.548	0.519	0.851	0.826	0.714	0.537	0.540	0.522	0.909	0.800	0.720
+P(0.0)	0.583	0.562	0.544	1.003	0.996	0.815	0.577	0.558	0.543	1.025	0.981	0.854
+P(0.5)	0.568	0.567	0.547	0.931	0.902	0.843	0.548	0.544	0.540	0.972	0.905	0.770
+P(1.0)	0.554	0.540	0.535	0.941	0.840	0.815	0.540	0.550	0.524	0.929	0.900	0.820
+P(1.5)	0.543	0.547	0.529	0.873	0.856	0.754	0.542	0.527*	0.514	0.861	0.812	0.755
+P(2.0)	0.548	0.537*	0.511	0.866	0.809	0.709	0.534	**0.518***	0.519	0.815	0.807	**0.708**
+V(0.0)	0.646	0.613	0.550	1.118	0.991	0.863	0.634	0.609	0.541	1.179	1.083	0.759
+V(0.5)	0.581	0.570	0.535	1.021	0.922	0.734	0.577	0.572	0.529	1.032	0.956	0.737
+V(1.0)	0.555	0.550	0.515	0.930	0.876	0.722	0.550	0.539	0.520	0.922	0.869	0.735
+V(1.5)	0.555	**0.536***	0.516	0.904	0.796	0.717	0.546	0.533	0.512	0.875	0.780	0.732
+V(2.0)	**0.540***	**0.536***	**0.510**	0.855	**0.785***	**0.696**	**0.531**	0.534	**0.511**	0.847	0.857	0.710
+PV(0.0)	0.647	0.648	0.595	1.209	1.097	0.981	0.639	0.648	0.587	1.163	1.118	1.004
+PV(0.5)	0.610	0.592	0.580	1.056	1.037	0.927	0.599	0.583	0.557	1.091	1.059	0.887
+PV(1.0)	0.577	0.562	0.554	0.934	0.885	0.823	0.564	0.548	0.542	0.967	0.875	0.822
+PV(1.5)	0.555	0.545	0.523	0.881	0.827	0.740	0.543	0.534	0.522	0.870	0.845	0.777
+PV(2.0)	0.545	0.538*	0.514	**0.839***	0.800	0.731	0.540	0.527	0.516	**0.798**	**0.739***	0.777

Table 2: Summary of the model mean-absolute-error for the fine narratives. DB: DistilBERT, BT: BERT, RB: RoBERTA. Best in column are in bold-face. * statistically significant difference with baseline (top row).

	Dataset: MOSEI					
	Baseline: Verbatim			Baseline: MONAH		
Input	DB	BT	DR	DB	BT	DR
Baseline	0.832	0.838	0.838	0.839	0.840	0.844
P(0.0)	0.834	0.838	0.840	0.838	0.836	0.840
P(0.5)	0.837*	0.836	0.839	0.841	0.839	0.841
P(1.0)	0.837*	0.838	0.840	0.840	0.839	0.840
P(1.5)	**0.841***	0.840	0.842*	0.842	0.840	0.843
P(2.0)	0.839*	0.840	0.837	**0.844**	0.841	0.844
V(0.0)	0.834	0.835	0.837	0.836	0.840	0.839
V(0.5)	0.839*	0.834	0.840	0.839	0.837	0.842
V(1.0)	0.834	**0.842***	0.842*	0.841	0.841	**0.845**
V(1.5)	0.839*	0.834	**0.843***	0.840	0.839	0.842
V(2.0)	0.839*	0.839	0.840	0.842	**0.845***	0.845
PV(0.0)	0.831	0.835	0.833	0.836	0.836	0.839
PV(0.5)	0.835*	0.837	0.840	0.840	0.840	0.839
PV(1.0)	0.836*	0.835	0.838	0.840	0.840	0.842
PV(1.5)	0.839*	0.841*	0.838	0.842	0.841	0.843
PV(2.0)	0.835*	0.841	0.839	0.842	0.841	0.843

Table 3: Summary of the classification accuracies for the fine narratives. Best in column are in bold-face. * statistically significant difference with baseline (top row).

of the guess the emotions section to ensure that the response quality is acceptable. When testing the hypothesis, we used a two-sample, one-sided t-test to compare the means.

