



# Deep Learning for Acoustic Irony Classification in Spontaneous Speech

Helen Gent<sup>1,3</sup>, Chase Adams<sup>1,3</sup>, Chilin Shih<sup>1</sup>, Yan Tang<sup>1,2</sup>

<sup>1</sup>Department of Linguistics, University of Illinois Urbana-Champaign, USA

<sup>2</sup>Beckman Institute for Advanced Science and Technology, USA

<sup>3</sup>SRI International, USA

{hmgent2, chasea2, cls, yty}@illinois.edu  
{helen.gent, chase.adams}@sri.com

## Abstract

Recognizing irony in speech and text can be challenging even for humans. For natural language processing (NLP) applications, irony recognition presents a unique challenge as irony alters the sentiment and meaning of the words themselves. Combining phonological insights from past literature on irony prosody and deep learning modeling, this research presents a new approach to irony classification in naturalistic speech data. A new corpus consisting of nearly five hours of irony-annotated, naturalistic, conversational speech data has been constructed for this study. A wide array of utterance-level and time-series acoustic features were extracted from this data and utilized in the training and fine-tuning of a series of deep learning approaches for irony classification. The best-performing model achieved an area under the curve of 0.811 in the speaker dependent condition, and 0.738 in the speaker independent condition, outperforming most irony classification models in the existing literature. In addition to the myriad real-world applications for this approach, its contribution to the understanding of prosodically-encoded augmentation of semantic content constitutes a significant step forward for research in the fields of linguistics and NLP.

**Index Terms:** irony classification, computational paralinguistics, deep learning, naturalistic speech data, prosody modeling

## 1. Introduction

Verbal irony is a blanket term for a broad array of rhetorical devices that alter the sentiment, intention, and meaning expressed in speech, often for the purpose of humor<sup>1</sup>. While many have argued that ironic utterances must have a meaning opposite to the literal meaning of the words that comprise them [1, 2, 3, 4, 5, 6], others have defined verbal irony more loosely as utterances where the intended meaning is simply different from the literal meaning [7, 8, 9, 10, 11]. Here, an utterance is considered ironic if the literal meaning of the words is in opposition to – or at least different from – the speaker’s intended meaning. This definition broadens the scope of admissible ironic utterances by not requiring direct contradiction. This is particularly crucial when using naturalistic, conversational speech data, in which the kind of experimental control necessary to ensure exact opposition cannot be exerted. This research examines verbal irony that is the result of successful, collaborative communication between the speaker and the listener; that is, irony that is both intended to be understood and successfully interpreted.

Although prosody has been found to be significant, alongside semantic and contextual information, in the identification of irony in speech [3, 8, 12, 13, 14, 15], the results of

recent phonological research into irony prosody indicate that there may be no single reliable set of acoustic cues constituting an “ironic tone of voice”. Rather, the variability between ironic and non-ironic speech may rely on the speaker’s communicative choices [7, 16]. Nonetheless, decades of research have identified a selection of acoustic features by which speakers may differentiate ironic from non-ironic speech, even if the directionality and degree of difference is not reliable. These acoustic features, most commonly measured at the utterance level, include differences in F0 mean and its standard deviation (SD) [4, 7, 9, 10, 16, 17, 18, 19, 20], amplitude mean and SD [4, 7, 17, 19], timing, stress, and duration [7, 9, 10, 14, 16, 17, 18, 19, 21], segmental reduction [19], and voice quality measures such as HNR [9, 10].

Machine learning approaches to irony classification have focused primarily on text-based data such as tweets [22, 23, 24]. Approaches using speech data have relied heavily on non-naturalistic sources in the form of either elicited irony [25] or acted speech from situational comedies [26, 27]. This reliance on acted or elicited irony, while practical, is far from ideal for the study of a speech phenomenon as intricate as irony, particularly given the significant differences between naturalistic and “posed” irony [4].

This study investigates the efficacy of various acoustic and text feature combinations for the classification of irony in naturalistic speech and the role of prior knowledge of a speaker in successfully identifying irony in their speech.

