A Multimodal Inquiry into the Expression of Humor

Adam Hamden
University of Southern California
ahamden@usc.edu

1 Introduction

Humor is an aspect of natural language that is becoming an increasingly relevant area of research. Human-computer interaction and human-robot interaction require a degree of robustness that incorporates all models of human cognitive abilities, including humor (Stock and Strapparava, 2005). Works exploring the effects of humor on communication show the importance of humor on likeability and rapport (Morkes et al., 1999). The question remains, however, whether we can learn humor markers from gesture and prosody to reliably classify humorous intent from audio, text, and video data.

Interactions between humans and computers where jokes were used resulted in participants rating the agent as competent and reported a greater sense of cooperation, demonstrating that humor contributes directly to the likeability of an interface (Morkes et al., 1999). This study further showed that humor did not distract from the communicative intent of the interaction, alleviating any concerns of humor not enhancing a dialogue (Morkes et al., 1999).

As humans, we respond strongly to social stimuli. Being exposed to laughter alone has been shown to be a sufficient stimulus for laughter and smiles (Provine, 1992). Designing models for humor is the key to understanding an aspect of human social interaction that will bridge the gap for robot and computer agents. By empowering agents with humor, we can build more personable agents that embody a social function.

2 Related Work

2.1 Humor Detection from Language

Traditional methods of humor analysis and detection have typically relied on text-based features as input (De Oliveira and Rodrigo, 2015). Table 1 showcases some language-only humor datasets including "16000 One Liners" (Mihalcea and Strapparava, 2005), "Pun of the Day" (Yang et al., 2015), "Ted Laughter" (Chen and Lee, 2017). Humor, however, is an act of great emotional expression—delivery is key. When we discuss delivery, it is important to consider multimodal inputs. The natural extension is to examine features that can be captured through audio and video.

In a study exploring the prosodic and multimodal markers of humor, it is noted that considering pause-based unit, pitch, volume, pause length features for audio and smiling, facial action units features for video are strong indicators of humorous expression (Gironzetti, 2017).

2.2 Humor Detection from Audio

Attempts at incorporating features beyond text started with audio-only models for sarcasm detection (a form of humorous expression) (Rakov and Rosenberg, 2013). The audio features considered here include speaking rate, standard deviation of pitch, and intensity unigrams. While this sarcasm detection approach performed with promising accuracy, the authors note that more work is do be done in classifying instances found to be ambiguous by human annotators, suggesting that speech may not be the only mode of communicating humorous intent (Rakov and Rosenberg, 2013). In addition, the scope of this humor detection is limited to sarcasm which does not offer a generalized model for humorous expression. Sarcasm is a specific, and often biting form of irony typically presented to express contempt; thus, it carries a dual intent that is not wholly humorous.

Other works have employed both audio and text features for the binary humor classification task. Bertero and Fung used a Convolutional Neural Network (CNN) model that combines word-level and acoustic-frame level features in the "Big Bang Theory" dataset. The authors maintain that the intent of the speaker can be informed by variations in

Dataset	size	modality	type
16000 One Liners	32000	*{l}*	joke
Pun of the Day	4846	*{l}*	pun
Ted Laughter	9452	*{l}*	joke
Big Bang Theory	43672	*{l, a}*	tv show
UR-FUNNY	16514	*{l, a, v}*	joke

Table 1: Survey of existing humor datasets from a variety of modalities including language, audio, and video.

pitch, loudness, and intonation. In addition, lexical, syntactic and structural, sentiment, antonyms, and speaker turn features were considered. With promising accuracy on this prediction task, incorporating visual features could enhance the performance as a smile while speaking strongly informs humorous intent (Gironzetti, 2017).

2.3 Humor Detection from Video

Non-verbal information complements a humorous expression just as strongly as verbal information captured through audio and text features. Wendt and Berg note that non-verbal humor based on gestures, facial expressions, or whole body movements has not been the target of HRI-research so far, reinforcing the lack of research into multimodal humor modeling. The authors posit that the long-term aim of the research should be aimed at understanding and defining what type of behavior is viewed as humorous, to what audiences, and in what contexts.

