Guidelines for designing and building an automated multimodal textual annotation system

Anonymous ACL submission

Abstract

Automatic multimodal textual annotation systems have the potential to advance research in conversational analysis and computational linguistics; however, there is little research in this area. In this short paper, we outline guidelines for designing an automated multimodal textual annotation system. We discussed the guidelines using three guiding questions. First, we share the four possible use-cases for the purpose of building an automated multimodal annotation system. Second, we share five guidelines for collecting and extracting multimodal data. Lastly, we share another four guidelines regarding what to present to the users.

1 Introduction

Multimodal textual annotations to verbatim transcripts are a distinctive addition to describe the nonverbal cues together with *what* is being said. Vivid nonverbal annotations help the user understand the conversation better and, in some cases, make a conversation accessible to someone with a disability. To help future researchers design an automated multimodal annotation system, we compiled a list of guiding questions so that the design exploration is more complete.

Automatic multimodal transcription systems are a new breed of transcription systems that have not received much attention. That said, manual multimodal transcription systems have been an area of active research in the linguistics field. With advances in Computer Vision and Speech Processing deep learning systems, we observe that this new breed of multimodal transcription is beginning to emerge (Moore, 2015; Umair et al., 2022; Kim et al., 2021b). There are many potential areas to explore in the system design and build. In this paper, we frame the guidelines into three sequential questions: (1) What is the purpose of designing and building the automated multimodal annotation system? (2) How to gather data and extract multimodal features from the raw input(s)? (3) What information to present, and what is the appropriate method of presentation?

2 Related Works

Moore (2015) investigated the potential of automated transcription for the use in Conversation Analysis; At the time of writing, an automated transcription tool could only do word recognition and silence, but not speaker segmentation, overlapping talk, phonetic representation, and prosody. Research in automated speech analysis has attempted to address these shortcomings separately. Speaker diarization helps separate different people's voices so that the transcription is marked with the correct speaker. Some approaches have used multimodal approaches that integrate both computer vision with automated speech recognition modalities to identify the correct source (Ephrat et al., 2018; Gao et al., 2018). There are also new prosody identification tools that have been open-sourced, such as Vokaturi (2019).

Outside of the dimensions mentioned by Moore (2015), tools that perform automated gestures detection (Marcos-Ramiro et al., 2013; Noori et al., 2019) and eye contact (or gaze) (Marcos-Ramiro et al., 2013; Noori et al., 2019) have been explored within the field of Computer Vision, but have not translated into multimodal annotation applications. In light of the rapid development of the aforementioned separate tools, some researchers attempted to integrate them together into a multimodal annotation system. Gailbot (Umair et al., 2022) is one example of an automated multimodal annotation system. Its name is closely associated with Gail Jefferson, who published the seminal Jefferson transcription glossary (Jefferson, 2004). This system is widely used in the linguistics field to annotate multimodal cues in a conversation. MONAH (Kim et al., 2019a, 2021b) is another example that uses video or audio inputs to weave in non-technical

phrases that describe the talkturn. In this paper, we focus on the use of annotation systems by humans for interpreting and analysing conversations.

3 What is the purpose of designing and building the automated multi-modal annotation system?

Multimodal transcription is an interpretative act. The vocabulary and granularity of the multimodal description are heavily influenced by the motivation of transcription. Who is the target user? What would the target user like to be aware of? How many speakers do we expect in the conversations? We explore plausible motivations through four different use cases.

Use Case 1: To recall a conversation on paper. Certain healthcare professionals, such as psychologists, would need to revisit earlier conversations while in session with the client. In such a setting, it would be more feasible to print the previous conversations and their multimodal annotations for the psychologist to glance through without access to a computer. That way, the psychologist can maintain a presence in the conversation. In addition to healthcare settings, multimodal annotations have been applied to document high-stakes conversations such as police interrogations (Stokoe, 2010) and courtroom settings (Komter, 2013) (see appendix A.1 and A.2).

Use Case 2: To quickly sift through multiple conversations. Customer service team leaders often have multiple customer service agents under their lead, charged with the responsibility of maintaining the quality of the conversations the team is having with the company's customers. Here, multimodal annotations could help the team lead sample audio recordings that suggest poor tone or rapport, allowing team leads to surface problematic cases more efficiently. For example, the use of the word "well" after a long silence is a well-known phenomenon of "dispreferred" talk-turn format (Pomerantz, 1984).