## 2. Methods

### 2.1. Data Gathering

Comedy podcasts were identified as an ideal pre-recorded, naturalistic data source, given that they are often unscripted, conversation-based, and likely to contain a high concentration of ironic utterances. The creator of the Sad Boyz Podcast [28] provided access to the original, premixed recordings of several podcast episodes in WAV format. A team of annotators was trained to perform irony annotation using the discourse-based irony labeling method introduced in [29]. For each segmented utterance in the long-form episodes, irony was annotated based on the conversational response that followed it. This allowed the annotators to interpret conversational participant comprehension and acknowledgement of irony at the time of recording, rather than using their own judgements as listeners outside the conversation. Each episode was labelled by two annotators, and only the samples for which both annotators agreed about the application of the irony labeling scheme were included in the corpus.

The final corpus, with a reduced number of non-ironic utterances selected to achieve class balance, comprises 4.68 total

<sup>1</sup>The various sub-types of verbal irony were not differentiated.

hours of naturalistic, conversational speech produced by 12 different speakers (7 female, 5 male). In total, there are 5,812 samples - 2,906 labeled ironic and 2,906 labeled non-ironic. Utterances ranged from 0.20 to 13.25 seconds (mean 2.90, SD 2.12). Guests to the podcast provided anywhere from 6 to 115 ironic utterances (mean 45.4, SD 34.57). Notably, the podcast hosts (both male) produced significantly more ironic utterances (1,495 and 957), since they featured in each episode. In selecting non-ironic utterances while preserving duration variance in the data, non-ironic samples were ordered by the absolute value of the difference between their duration and the mean duration for all non-ironic utterances. Then a number of non-ironic samples equal to the number of ironic samples were selected at equally spaced points along this ordered list, irrespective of speaker identity.

## 2.2. Acoustic Feature Selection and Extraction

Preprocessing and acoustic feature extraction were performed using Python. Any samples recorded in stereo were down-mixed to a single channel. The files were then downsampled to 16 kHz, followed by normalizing their root-mean-square (RMS) amplitude to 0.04. Utterance-level feature selection was primarily motivated by the results of past phonology literature on the topic of irony prosody. Automatic speech recognition (ASR) output from Wit.ai [30] provided textual inputs to the Penn Forced Aligner [31], the alignments from which facilitated the extraction of segment-level features which were normalized by speaker per utterance. The following 37 utterance-level acoustic features (Fts) in five acoustic categories were extracted from each sample:

- F0 (7 Fts): mean, SD, range, median; mean for sonorant consonants, stressed vowels, and unstressed vowels
- Formants 1-3 (9 Fts): mean values for sonorant consonants, stressed vowels, and unstressed vowels
- Timing (10 Fts): sound-to-silence ratio, total utterance duration, total number of pauses, syllables per second, consonant-to-vowel ratio; average duration of words, silences, consonants, vowels, and stop consonants
- Energy (4 Fts): SD and range of windowed RMS (10-ms windows), dynamic range, intensity
- Voice Quality (7 Fts): HNR mean, SD, and range; center of gravity, skewness, kurtosis, and SD of fricative consonants' spectra

In addition to these utterance-level features, a selection of time-series acoustic features were extracted, given the promising results in [29] examining time-series differences between ironic and non-ironic speech in terms of F0 contour. The following 4 time-series acoustic features were extracted from each 10-ms window, for a total of 24 values per window:

- F0
- HNR
- 1<sup>st</sup>-13<sup>th</sup> Mel-Frequency Cepstral Coefficients
- 1<sup>st</sup>-9<sup>th</sup> Relative Spectral Transform - Perceptual Linear Prediction Coefficients

## 2.3. Model Input Preprocessing

Three configurations of utterance-level acoustic features were tested. The 37 utterance-level features were first transformed to a matching number of principle components (PC) by a Principal

Components Analysis. Two separate feature sets were then created using the first 3 and 30 PCs. The decision to use only the first 3 PCs was prompted by the motivation to be able to interpret meaningful groupings of acoustic features. F0, HNR and formants, and timing/duration measures were found to be the primary contributors in the first 3 PCs, respectively. The first 3 PCs, however, only explained 33.1% of the data variance. Therefore, the second configuration using 30 PCs, explaining 97.5% of the data variance, was also evaluated. Finally, the ComParE 2016 feature set was also tested as a baseline with other work in this problem space [32, 33].