Katevas et al. conducted the Robot Comedy Lab experiment which used a robot to perform a standup comedy routine. The robot's gesture and gaze were manipulated throughout the performance so as to understand their effect on the audience response. Katevas et al. hypothesize that the there may have been difficulty in interpreting the gestures or that gestures do not function in the way they expected them to. This highlights the need to understand the relationship between humor and gesture. While naïve observational methods have been used to generalize how humor is expressed, there is an opportunity to create a refined model for humor gesturing.

The UR-FUNNY dataset presents a multimodal approach for punchline detection (Hasan et al., 2019). Punchline detection is a task used to determine the likelihood of an expression being the punchline to the joke given the context leading up to it. The UR-FUNNY paper uses GloVe word embedding from text (Pennington et al., 2014); COVAREP features from audio (Degottex et al., 2014); facial Action Units (Rosenberg and Ekman, 2020) and rigid/non-rigid facial parameters from video (Baltrušaitis et al., 2016). Hasan et al. design a Multimodal Context Network (MCN) used to learn a multimodal representation of the context that employs a Transformer-based encoder in the UR-FUNNY dataset task (Hasan et al., 2019; Vaswani et al., 2017).

2.4 Gesture Modeling

With the importance of considering gesture in humor expression, works on gesture modeling become especially relevant even in contexts beyond humor.

Ginosar et al. constructs a model to generate plausible gestures from audio speech input for a model fine-tuned on a particular subject. This cross-modal translation task is used to learn how an individual gestures as they speak. Interestingly, they do not include text-based features as input, claiming that previous work incorporating text is used to train gestures for virtual agents with datasets that are curated for a more rigidly defined task (Ginosar et al., 2019). Instead, the in-the-wild analysis used here focuses only on raw audio signals. The authors use a fully convolutional network with an audio encoder and a 1D UNet that maps the 2D log-mel spectrogram of the audio to a temporal stack of pose vectors (Ronneberger et al., 2015; Isola et al., 2017).

Ginosar et al. uncover that a model trained on a different speaker is on average better at predicting gestures when compared to predicting random motion, but still significantly worse than predicting just the median pose of the true speaker; thus implying that there are idiosyncrasies in gesturing (Ginosar et al., 2019). The scope of the gesturing here is general in topic, raising the question as to whether humor gesturing is as idiosyncratic or if general methods for humor gesturing exist commonly among speakers.

3 Dataset

The PATS dataset is an aligned Pose, Audio, Transcript, and Style dataset that contains aligned segments from 25 different speakers (Ahuja et al., 2020; Ahuja et al.; Ginosar et al., 2019). The dataset present a diverse set of styles with speakers' backgrounds including talk show hosts, lecturers, YouTubers, and televangelists.

3.1 Dataset Features

The following set of features are used from the dataset:

Language: The language features used for this task are the fixed BERT embeddings (768-d vectors) for each word in the aligned sentence according to the bert_base_uncased pre-trained model from Hugging Face (Devlin et al., 2018; Wolf et al., 2020). The transcripts are extracted from the original video using Google Automatic Speech Recognition (Chiu et al., 2018). Naturally, there is inherent error in this automatic transcription process, which was estimated to be .29 (Word Error Rate) (Ahuja et al., 2020).

Audio: The acoustic feature used for this task is the log-mel spectrogram (128-d vector) of each interval in the dataset. The log-mel spectrogram is a more salient representation of the audio that could be more useful in learning tasks (Ginosar et al., 2019).

Pose: The pose features used for this task are the 52 2D skeletal joint positions (normalized) from the upper body. These were extracted using OpenPose at 15 frames-per-second (Cao et al., 2017).

3.2 Humor Label

For the humor classification task, we need to generate binary labels for each interval indicating examples of humor and non-humor. The 15 talk show hosts in the PATS dataset are labelled as examples of humor while the remaining 10 (lecturers, YouTubers, televangelists) are annotated as examples of non-humor (Ahuja et al., 2020). Talk show hosts were chosen as examples of humor since they serve to make their audiences laugh and often present topical news material with a humorous spin.

4 Multimodal Humor Classification

4.1 Problem Formulation

The PATS dataset presents us with three modalities which we will represent as $M = \{t, a, p\}$. Each modality is presented in an aligned sequential form (Ahuja et al., 2020).

Each datapoint can be represented as a tuple (S, l), of the features $S = \{S_m : m \in M\}$ and the label $l \in 0, 1$. Since each datapoint represents a 4.3 second interval sampled at 15 frames-per-second, every modality in S is a sequence of length 64. Thus S_t has a dimension of $(64, 768)$, S_a has a dimension of $(64, 128)$, and S_p has a dimension of $(64, 104)$.