6 Experimental Results

6.1 Ablation analysis

We first discuss the results of the sentiment regression task shared by both MOSEI and MOSI before discussing the results of the multi-label classification task from MOSEI. Table 2 and 3 summarizes the model performances for the ablation tests. In summary, the addition of pitch or volume annotation (but not necessarily both) improves the performance on the regression task and the classification across all three variants of the Bert models.

Since there are three models, two sentiment tasks (MOSEI, MOSI) and two baselines (Verbatim and MONAHv2), we present 3x2x2=12 columns in Table 2. Out of the 12 combinations, the addition of pitch and volume annotations improved supervised learning significantly in six combinations. Four of the six combinations are significant improvements over the Verbatim baseline, and two are significantly better over the MONAHv2 baseline.

Since there are three models, one classification task (MOSEI) and two baselines (Verbatim and MONAHv2), we present 3x1x2=6 columns in Table 3. Out of the 6 combinations, the addition of pitch and volume annotations improved supervised learning significantly in four combinations. All three combinations under the verbatim baseline are significant, and one is significantly better over the MONAHv2 baseline.

6.2 Human ratings

The results of the hypotheses testing results are summarized in Table 4.

For **Q2A**, MONAH v2 and 3 both outperformed verbatim significantly, but Jefferson did not outperform verbatim significantly. For **Q2B**, the accuracy of guess the emotions using video significantly outperformed Jefferson and verbatim transcript, but not MONAH v2 nor v3. Lastly, for **Q2C**, although the MONAHv3 accuracy (0.558) outperformed the MONAHv2 (0.497), the difference is not significant.

For **Q3A**, users do find the MONAHv3 system significantly more usable than the Jefferson system.

For **Q4A**, Jefferson significantly outperformed MONAHv3 in chronemics score. For **Q4B**, MONAHv3 significantly outperformed Jefferson in kinesics score. For **Q4C**, Jefferson significantly outperformed the MONAHv2, but did not outperform MONAHv3. Lastly, for **Q4D**, MONAHv3 significantly outperformed MONAHv2 in paralinguistic score.

7 Discussion

7.1 Ablation analysis

For supervised learning, we observed that the impact of the pitch and volume annotations are more significant over the Verbatim baseline (4 sentiment combinations, and 3 classification combinations significant) than it is over the MONAHv2 baseline (2 sentiment combinations, and 1 classification combination significant). This suggests that pitch and volume annotations are more valuable when added to a transcript without any nonverbal annotations. As MONAHv2 have other nonverbal talkturn-level annotations, the additional information added by pitch and volume annotations is less significant.

In addition, we observed that by model, BERT has a higher count of significant differences (6) as compared to DistilBERT (3) or DistilRoBERTA (1). This is interesting because the tokenizer of BERT and DistilBERT behaves identically, as seen in the Appendix A.3. Therefore, the tokenizer is not a reason behind this difference. The plausible reason is that both DistilBERT and DistilRoBERTA are distilled models and have less ability to take advantage of the pitch (up/down arrows) and volume (capital letters) annotations.

7.2 Human ratings

In this section, we will be discussing our three hypotheses areas illustrated in Fig. 2.

Does it improves the user's accuracy in guess-the-emotion? Compared to Kim and Yacef (2023a), where the authors performed a user-study of MONAHv2 did not find a significant difference in **Q2A** with 104 completions. Our study has a larger number of completions (126), and we found statistically significant differences in both MONAH v2 and v3 outperforming the verbatim transcripts. In contrast, the mean of Jefferson transcripts in this study is lower than that of verbatim, which could be explained by the low usability score and the users did not understand the annotations.

As for **Q2B**, previously, the authors found that video is significantly better than all variants of transcripts, but in this paper, we found that video is only significantly better than the Jefferson and the verbatim transcript. This is interesting because the video file encodes a lot more information compared to the pure textual MONAH (v2 or v3) transcripts.