Conversation skills trainers in medical school would also face similar pain points (Liu et al., 2016). Given a large class of medical students who would like to improve their conversation skills through multiple practice sessions, a more efficient search can be made possible through multimodal annotations.

Use Case 3: To improve the accessibility of the conversation. The Audio Description Service is an additional audio track broadcast that describes visual elements for those who are blind or visually impaired (Mikul, 2010). Similarly, multimodal annotations are textual descriptions of visual non-verbal cues in a conversation. For example, "the boy *smiled* and said, 'how are you?'" If one is designing for this use case (visually impaired), one would need to convert the smiling event into another modality, say, a *ding* sound (Shi et al., 2019), or haptic vibration (Krishna et al., 2010).

On the other hand, if the use case is to design for the hearing impaired, the selected multimodal annotations would be different. Shiver and Wolfe (2015) found that deaf students preferred to watch videos with automated captioning than with no captions. It would be reasonable to extend automated captioning to annotate cues such as out-breaths (sighs), elongation of syllabus, and pitch for the deaf user to appreciate the prosody context better.

Use Case 4: Two-pass approach to reduce individual biases. Manual annotation, because it is an interpretive act, might contain unconscious bias that stems from the annotator's culture and background (Billig, 1999). To combat the individual bias, it might be a good idea to have the first pass done automatically (Bokhove and Downey, 2018) and do a manual second pass in high-stakes annotations. When the first pass is done by an automated system, biases stemming from system settings (rules-based generation) or training data (supervised learning based generation) will be in the annotations. The second human pass will minimise individual confirmation bias and system bias.

4 How to gather data and extract multimodal features from the raw input(s)?

Human detection and annotation of multimodal features is a significantly time-consuming process. It is estimated that for every hour of recording, it takes about four to five hours of human effort to transcribe the multimodal annotations (Walford, 2001; Punch and Oancea, 2014). Furthermore, depending on the answer to the previous question on the motivation(s) for building the multimodal annotation system, the required quality of the multimodal annotations could vary. Based on our experience and the literature, we present the following guidelines for collecting and extracting multimodal information from the unstructured data recording.

Guideline 1: Use one audio recording channel

for each speaker. Although humans can easily separate voices for each speaker, ASR, and diarization systems at this point do not perform as well as humans (Kim et al., 2019b; Prokopalo et al., 2022). By using one channel for each speaker in recordings, overlapping speech is cleaner to be analyzed, and delays between talkturns and the construction of talkturns can be done more accurately (Umair et al., 2022). In our experience of attempting speaker diarization from one channel, there were two problems – (1) overlapping speech is identified and assigned as one person's speech, and (2) transcription of the overlapped speech into words had poor accuracy.

Guideline 2: Calibrate captures such as camera angles and microphone distances. Algorithms that extract multimodal signals may be trained on a well-calibrated collection of videos; therefore, non-calibrated captures may result in poor multimodal feature extraction. Use-cases that involve a one-to-one video conference have less variables to control for. But, in settings that involve a physical space as well as multiple speakers, conversational data capture would have additional considerations, such as using a microphone array (Sun and Canny, 2012), a 360-degree camera (McIlvenny, 2018), or multiple cameras (Xiong and Quek, 2005).

Guideline 3: Use multiple existing research datasets to minimize blind spots dealing with multimodal datasets Before spending the effort to collect the dataset from scratch, the experience of applying the extraction algorithms on existing datasets could help inform the pitfalls to avoid when assembling a new dataset. Poria et al. (2017) has put together a valuable collection of multimodal datasets. For example, when we used the IEMOCAP dataset (Busso et al., 2008) to extract multimodal features, we realized that capturing the actors at a 45 degrees angle made it difficult to discern between head nodding and shaking (appendix A.3) because the nose landmark exhibited both vertical and horizontal displacement in both cases. In the multimodal setting, there are many areas that could go wrong – from the bit rate, dimensions, the encoding of the file, large file storage requirements, audio-visual synchronization, or even physical lighting. Applying the multimodal extraction algorithm on multiple datasets helps ensure the system is resilient as each dataset would usually have many of these variables controlled.