For the time-series features, input lengths were normalized to a uniform value by grouping time-series information for each speech utterance into bins representing 10% of the total utterance duration, irrespective of the number of 10-ms windows in each bin. The means and SDs of the 24 time-series features in each bin were used as the model inputs, resulting in an input tensor of shape (10, 48).

For the textual features, the transcriptions acquired by ASR were incorporated through the use of pre-trained word2vec embeddings from the Google News corpus [34]. Prior to being converted to word embeddings, text inputs were vectorized, then truncated or zero-padded to a uniform length of 25 words per utterance<sup>2</sup>.

## 2.4. Model Architectures

The features extracted above correspond to three input modalities: utterance-level acoustic data, time-series acoustic data, and text embeddings. Each modality was evaluated independently and in combination, using simple architectures to provide a baseline for future comparison. A basic feed-forward neural network was selected for the utterance-level acoustic inputs, a long short-term memory neural network (LSTM) for the time-series acoustic inputs, and a convolutional neural network (CNN) for the text inputs, as detailed in Fig. 1.

The models using only one input modality (e.g. utterance-level acoustic features only) are here referred to as *unimodal* models while models combining two or three input modalities are referred to as *bimodal* and *trimodal* models, respectively. All models were implemented using the Keras toolkit [35]. In all cases, model training was performed using an Adam optimizer and a sparse categorical crossentropy loss function. A batch size of 64 was employed over 150 epochs per model. A rectified linear unit (ReLU) was used as the activation function for all the hidden layers. A full treatment of hyper-parameter tuning over all input feature and modality combinations was deemed cost-prohibitive. As such, tuning was not performed until after the best combination of input features had been identified. Model implementation, preprocessing, training, and evaluation code can be found at <https://github.com/helengent/Irony-Recognition>.

Fig. 1 visualizes the architecture for all three input modalities, and their combination. For clarity, bimodal models simply excluded a single modality at concatenation. The unimodal models use largely the same architecture for each of their individual branches as in Fig. 1, without the need to reach a uniform intermediate size for concatenation. Fig. 1 represents the model architecture after hyper-parameter tuning, and is not reflective of the precise architecture from initial modality comparison experiments.

<sup>2</sup>25 words represents the corpus' 95th percentile utterance length

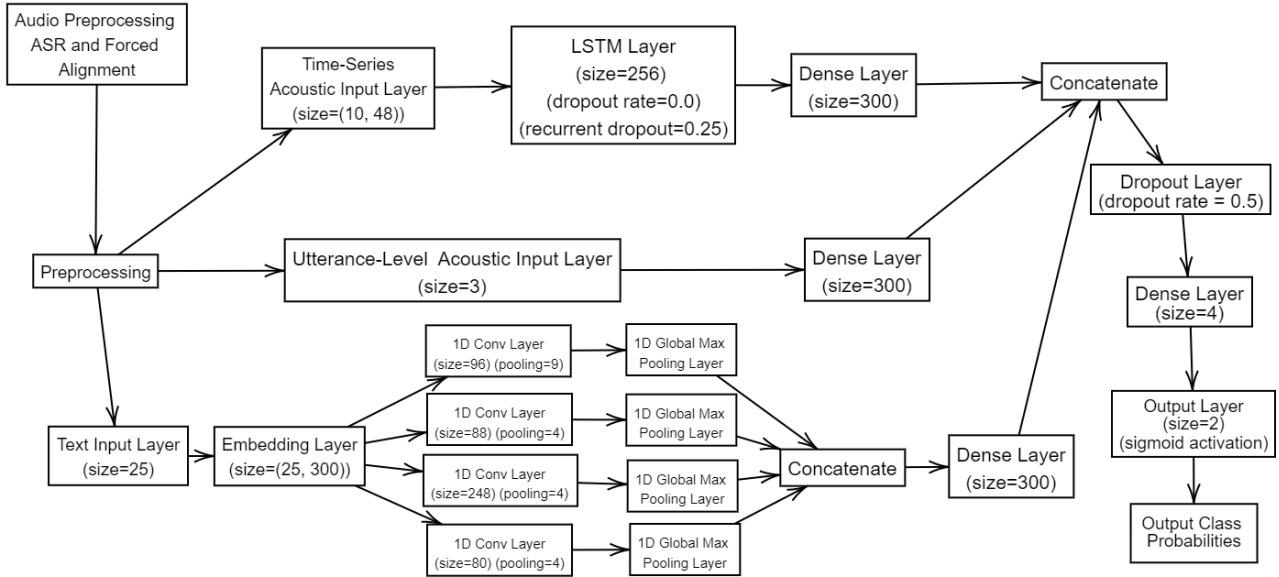


Figure 1: Flowchart representing the architecture of the trimodal model after hyper-parameter tuning