Is each system easy to use? Previously, MONAHv2 was found to be more usable than the Jef-

Hypothesis	CLG	BSE	Avg CLG	N CLG	Std CLG	Avg BSE	N BSE	Std BSE	p-value
Q2A *The accuracy of guess the emotions using MONAH and Jefferson transcripts outperform verbatim significantly.*	j	t	0.456	581	0.498	0.447	591	0.498	0.373
	m2	t	0.526	591	0.500	0.447	591	0.498	0.003*
	m3	t	0.558	581	0.497	0.447	591	0.498	0.001*
Q2B *The accuracy of guess the emotions using video would outperform all transcripts significantly.*	v	t	0.566	581	0.496	0.447	591	0.498	0.001*
	v	j	0.566	581	0.496	0.456	581	0.498	0.001*
	v	m2	0.566	581	0.496	0.526	591	0.500	0.084
	v	m3	0.566	581	0.496	0.558	581	0.497	0.384
Q2C *MONAH would significantly outperform its previous iteration in guess-the-emotion score.*	m3	m2	0.558	581	0.497	0.526	591	0.500	0.14
Q3A *Users would find the MONAH system significantly more usable than the Jefferson system.*	m3	j	58.5	48	13.9	49.7	48	16.1	0.003*
Q4A *Jefferson significantly outperform MONAH in chronemics score.*	j	m3	3.92	720	0.982	3.79	720	1.00	0.007*
Q4B *MONAH significantly outperform Jefferson in kinesic score.*	m3	j	3.75	720	1.10	3.54	720	1.24	0.001*
Q4C *Jefferson significantly outperform MONAH in paralinguistic score.*	j	m2	3.87	720	0.975	3.70	720	1.12	0.001*
	j	m3	3.87	720	0.975	3.95	720	0.944	0.950
Q4D *MONAH significantly outperform its previous iteration in paralinguistic score.*	m3	m2	3.95	720	0.944	3.70	720	1.12	0.001*

Table 4: Summary of hypotheses testing results. CLG: Challenger, BSE: Baseline, j: Jefferson, m3: MONAHv3, m2: MONAHv2, t: Verbatim, v: Video.

ferson system. MONAHv3 remains significantly more usable than the Jefferson system with the addition of one type of symbol from the Jefferson transcription system. In both systems, the up/down arrows denote high/low pitch.

Is each system thorough in annotating nonverbal events? In MONAHv3, we did not change the chronemics and kinesic annotations. Therefore, we expected the previous findings to remain valid in this study, and it did – Jefferson still significantly outperformed MONAHv3 in chronemics (**Q4A**) and MONAHv3 still significantly outperformed Jefferson in kinesic (**Q4B**). The addition of pitch and volume annotations should improve the paralinguistic score, and the results confirmed it. Jefferson was found to outperform MONAHv2, and it continued to outperform MONAHv2 in the current study in **Q4C**. However, Jefferson no longer significantly outperforms MONAHv3 in **Q4C**. Lastly, we observed that the additional annotations in MONAHv3 significantly improved upon the paralinguistics score as MONAHv3 significantly outperformed MONAHv2 (**Q4D**).

8 Conclusion

In this paper, we produce a state-of-the-art system in terms of the number of nonverbal aspects being annotated, and have automatically annotated prosody related annotations, elevating the thoroughness of human-rated paralinguistics score to be comparable to the manual, and time-consuming Jefferson transcripts. In addition, automated supervised learning and manual human ratings have improved significantly from these additions, demonstrating the importance of our additions.

Future works could focus on chronemics where Jefferson still significantly outperforms MONAHv3. As for limitations, there are many other analysis lenses such as toxicity, lie, or sacarsm detection that could be explored in future works. The risks of this paper include the over-generalization of findings, e.g., improving the accessibility to vision or hearing impaired users require more specific user testing.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Frank Boers, Paul Warren, Gina Grimshaw, and Anna Siyanova-Chanturia. 2017. On the benefits of multimodal annotations for vocabulary uptake from reading. *Computer Assisted Language Learning*, 30(7):709–725.

John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, 189(3):189–194.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Jennifer Cole. 2015. Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1-2):1–31.

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*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network (2015). *arXiv preprint arXiv:1503.02531*, 2.

Gail Jefferson. 2004. Glossary of transcript symbols. *Conversation analysis: Studies from the first generation*, pages 24–31.

Joshua Y Kim and Kalina Yacef. 2023a. An empirical user-study of text-based nonverbal annotation systems for human–human conversations. *International Journal of Human-Computer Studies*, page 103082.

Joshua Y Kim and Kalina Yacef. 2023b. Guidelines for designing and building an automated multimodal textual annotation system. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, ICMI '23 Companion, page 330–336, New York, NY, USA. Association for Computing Machinery.