4.2 Unimodal Feature Encoding

Each modality in M is individually encoded using a separate LSTM, resulting in three separate representations U_t, U_a, U_p each with hidden size h_t, h_a, h_p, respectively. We can observe this encoding scheme in Figure 1. We then concatenate all three modality representations along the time dimension, yielding a combined representation H with shape $(64, h_t + h_a + h_p)$.

4.3 Multimodal Feature Encoding

The combined unimodal representations are passed into a multimodal feature encoder that attempts to learn some H' that encodes the features along both the time and feature dimension.This scheme is inspired by the network architecture proposed by Hasan et al. for the UR-FUNNY dataset. The multimodal feature encoder uses a Transformer encoder architecture which is able to learn temporal relationships in input sequences through the self-attention mechanism (Vaswani et al., 2017). The output of the Transformer encoder, H' is fed into a simple feed-forward neural network classifier head.

5 Experiments

We conduct several humor classification experiments to better understand which feature or combination of features has the best performance on the binary humor classification task outlined in Section 4.1. The outcome of these experiments may reveal how important pose features are in the humor classification task.

5.1 Baseline

The baseline models were trained with the unimodal and multimodal feature encoding using a subset of the modalities provided. Table 2 details the results of the respective models on the test set. The performance was evaluated using F1 micro score which is calculated as the percentage of predicted labels that match their true values. Using the audio features alone leads to the highest binary accuracy among other feature combinations.

The baseline models were trained on all 25 speakers in the PATS dataset, with the talk show host being annotated as examples of humor, respectively. The models also used the predefined train,

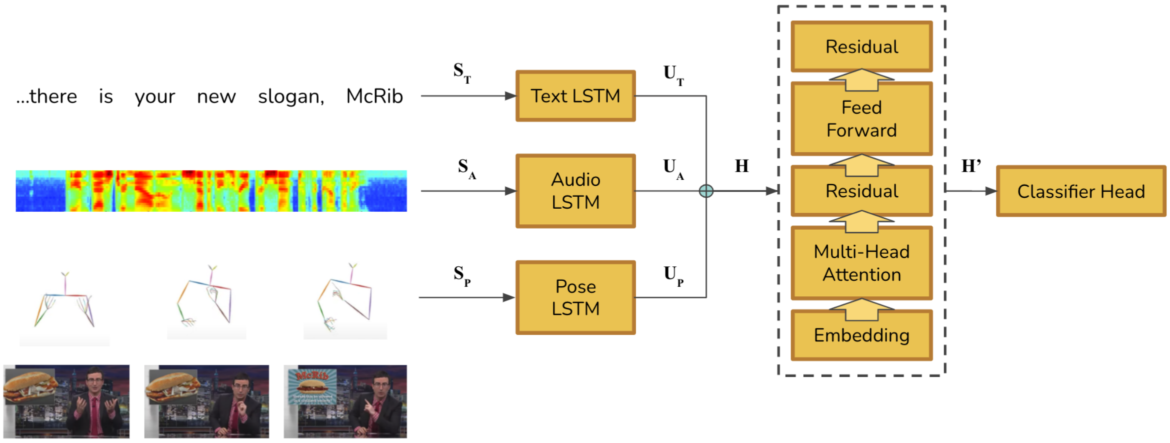

Figure 1: Overall system architecture showing the raw features encoded into unimodal features through the individual LSTMs, combined encoding through the Transformer encoder, and the classifier head.

Models	F1 micro
Multimodal Feature Encoding	
Text + Audio + Pose	.9507
Text + Audio	.9630
Text + Pose	.9317
Audio + Pose	.9855
Text	.9159
Audio	**.9899**
Pose	.8504

Table 2: Performance (F1 micro) of Multimodal Feature Encoding baselines using one or more modalities.

Models	F1 micro
Multimodal Feature Encoding	
Text + Audio + Pose	0.8964
Text + Audio	0.9564
Text + Pose	0.9233
Audio + Pose	0.9561
Text	0.8905
Audio	**0.9710**
Pose	0.7954

Table 3: Training Performance of Multimodal Context Network model

validation, and test splits from the original dataset (Ahuja et al., 2020).