Table 1: F1 comparison for unimodal and bimodal models using time-series acoustic features and text embeddings

| Dependent | Independent | Time-series | Text |
|-----------|-------------|-------------|------|
| 0.714     | 0.611       | x           |      |
| 0.667     | 0.616       |             | x    |
| 0.725     | 0.655       | x           | x    |

Table 2: F1 comparison of unimodal and trimodal models using different configurations of utterance-level acoustic features

|         | Unimodal  |             | Trimodal     |              |
|---------|-----------|-------------|--------------|--------------|
|         | Dependent | Independent | Dependent    | Independent  |
| ComParE | 0.712     | 0.585       | 0.601        | 0.644        |
| 30 PCs  | 0.696     | 0.551       | 0.721        | 0.653        |
| 3 PCs   | 0.617     | 0.338       | <b>0.728</b> | <b>0.664</b> |

### 3. Results

The performance of all models was evaluated in terms of precision, recall, F-measure (F1), and accuracy in addition to the area under the curve (AUC) of the Receiver operating characteristic and its corresponding Equal Error Rate (EER). A 5- and 4-fold cross-validation strategy was adopted for the speaker dependent and speaker independent conditions, respectively. In comparing these metrics across the unimodal, bimodal, and trimodal models, the best combination of input modalities was, narrowly, the trimodal model combining 3 utterance-level PCs, time-series acoustic features, and text embeddings. For brevity, F1 metrics are presented to quantify the differences in model configuration. A full treatment of these metrics is given in Tab. 3 for the best model after hyper-parameter tuning.

Tab. 1 shows performance results for the best unimodal and bimodal models - those using time-series acoustic features and text embeddings. Note the marked performance decrease between the dependent and independent conditions using only time-series acoustic features as compared with using only text. This performance decrease is mitigated when the two input modalities are combined.

Table 3: Average performance combining 3 utterance-level PCs, time-series acoustic features, and text embeddings (s.d.)

| Cross Validation Strategy | Speaker Dependent | Speaker Independent |
|---------------------------|-------------------|---------------------|
| Precision                 | 0.708 (0.044)     | 0.665 (0.101)       |
| Recall                    | 0.782 (0.094)     | 0.631 (0.102)       |
| F1                        | 0.739 (0.026)     | 0.636 (0.022)       |
| Accuracy                  | 0.725 (0.018)     | 0.679 (0.029)       |
| AUC                       | 0.811 (0.015)     | 0.738 (0.027)       |
| EER                       | 0.267 (0.008)     | 0.320 (0.025)       |

Tab. 2 shows a clear trend in the performance of the unimodal models increasing with the addition of more features. In the trimodal models, however, this trend is reversed.

Hyper-parameter tuning was conducted for the best combination of input modalities. A random search optimizing for validation accuracy was performed across layer sizes, number of convolutional layers, pooling sizes, dropout rates, optimizer type, and activation function for the output layer over 400 trials. The sizes of input layers, embedding layers, and the output layer were not changed, nor were the activation functions of hidden layers. Fig. 1 shows the best model architecture after hyper-parameter tuning, while Tab. 3 displays the results across cross-validation folds of the best-performing combination of input modalities after hyper-parameter tuning<sup>3</sup>.

### 4. Discussion

The impact of prior speaker knowledge on performance was assessed through the use of the dependent and independent conditions. Examining the time-series acoustic and text embedding unimodal results, the difference in performance decrease between conditions indicates that text-exclusive irony classification is less sensitive to speaker-specific artifacts than acoustic-exclusive approaches. The bimodal combination of these two

<sup>3</sup>Excluding either of the two hosts did not result in a major decrease in performance as evidenced by low SD in the independent condition suggesting that the over-representation of these speakers in the data does not unduly bias the model for them.