Guideline 4: Resumable processes are important for slow multimodal extraction algorithms. Multimodal extraction algorithms are slow because they are disk-intensive. For example, to extract the timestamped facial landmarks coordinates from the video, the OpenFace (Baltrusaitis et al., 2018) algorithm extracts a still image file for every frame, runs the landmark analysis, and collates the data over every still image extracted (see appendix A.4). Therefore, we recommend ensuring the process could be resumed upon failure without restarting the extraction process, reducing the time required to complete the process. In our experience, the compute time for visual landmarks extraction is just as long as the recording length. Having to restart the visual extraction because some other modules have failed would be a significant time-sink.

Guideline 5: Consider the ethical implications of the annotations When training the system to generate textual annotations, the training annotations could be sourced from crowd-sourcing platforms or hire experts. Crowd-sourcing platforms are not suitable for sensitive conversations (e.g., police interrogations) unless the content can be anonymized. For anonymization, we recommend trained named entity recognition models like Peters et al. (2017) to identify person names and hashing the names at the word level so that pitch and volume annotations at the word level would not be affected. On the other hand, hired annotators can be appropriate for sensitive content where simple name anonymization is not sufficient. Named annotators can be included in the ethics application upon screening, which improves accountability and correctness of the interpretations – an important consideration for high-stakes conversations.

5 What information to present and what is the appropriate method of presentation?

On the one hand, in the Linguistics community of researchers, besides the Jefferson transcription system (Jefferson, 2004), there are a few variants (Mondada, 2018; Rossano et al., 2009; Streeck, 1993) developed that annotates *what* is being said together with *how* it is being said. For example, Mondada (2018) annotates temporal bodily actions by spatially aligning the annotations with the verbal transcript (Appendix A.5).

On the other hand, there is a general lack of tex-

tual multimodal transcription systems in the Computer Visualization community of researchers. Kim et al. (2021a) conducted a review of sixteen different systems that visualizes dyadic conversations and found that there was only one study that visualized the multimodal context *together* with the textual representation of what is being said. In this asymmetric state of research between the abundance of manual expert annotations and the lack of computer-automated annotations, we outline four guidelines for deciding what to present to the users.

Guideline 6: Decide on the trade-off between technical and layman annotations. Multimodal systems in the linguistics community are often technical; the annotations are specific, often using symbols to represent a particular event. For example, colons (:) represent a particular syllabus's elongation. Kim and Yacef (2021) reported that while users of the Jefferson system found that the annotations are precise, both students with and without a linguistics background found that the Jefferson system was harder to use compared to annotations that describe the multimodal annotation in layman language. Knowing your audience and the purpose of the system will inform the trade-off between the precise annotations of a technical system (e.g., Jefferson) against layman annotations.

Guideline 7: Choose Rule-Based Generation over Deep Learning Generation if training data is missing or precise control is required. There are two considerations for deep end-to-end learning natural language generation (NLG). Firstly, there is a lack of sufficiently large datasets required for a model to take a video as input and output text, and the lack of the different variants of output texts if we seek precise control over the categories of actions to be annotated. There are datasets (Damen et al., 2020; Shao et al., 2020; Chung et al., 2021) that could help tangentially. These datasets contain temporal action detection tasks (Hutchinson and Gadepally, 2021) (the action sequence is shorter than the video).

Secondly, it is difficult to precisely control the output of a deep learning NLG model. The current degree of control (Keskar et al., 2019; Liu et al., 2021) still does not have the fine-grained controls like having insertion rules over extracted multimodal features. Because of these two limitations over precise control and training data requirement, consider rule-based generation (Moore, 2015; Umair et al., 2022; Kim et al., 2021b) rather than deep learning generation.

Guideline 8: Test early and cheaply with supervised learning. Conducting human user experience research is expensive (Sauro and Lewis, 2016; Law et al., 2014). One could, therefore, consider using supervised learning to investigate whether the newly added annotations help improve performance in a supervised learning task, say, automatic emotion recognition. This ablation test allows the researcher(s) to gain confidence that the newly added features are helpful to algorithms before testing with human users.

In our experience, we saved months of effort by conducting supervised learning before human user tests, considering the months required for internal ethics approval and external liaison with university faculties (about half of which rejected external requests). Supervised learning allowed us to sidestep such efforts and quantified the usefulness of each added annotation and optimize the set of annotations to present during user testing.