5.2 Subset Training

The subset training scheme trains several models on a subset of the speakers and is evaluated on a different subset. In this setup, we hope to understand the extent to which the model is overfitting, or memorizing, specific speaker styles leading to strong performance on the test set seen in the baseline. Since the model is evaluated on speakers unseen during training, we expect a performance decline from the baseline in which the model had seen all the speakers. The model was trained on a subset of seven of the PATS speakers: fallon, almaram, lec_hist, colbert, rock, ytch_prof, lec_evol. This subset ensured that the train set was balanced with 50.34% as examples of humor and 49.66% as examples of non-humor. Furthermore, the remaining set of speakers in the test set result in a 54.4%/45.6% humor/non-humor split. Table 3 details the training performance, evaluated on the subset of speakers seen during training. Table 4 details the results of the respective models on the test set.

6 Results and Discussion

6.1 Baseline Results

The performance of the baseline on the various feature combinations shows overall promising initial results. Audio alone yielded an accuracy of 98.99% which makes additional improvements on the baseline model a difficult task. The dominating performance of the audio feature is likely due to it being the most fine-grained featured sampled from the interval, thus potentially containing the most salient information about the speaker. Table 2 highlights this, as all the models containing audio outperform the rest. Furthermore, we hypothesize that the auditory dynamicism of talk show hosts that is meant to keep the speaker engaged, creates easy ways for the model to determine whether a

Models	F1 macro	F1 micro	F1 weighted	Humor Precision	Humor Recall
Multimodal Context Network					
Text + Audio + Pose	0.7177	0.7661	0.7550	0.7822	0.8945
Text + Audio	0.8839	0.8953	0.8955	0.9245	0.9162
Text + Pose	0.7974	0.8058	0.8106	0.9281	0.7649
Audio + Pose	**0.8933**	**0.8996**	**0.9016**	**0.9784**	0.8670
Text	0.8130	0.8253	0.8283	0.9062	0.8201
Audio	0.8366	0.8606	0.8566	0.8595	**0.9428**
Pose	0.5920	0.6491	0.6407	0.7161	0.7755

Table 4: Performance of Multimodal Feature Encoding model trained on a subset of speakers and tested on a different subset of speakers

particular speaker is a talk show host, and thus an example of humor in our data.

Interestingly, pose alone has 85.04% accuracy in predicting humorous action even though it is a coarser feature that is typically not the primary source of information communication. Furthermore, the only modality whose performance is improved when pose is added is the text modality since text alone is 91.59% whereas text + pose is 93.17%. The lack of performance gain in the other modality combinations is likely due to the dominating aspect of the audio feature. Anything that dilutes or takes away from the importance of the audio feature will cause the model performance to waver.

6.2 Baseline Shortcomings

The stellar performance of the baseline models does raise some questions regarding the extent to which the baseline models have overfit to the particular styles of the respective speakers. Since the examples of humor we chose from the PATS dataset are all talk show hosts, the models may have just learned the styles of the respective speakers as opposed to what humorous language, audio, and pose entail. Furthermore, the models were trained on every speaker and tested on every speaker, further increasing the chance that individual styles were learned as opposed to general methods for humor.

6.3 Subset Training Results

The results from the subset training scheme are quite promising in affirming the models generalizability. Given that we evaluated these models on a set of speakers unseen during training, a performance decline was expected. Furthermore, considering that the size of the training set was about a third of the size of the original train set seen in the baseline, the models continue to perform well.

Table 3 details the training performance. These results show a slight decline in performance when compared to the baseline. Audio alone continues to be the best single modality, and overall modality, when compared to the other models, yielding an accuracy of 97.10%. Since the speakers in the train set were not used in the test set, we were able to use the datapoints that were previously withheld for the test set, thus creating a slightly larger train set. However, training on a subset of the speakers still results in a train set that is nearly a third of the size of the original baseline train set.

Table 4 details the results on the test set. The audio + pose modality is consistently the best performing model across the various evaluation metrics with an accuracy of 89.33%. Following closely behind is the text + audio modality with an accuracy of 88.39%. This high performance of audio pair modalities is also reflected in the baseline performance. The audio + pose model leads in precision (by over a 5% margin) for the humor class with 97.84% accuracy. Interestingly, audio—which had dominated the baseline models with 97.10% accuracy—was the third most performant modality. However, audio remained as the best single modality for the binary humor prediction task.