Table 4: Performance comparison of top models with existing literature

|                              | Data Source                        | Language | Features  | Model                                  | Speaker Dependent Performance               | Speaker Independent Performance       |
|------------------------------|------------------------------------|----------|---|--|---|---------------------------------------|
| Burkhardt et al. (2017)[25]  | Elicited irony (crowd-sourced app) | German   | ComParE   | SVM with linear kernel                 | UAR: 0.693                                  | UAR: 0.614                            |
| Castro et al. (2019)[26]     | Acted (sit-coms)                   | English  | Hand-engineered acoustic (including time-series), text embeddings, video cues | SVM                                    | text+video F1: 0.716                        | text+audio F1: 0.631                  |
| Rakov & Rosenberg (2013)[27] | Acted (Daria)                      | English  | Hand-engineered acoustic  | SimpleLogistic (LogitBoost) classifier | Accuracy: 0.8157 (112 samples, one speaker) | N/A                                   |
| Yang et al. (2019)*[36]      | Naturalistic (gameplay videos)     | Chinese  | Acoustic (baseline set and selected time-series) + text                       | CNN                                    | AUC: 0.751 (one speaker)                    | N/A                                   |
| <i>This study</i>            | Naturalistic (podcasts)            | English  | Hand-engineered acoustic (including time-series), text embeddings             | LSTM + CNN                             | AUC: <b>0.811</b><br>F1: <b>0.739</b>       | AUC: <b>0.738</b><br>F1: <b>0.636</b> |

outperforms all unimodal models while mitigating the effect of speaker dependency in the acoustic data.

The addition of utterance-level PCs to the trimodal model improved only slightly on the overall bimodal performance using time-series acoustic features and text embeddings. While time-series acoustic features and text embeddings together likely contain the most information for this task, the introduction of redundancy in the form of utterance-level acoustic features led to marginal performance gains.

Tab. 4 places the performance of the aforementioned best models by condition in context with results previously reported for this task. An exact one-to-one comparison is impossible due to varying evaluation metrics across past studies, different data sources, and even different languages under investigation. However, given how understudied the question of irony classification in speech data has been, these remain the best models against which to compare results.

The use of acted, elicited, or naturalistic speech data is expected to impact the results of ironic speech studies. Machine learning classifiers using only acted speech are expected to achieve better overall performance metrics, mirroring the observation that human listeners more easily discern acted irony than naturalistic irony [4]. Thus, the use of naturalistic speech should be considered a factor making this task more challenging than using acted, or even elicited, speech.

In general, the models developed in this study perform competitively with models from past literature for both conditions. This bears out when comparing the reported Unweighted Average Recall (UAR)<sup>4</sup> in [25] with the F1 and accuracy scores reported in Tab. 3. The model in [25] used the ComParE feature set, yet the results of modality comparison experiments on the trimodal models indicate that the ComParE feature set does not perform as effectively on the data gathered for this research as knowledge-driven, task-specific features. While the differences in data sources and languages between these two models prevent one-to-one comparison, the fact that the models using hand-selected acoustic features in this study outperform Burkhardt et al.’s best models using ComParE may serve as evidence that tailoring acoustic feature selection to the phenomenon under consideration has more potential for classifier success than using a one-size-fits-all feature set.

While Rakov and Rosenberg reported an impressive speaker dependent accuracy of 0.816 [27], this performance must be contextualized by its corpus of only 112 acted speech samples chosen explicitly by the authors from an animated show in hopes of extracting samples with particularly “exaggerated acted speech.” These factors are expected to lead to higher performance without generalizing to real-world human speech.

Castro et al. also used acted speech, from live-action situ-

<sup>4</sup>UAR was deemed unnecessary due to this corpus’ class balance.

ational comedies, but had a larger corpus (690 samples) [26]. Their best model for the speaker independent condition used text and audio data, including both utterance-level and time-series acoustic features; this study corroborates the finding that these are the most predictive input features for irony classification. The models reported in Tab. 3 outperform the results reported by [26] across conditions. It is notable that a model using naturalistic speech outperformed a model using acted speech, particularly given that hand-engineered acoustic features were also used in [26], perhaps accounted for by a larger data set or the use of deep learning rather than support vector machines.