Guideline 9: Support supervised learning augmentation with user study. After the contents of the multimodal transcriptions are prepared, researchers could potentially improve the support to the conversational analyst through supervised learning and augment the transcription with machine learning attention (Shen and Lee, 2016), or Shapley value contribution (Nowak and Radzik, 1994) from each word (Legesse et al., 2016), assuming that there is data regarding the target label. The method of presentation feature importance in a conversational analysis setting should be supported with a user study. For example, Kim and Yacef (2021) enlarged the font size of words having high attention weights and reported that users mistook the larger words as words spoken with a louder voice (appendix A.6).

6 Conclusion

An automated multimodal annotation system would advance research in both computational linguistics and linguistics communities. From the intersection of conversational analysis, advances in deep learning models, and system usability research, we proposed a set of guidelines to help researchers design and build an automated multimodal annotation system. The limitation is that we made these proposals from the experience in building and analyzing multi-modal narrative systems.

References

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Michael Billig. 1999. Whose terms? whose ordinariness? rhetoric and ideology in conversation analysis. *Discourse & Society*, 10(4):543–558.

Christian Bokhove and Christopher Downey. 2018. Automated generation of 'good enough'transcripts as a first step to transcription of audio-recorded data. *Methodological innovations*, 11(2):2059799118790743.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. 2021. Haa500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13465–13474.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2020. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Ruohan Gao, Rogerio Feris, and Kristen Grauman. 2018. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53.

Matthew S Hutchinson and Vijay N Gadepally. 2021. Video action understanding: A tutorial. *IEEE Access*.

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

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Joshua Y Kim, Rafael A Calvo, NJ Enfield, and Kalina Yacef. 2021a. A systematic review on dyadic conversation visualizations. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 137–147.

Joshua Y Kim, Greyson Y Kim, and Kalina Yacef. 2019a. 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. 2019b. A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. *arXiv preprint arXiv:1904.12403*.

Joshua Y Kim and Kalina Yacef. 2021. An empirical user-study of text-based nonverbal annotation systems for human-human conversations. *arXiv preprint arXiv:2112.14928*.

Joshua Y Kim, Kalina Yacef, Greyson Kim, Chunfeng Liu, Rafael Calvo, and Silas Taylor. 2021b. 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.

Martha Komter. 2013. 30 conversation analysis in the courtroom. *The handbook of conversation analysis*, page 612.

Sreekar Krishna, Vineeth Balasubramanian, and Sethuraman Panchanathan. 2010. Enriching social situational awareness in remote interactions: insights and inspirations from disability focused research. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1275–1284.

Effie Lai-Chong Law, Paul Van Schaik, and Virpi Roto. 2014. Attitudes towards user experience (ux) measurement. *International Journal of Human-Computer Studies*, 72(6):526–541.

Meshesha Legesse, Gabriele Gianini, and Dereje Teferi. 2016. Selecting feature-words in tag sense disambiguation based on their shapley value. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 236–240. IEEE.

Chunfeng Liu, Karen M Scott, Renee L Lim, Silas Taylor, and Rafael A Calvo. 2016. Eqclinic: a platform for learning communication skills in clinical consultations. *Medical education online*, 21(1):31801.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Alvaro Marcos-Ramiro, Daniel Pizarro-Perez, Marta Marron-Romera, Laurent Nguyen, and Daniel Gatica-Perez. 2013. Body communicative cue extraction for conversational analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE.

Paul McIlvenny. 2018. Inhabiting spatial video and audio data: Towards a scenographic turn in the analysis of social interaction. *Social Interaction. Video-Based Studies of Human Sociality*, 2(1).

Chris Mikul. 2010. Audio description background paper. *Ultimo NSW: Media Access Australia*.

Lorenza Mondada. 2018. Multiple temporalities of language and body in interaction: Challenges for transcribing multimodality. *Research on Language and Social Interaction*, 51(1):85–106.

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

Farzan Majeed Noori, Benedikte Wallace, Md Uddin, Jim Torresen, et al. 2019. A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network. In *Scandinavian conference on image analysis*, pages 299–310. Springer.

Andrzej S Nowak and Tadeusz Radzik. 1994. The shapley value for n-person games in generalized characteristic function form. *Games and Economic Behavior*, 6(1):150–161.

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

Anita Pomerantz. 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shaped.

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.