Audio alone does not completely falter with the highest recall for the humor class of 94.28%. This performance exceeds the recall of the audio + pose modality by a margin of 7.58%. However, since false negatives may be less of a concern in a low stakes humor classification setting, this is a less useful trade-off.

The consistently strong results, even when trained on less data and tested on unseen speakers, indicates that the models are likely not overfitting to the individual styles of the speakers, but perhaps learning general styles for humor/non-humor from the various modalities.

6.4 Qualitative Analysis of Results

With the strong performance of the subset training task, it is worth examining the performance of the most accurate model (audio + pose) on the individual speakers. We examine the speakers with the highest and lowest recall from the humor/non-humor classes, respectively. For the humor class, conan has a recall of 98.77% while ellen bottoms out at 71.69%. On the other hand, for the non-humor class, lec_law leads in performance with a recall of 97.71% while chemistry is the worst with 90.75%. For context, conan and ellen are both talk show hosts while lec_law and chemistry represent academic lectures given by professors.

Figure 2 shows the relationship between spatial extent average and the lexical diversity of the various speakers in the PATS dataset. The figure shows a natural clustering between the the various domains of speakers. Note that tv shows are represented in the humor class while lecturers, televangelists, and YouTube are represented in the non-humor class.

6.4.1 Humor Speakers

In our test set, the humor speaker with the highest recall was conan. Upon visual inspection of Figure 2, we can see that conan is the nearest neighbor to colbert, a speaker present in the train set. This figure specifically identifies the lexical diversity and the spatial extent of each speaker. In terms of lexical diversity, the broad topic covered by by a talk-show host such as Conan O'Brien would lead to high lexical diversity. In addition, most of the the tv show speakers have a higher spatial extent average. Conan typically performs with very large gestures that vary greatly for comedic effect and other signalling. Stephen Colbert and Jimmy Fallon fall into a similar type of humorous performance that involves these large gestures.

On the other end of humor, we see the audio + pose model struggle with ellen. From Figure 2 we can see that of all the tv show speakers, ellen is out of cluster. Instead, ellen falls in the middle of the range for both lexical diversity and spatial extent. The nearest neighbors to ellen are YouTubers. This is somewhat expected as Ellen DeGeneres has a much different style of performance than the aforementioned speakers. Ellen tend to be more reserved in terms of gestures (as seen by the smaller spatial extent average) and the topic of the show tends to be limited in scope to celebrity interviews and is much less political. As such, there tends to be less linguistic diversity. While Ellen is light-hearted and enthusiastic, her performance does not necessarily constitute a humorous act. As such, one could argue that perhaps ellen would fall into a separate category that is happy, bubbly, but not necessarily humorous. Alternatively, the model was not exposed to the type of humorous performance that Ellen employs, thus stressing the need for a more diverse humor corpus.

6.4.2 Non-humor Speakers

We can use similar qualitative analysis to examine the performance of the model on non-humorous examples. According to the audio + pose model, lec_law leads has the highest performance with a recall of 97.71%. This is expected as it is the nearest neighbor of lec_evol, lec_hist, and ytch_prof (Figure 2), all three of which were speakers included in the train set. Conversely, chemistry is has the worst recall for the non-humor speakers. Similar to ellen, chemistry lies far away from any other of the speakers in the PATS dataset, much less the ones used in the train set. It seems that the chemistry lecturer had the high expressiveness of the humor speakers but the low lexical diversity of the non-humor speakers, thus leading to lower recall.

6.4.3 Qualitative Analysis Conclusion

Ultimately, the qualitative analysis of the individual speakers has affirmed that the lexical diversity and spatial extent average of the individual speakers results in a set of clusters separated largely on lines of humor/non-humor. Intuitively, speakers with humorous intent may have a degree of dynamicism in their gesturing that is meant to signal excitement, happiness, or overall an emphasis of a punchline.