Joshua Y Kim, Kalina Yacef, Greyson Kim, Chunfeng Liu, Rafael Calvo, and Silas Taylor. 2021. Monah: Multi-modal narratives for humans to analyze conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 466–479.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

François Mairesse, Joseph Polifroni, and Giuseppe Di Fabbrizio. 2012. Can prosody inform sentiment analysis? experiments on short spoken reviews. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5093–5096. IEEE.

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

Rupal Patel and William Furr. 2011. Readn'karaoke: Visualizing prosody in children's books for expressive oral reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3203–3206.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309.

Muhammad Umair, Julia Beret Mertens, Saul Albert, and Jan P de Ruiter. 2022. Gailbot: An automatic transcription system for conversation analysis. *Dialogue & Discourse*, 13(1):63–95.

Amir Zadeh and Paul Pu. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Long Papers)*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

A Appendices

A.1 Z threshold sensitivity analysis

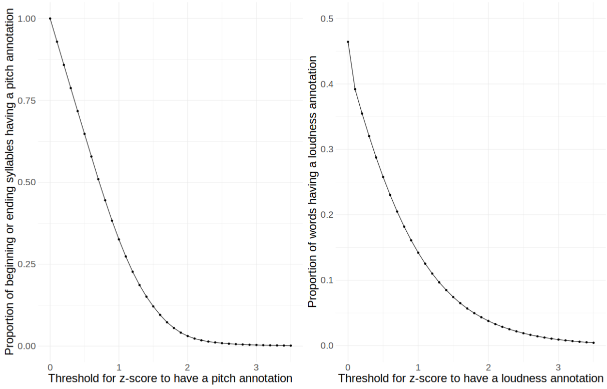


Figure 4: Sensitivity analysis on the impact of changing the threshold on the percentage of text being annotated. Left: Since pitch annotations can be added to the first or last syllabus of the word, we consider all beginning and ending syllabus as candidates. We observe that when the threshold is zero, all syllables have a pitch annotation. Right: Since loudness annotations are added at the word-level, and we only annotate loud words, and not soft words, we observe that when the threshold is zero, nearly half (50%) of the words have a loudness annotation.

A.2 Tokenizer behavior

Since we are adding up and down arrow sumbols as well as capitalising to the characters, it would be important to study the effects of these additions on the behaviour of the tokenizer, these are illustrated in Table 5.

A.3 Hyperparameter tuning

We used a random search to tune the hyperparameters, and each configuration was given 15 trials. In each trial, we pick a random value that is listed in Table 6.

A.4 Jefferson Reference

Table 7 is provided to all participants to help them familiarize themselves with the Jefferson Transcription System.

A.5 System Usability Scale Questions

Figure 5 details the set of the ten system usability scale questions.

Table 5: Tokenization outcomes with different annotations by pre-trained models. BERT and DistilBERT have the same tokenization outcomes. For ROBERTa and Reformer, we added the randomly initialized up and down arrows tokens embeddings.

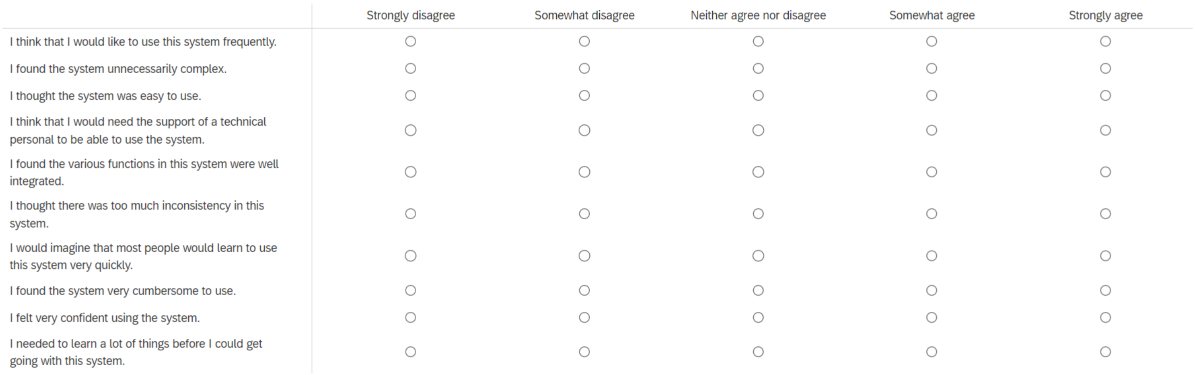
	Tokens from tokenizer		
Input	**BERT**	**DistilBERT**	**Distil ROBERTa**
happy days	happy, days	happy, days	happy, Ġdays
↑happy days	↑, ##ha, ##ppy, days	↑, ##ha, ##ppy, days	↑, happy, Ġdays
happy↓ days	happy, ##↓, days	happy, ##↓, days	happy, ↓, Ġdays
↑happy↓ days	↑, ##ha, ##ppy, ##↓, days	↑, ##ha, ##ppy, ##↓, days	↑, happy, ↓, Ġdays
HAPPY days	H, ##AP, ##P, ##Y, days	H, ##AP, ##P, ##Y, days	H, APP, Y, Ġdays
↑HAPPY days	↑, ##HA, ##PP, ##Y, days	↑, ##HA, ##PP, ##Y, days	↑, H, APP, Y, Ġdays
HAPPY↓ days	H, ##AP, ##P, ##Y, ##↓, days	H, ##AP, ##P, ##Y, ##↓, days	H, APP, Y, ↓, Ġdays
↑HAPPY↓ days	↑, ##HA, ##PP, ##Y, ##↓, days	↑, ##HA, ##PP, ##Y, ##↓, days	↑, H, APP, Y, ##↓, Ġdays