In [36], Yang et al. presented a study of general “humor” rather than specifically irony, but these tasks have a great deal of overlap. Their model was included in Table 4 because it utilized naturalistic speech, something none of the prior attempted irony classifiers did, making it one of the best models against which to compare these results. An AUC of 0.751 for a single speaker in the task of “humor” detection was reported in [36]. While this is not directly comparable with irony, it still demonstrates that the models detailed in Tab. 3 perform competitively with other uses of naturalistic speech for a similar task.

While this study presents a powerful combination of features and techniques, advancing the state of the art for verbal irony classification, much work remains. Due to the nature of the annotation scheme, no comparison can yet be made against naïve listener judgement on naturalistic speech data. There is some evidence from the literature that conversational context from the preceding turn may also be informative [37]. Furthermore, the incorporation of video data may allow for the analysis of facial expression and body language cues for irony [26]. The combination of such novel analyses alongside the presented framework is expected to improve further upon these results.

## 5. Conclusion

One of the unique challenges verbal irony classification presents to researchers is the necessity for naturalistic speech data which can generalize to real-world applications. While naturalistic speech can be difficult to gather and is less controlled than acted speech, this study demonstrates that using such data is as viable or more so than following past trends of using acted speech data. The corpus of naturalistic speech detailed here should be used as a benchmark for future research in this space. Additionally, the combination of theoretically-founded, task-specific features with modern deep learning architectures leads to significant performance increases in this space.

**Acknowledgements** We express sincere thanks to Jarvis Johnson for access to The Sad Boyz Podcast, without which this research could not happen. We also thank the annotation team who painstakingly prepared the corpus for analysis and our anonymous reviewers for their helpful feedback.