7 Future Work

While the results of these experiments show promising performance for the binary humor classification task, there remains a need for more robust multimodal humor data. Talk-show hosts present a very limited scope of humor that falls into a certain standard of performance. Humor, however, is a diverse expression in subject matter, performance, and setting. As such, a multimodal humor

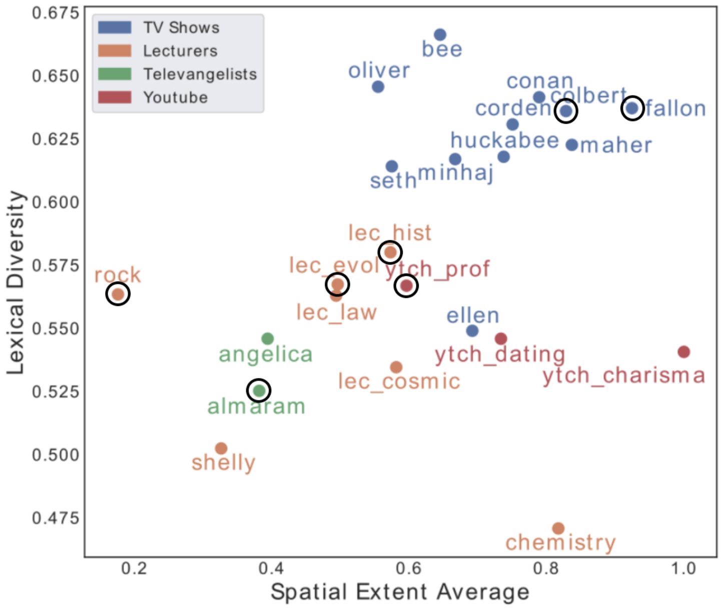

Figure 2: Spatial Extent Average vs. Lexical Diversity. The speakers circled in black were included in the train set, while the remaining set of speakers was withheld for the test set. This figure was adapted from the work of Ahuja et al.

.

dataset should include a wide array of different comedians as seen in the UR-FUNNY dataset (Hasan et al., 2019). It is important, however, that the speakers be comedians as it makes the alleviates concerns surrounding humorous intent. In other words, the task is more specified when predicting humorous intent as opposed to humor. Everyone perceives humor differently, thus making a pure humor prediction task extremely subjective. Several comedic performances can be found online and can be extracted in a manner similar to PATS which includes aligned pose, audio, and transcript information (Ahuja et al., 2020; Ahuja et al.; Ginosar et al., 2019). It would also be useful to include facial expression information similar to the UR-FUNNY dataset as the face can provide addition salient gesture information (Hasan et al., 2019).

8 Ethical Implications

It is important to consider the ethical implications of gesture modeling for humor. As gestures can be used to enhance the representation of an individual through video, one must be aware of potentially malicious uses of such a model.

The ability of a gesturing agent to establish rapport with a user could lead to a false sense of security in the user that is especially dangerous when used in a surveillance context. Users may also be more inclined to share private information under this guise.

It is also important to consider that humorous intent varies across culture and language. A humorous gesture model trained on American subjects speaking English may have learned gestures that do not necessarily translate into humor in other linguocultural contexts.

References

Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895.

Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach.

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

Dario Bertero and Pascale Fung. 2016. Deep learning of audio and language features for humor prediction. In *Proceedings of the Tenth International Conference* *on Language Resources and Evaluation (LREC'16)*, pages 496–501.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.

Lei Chen and Chungmin Lee. 2017. Predicting audience's laughter during presentations using convolutional neural network. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 86–90.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE.

Luke De Oliveira and Alfredo L Rodrigo. 2015. Humor detection in yelp reviews. *Retrieved on December*, 15:2019.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

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*.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506.

Elisa Gironzetti. 2017. Prosodic and multimodal markers of humor. In *The Routledge handbook of language and humor*, pages 400–413. Routledge.

Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Kleomenis Katevas, Patrick GT Healey, and Matthew Tobias Harris. 2015. Robot comedy lab: experimenting with the social dynamics of live performance. *Frontiers in psychology*, 6:1253.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538.

John Morkes, Hadyn K Kernal, and Clifford Nass. 1999. Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of srct theory. *Human-Computer Interaction*, 14(4):395–435.

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.

Robert R Provine. 1992. Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bulletin of the Psychonomic Society*, 30(1):1–4.

Rachel Rakov and Andrew Rosenberg. 2013. " sure, i did the right thing": a system for sarcasm detection in speech. In *Interspeech*, pages 842–846.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Erika L Rosenberg and Paul Ekman. 2020. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Oliviero Stock and Carlo Strapparava. 2005. Hahacronym: A computational humor system. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 113–116.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Cornelia S Wendt and Guy Berg. 2009. Nonverbal humor as a new dimension of hri. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 183–188. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2367–2376.