Yevhenii Prokopalo, Meysam Shamsi, Loïc Barrault, Sylvain Meignier, and Anthony Larcher. 2022. Active correction for incremental speaker diarization of a collection with human in the loop. *Applied Sciences*, 12(4):1782.

Keith F Punch and Alis Oancea. 2014. *Introduction to research methods in education*. Sage.

Federico Rossano, Penelope Brown, and Stephen C Levinson. 2009. Gaze, questioning and culture. *Conversation analysis: Comparative perspectives*, 27:187–249.

Jeff Sauro and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625.

Sheng-syun Shen and Hung-Yi Lee. 2016. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. *Interspeech 2016*, pages 2716–2720.

Lei Shi, Brianna Tomlinson, John Tang, Ed Cutrell, Daniel McDuff, Gina Venolia, Paul Johns, and Kael Rowan. 2019. Accessible video calling: Enabling nonvisual perception of visual conversation cues. In *Computer-Supported Cooperative Work 2019*. ACM, ACM.

Brent N Shiver and Rosalee J Wolfe. 2015. Evaluating alternatives for better deaf accessibility to selected web-based multimedia. In *Proceedings of the 17th international ACM SIGACCESS conference on computers & accessibility*, pages 231–238.

Elizabeth Stokoe. 2010. 'i'm not gonna hit a lady': Conversation analysis, membership categorization and men's denials of violence towards women. *Discourse & Society*, 21(1):59–82.

Jürgen Streeck. 1993. Gesture as communication i: Its coordination with gaze and speech. *Communications Monographs*, 60(4):275–299.

David Sun and John Canny. 2012. A high accuracy, low-latency, scalable microphone-array system for conversation analysis. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 290–300.

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.

Vokaturi. 2019. Vokaturi Overview.

Geoffrey Walford. 2001. *Doing qualitative educational research*. Bloomsbury Publishing.

Yingen Xiong and Francis Quek. 2005. Meeting room configuration and multiple camera calibration in meeting analysis. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 37–44.

A Appendices

A.1 Courtroom Example

```
DA:  She was in a pardying mood wasn't she.
     (2.1)
V:   Yes.
DA:  So were you weren't you.
     (2.0)
V:   Yes.
     (25.5)
```

Figure 1: The following conversational analysis extract (Komter, 2013), marked up with Jefferson annotations, detailed a snippet from a rape trial where the alleged victim (V) was asked a series of questions by the defense attorney (DA). The snippet annotated the 25.5 seconds pause after V confirmed the DA's suggestion that both her friend and she herself were "partying", allowing time for the jury to internalize what is being said.

A.2 Police Interrogation Example

```
P:    What about Jane.=what about abuse towards Jane:.
            (0.8)
S:    (Well-) when she started abusing m:e I believe I
      said some things ba:ck.=Yet when she says he's got
      a right to go where he wants, he's with me: well,
      (.) I believe I said some things back yeh.
            (0.2)
S:    D'you know warramean.
            (0.3)
P:    You threaten 'er at all.
            (0.4)
S:    No I didn't threaten 'er.
            (0.8)
?:    .hhhhh
S: →  I've got no reason to threaten 'e:r, I've never 'it a
      woman in my life.=an' I never will 'it a woman in my life.
            (0.8)
P:    ( ) heard the front door. ((reading from statement))
```

Figure 2: The following conversational analysis extract (Stokoe, 2010), marked up with Jefferson annotations, details a snippet from a police interrogation where the suspect (S) was asked a series of questions by the police officer (P).

A.3 IEMOCAP Example

Figure 3: Camera angle at a 45 degrees angle makes it difficult to discern between head nodding and shaking.

A.4 OpenFace Example

Figure 4: The OpenFace algorithm (Baltrusaitis et al., 2018) first exports each frame into a folder of image files, then run the landmark detection on the static image file to identify the coordinates of each landmark at that specific point in time.

A.5 Mondada Example

```
°well° I will    take from you     some eggs
                  +....+points eggs+,,,+
              †...†turns head-†,,†
```

Figure 5: Example of a Mondada (2018) transcript, where temporal gestures and gaze are aligned with the verbal transcript.

A.6 MONAH Example

Figure 6: The authors intended that the larger font size to indicate higher weight from the machine attention layer (i.e., the classifier made the decision due to these contributions). However, the human users were confusing larger font size to indicate a louder volume.