## 6. References

- [1] G. A. Bryant and J. E. Fox Tree, "Is there an ironic tone of voice?" *Language and speech*, vol. 48, no. 3, pp. 257–277, 2005.
- [2] A. Cutler, "On saying what you mean without meaning what you say," in *Tenth Regional Meeting, Chicago Linguistic Society*. CLS, 1974, pp. 117–127.
- [3] L. M. Milosky and J. A. Ford, "The role of prosody in children's inferences of ironic intent," *Discourse Processes*, vol. 23, no. 1, pp. 47–61, 1997.
- [4] P. Rockwell, "Lower, slower, louder: Vocal cues of sarcasm," *Journal of Psycholinguistic research*, vol. 29, no. 5, pp. 483–495, 2000.
- [5] P. Rockwell and E. M. Theriot, "Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis," *Communication Research Reports*, vol. 18, no. 1, pp. 44–52, 2001.
- [6] J. Tepperman, D. Traum, and S. Narayanan, "'yeah right': Sarcasm recognition for spoken dialogue systems," in *Ninth international conference on spoken language processing*, 2006.
- [7] L. Anolli, R. Ciceri, and M. G. Infantino, "From 'blame by praise' to 'praise by blame': Analysis of vocal patterns in ironic communication," *International Journal of Psychology*, vol. 37, no. 5, pp. 266–276, 2002.
- [8] C. A. Capelli, N. Nakagawa, and C. M. Madden, "How children understand sarcasm: The role of context and intonation," *Child Development*, vol. 61, no. 6, pp. 1824–1841, 1990.
- [9] H. S. Cheang and M. D. Pell, "The sound of sarcasm," *Speech communication*, vol. 50, no. 5, pp. 366–381, 2008.
- [10] R. R. Schaffer, "Vocal cues for irony in english," Ph.D. dissertation, The Ohio State University, 1982.
- [11] D. Tannen *et al.*, *Conversational style: Analyzing talk among friends*. Oxford University Press, 1984.
- [12] B. P. Ackerman, "Form and function in children's understanding of ironic utterances," *Journal of Experimental Child Psychology*, vol. 35, no. 3, pp. 487–508, 1983.
- [13] L. Anolli, R. Ciceri, and M. G. Infantino, "Behind dark glasses: Irony as a strategy for indirect communication," *Genetic, social, and general psychology monographs*, vol. 128, no. 1, p. 76, 2002.
- [14] G. A. Bryant and J. E. Fox Tree, "Recognizing verbal irony in spontaneous speech," *Metaphor and symbol*, vol. 17, no. 2, pp. 99–119, 2002.
- [15] C. Nakassis and J. Snedeker, "Beyond sarcasm: Intonation and context as relational cues in children's recognition of irony," in *Proceedings of the twenty-sixth Boston University conference on language development*. Cascadilla Press, Somerville, MA, 2002, pp. 429–440.
- [16] G. A. Bryant, "Prosodic contrasts in ironic speech," *Discourse Processes*, vol. 47, no. 7, pp. 545–566, 2010.
- [17] L. Anolli, R. Ciceri, and M. G. Infantino, "Irony as a game of implicitness: Acoustic profiles of ironic communication," *Journal of Psycholinguistic Research*, vol. 29, no. 3, pp. 275–311, 2000.
- [18] P. Rockwell, "Vocal features of conversational sarcasm: A comparison of methods," *Journal of psycholinguistic research*, vol. 36, no. 5, pp. 361–369, 2007.
- [19] O. Niebuhr, "'a little more ironic'—voice quality and segmental reduction differences between sarcastic and neutral utterances," in *7th International Conference on Speech Prosody, Dublin, Ireland, Proceedings*, 2014, pp. 608–612.
- [20] S. Attardo, J. Eisterhold, J. Hay, and I. Poggi, "Multimodal markers of irony and sarcasm," *Humor*, vol. 16, no. 2, pp. 243–260, 2003.
- [21] R. J. Kreuz and R. M. Roberts, "Two cues for verbal irony: Hyperbole and the ironic tone of voice," *Metaphor and symbol*, vol. 10, no. 1, pp. 21–31, 1995.
- [22] C. Van Hee, E. Lefever, and V. Hoste, "Semeval-2018 task 3: Irony detection in english tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 39–50.
- [23] T. Vu, D. Q. Nguyen, X.-S. Vu, D. Q. Nguyen, M. Catt, and M. Trenell, "Nihrio at semeval-2018 task 3: A simple and accurate neural network model for irony detection in twitter," in *Proceedings of the 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2018, pp. 525–530.
- [24] C. Baziotis, N. Athanasiou, P. Papalampidi, A. Kolovou, G. Paraskevopoulos, N. Ellinas, and A. Potamianos, "Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns," in *Proceedings of the 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2018.
- [25] F. Burkhardt, B. Weiss, F. Eyben, J. Deng, and B. Schuller, "Detecting vocal irony," in *International Conference of the German Society for Computational Linguistics and Language Technology*. Springer, Cham, 2017, pp. 11–22.
- [26] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an 'Obviously' perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4619–4629. [Online]. Available: <https://aclanthology.org/P19-1455>
- [27] R. Rakov and A. Rosenberg, "'sure, i did the right thing': a system for sarcasm detection in speech," in *Interspeech*, 2013, pp. 842–846.
- [28] J. Johnson, "Sad boyz," 2017. [Online]. Available: <https://sadboyzpod.com/>
- [29] H. Gent, "F0 as a cue for irony in spontaneous speech," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Australasian Speech Science and Technology Association Inc., 2019, pp. 711–715.
- [30] Wit.ai, Inc, "Wit.ai," 2020. [Online]. Available: <https://wit.ai/>
- [31] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics '08*, 2008.
- [32] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Wenginger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.
- [33] F. Eyben, M. Wullmer, and B. O. Schuller, "the munich versatile and fast open-source audio feature extractor," *Proceedings ACM Multimedia (MM)*, pp. 1459–1462, 2018.
- [34] Google Code Archive, "word2vec," July 2013. [Online]. Available: <https://code.google.com/archive/p/word2vec/>
- [35] F. Chollet *et al.*, "Keras," <https://keras.io>, 2015.
- [36] Z. Yang, B. Hu, and J. Hirschberg, "Predicting humor by learning from time-aligned comments," in *INTERSPEECH*, 2019, pp. 496–500.
- [37] D. Ghosh, A. R. Fabbri, and S. Muresan, "Sarcasm analysis using conversation context," *Computational Linguistics*, vol. 44, no. 4, pp. 755–792, 2018.