Table 6: Hyperparameter tuning

Hyper parameter	**Min.**	**Max.**	**Scale**
Context Size	2	10	Linear
Learning Rate	1e-6	1e-4	Loglinear
Batch Size	8	18	Linear
Warmup Ratio	0.0	0.5	Linear

Symbol	Defintion and use
[yeah] [ok]	Overlapping talk
=	End of one TCU and beginning of next begin with no gap/pause in between (sometimes a slight overlap if there is speaker change). Can also be used when TCU continues on new line in transcript.
(.)	Brief interval, usually between 0.08 and 0.2 seconds
(1.4)	Time (in absolute seconds) between end of a word and beginning of next. Alternative method: “none-one-thousand-two-one-thousand…”: 0.2, 0.5, 0.7, 1.0 seconds, etc.
Word [first letter underlined] Wo:rd [colon underlined]	Underlining indicates emphasis. Placement indicates which syllable(s) are emphasised. Placement within word may also indicate timing/direction of pitch movement (later underlining may indicate location of pitch movement)
wo::rd	Colon indicates prolonged vowel or consonant. One or two colons common, three or more colons only in extreme cases.
↑word ↓word	Marked shift in pitch, up (↑) or down (↓). Double arrows can be used with extreme pitch shifts.
WORD	Upper case indicates syllables or words louder than surrounding speech by the same speaker
°word°	Degree sign indicate syllables or words distinctly quieter than surrounding speech by the same speaker. Pre-positioned left carat indicates a hurried start of a word, typically at TCU beginning
word-	A dash indicates a cut-off. In phonetic terms this is typically a glottal stop
>word<	Right/left carats indicate increased speaking rate (speeding up). Left/right carats indicate decreased speaking rate (slowing down)
.hhh	In/out breath. Three letters indicate ‘normal’ duration. Longer or shorter inbreaths indicated with fewer or more letters.
whhord	Can also indicate aspiration/breathiness if within a word (not laughter)
w(h)ord	Indicates abrupt spurts of breathiness, as in laughing while talking
£word£	Pound sign indicates smiley voice, or suppressed laughter
word	Hash sign indicates creaky voice
word	Tilde sign indicates shaky voice (as in crying)
(word)	Parentheses indicate uncertain word; no plausible candidate if empty
(( ))	Double parentheses contain analyst comments or descriptions

Table 7: The Jefferson Transcription System.



	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I think that I would like to use this system frequently.	○	○	○	○	○
I found the system unnecessarily complex.	○	○	○	○	○
I thought the system was easy to use.	○	○	○	○	○
I think that I would need the support of a technical personal to be able to use the system.	○	○	○	○	○
I found the various functions in this system were well integrated.	○	○	○	○	○
I thought there was too much inconsistency in this system.	○	○	○	○	○
I would imagine that most people would learn to use this system very quickly.	○	○	○	○	○
I found the system very cumbersome to use.	○	○	○	○	○
I felt very confident using the system.	○	○	○	○	○
I needed to learn a lot of things before I could get going with this system.	○	○	○	○	○

Figure 5: System Usability Scale