UNVEILING SARCASM THROUGH SPEECH: A CNN-BASED APPROACH TO VOCAL FEATURE ANALYSIS

Blanca Sabater Vilchez[⋆] *Raphaël G. Gillioz*[⋆]

Pierre S. F. Kolingba-Froidevaux[⋆] *Florian Morgner*[⋆]

[⋆] Fachgebiet Audiokommunikation Technische Universität Berlin

ABSTRACT

Sarcasm, a prevalent yet complex form of expression, challenges both voice-assisted technologies and individuals with Autism Spectrum Disorders. Understanding sarcasm in speech is crucial for enhancing human-computer interactions and aiding nuanced communication. This study aims to replicate and extend Gao, Nayak, and Coler [1] research on sarcasm detection through vocal features using convolutional neural networks (CNNs), exploring the impact of advanced preprocessing techniques. Utilizing the MUStARD dataset, extensive preprocessing including audio extraction, denoising, and augmentation was performed, alongside employing the VGGish model in a transfer learning setup, augmented with layers tailored to sarcasm detection. Despite achieving a convergence in loss and accuracy metrics, with the highest F1-score of 0.64 noted for denoised and augmented data, the improvements were modest, underscoring the complexities of sarcasm detection. The study suggests future research focus on tempo over pitch and diversify data sources beyond sitcoms to enhance detection capabilities, indicating promising directions for advancing sarcasm recognition technologies.

Index Terms— Sarcasm recognition, voice information retrieval, tranfer learning

1. INTRODUCTION

Human communication is rich with layers of meaning beyond the literal. Irony, a form of non-literal expression, plays a significant role in everyday interactions [2]. In this paper, we focus on sarcastic utterances, which are a specific type of verbal irony that employs a discrepancy between the intended meaning and the literal meaning to express criticism or negativity towards a person or event [3]. This deliberate discrepancy creates a complex communication tool, serving various purposes [4]. Sarcastic utterances can soften a critical statement [5], add humor [6], [7], or convey layered meaning [8]. However, sarcasm can also be a source of misunderstanding. Studies highlight its potential for negativity and victimization [9]. This complexity becomes particularly challenging for individuals with Autism Spectrum Disorders (ASDs) who struggle with rapid communication despite possessing some ability to recognize sarcasm [10]. Understanding sarcasm offers a key to enhancing communication, not only for humans but also for future voice assistant interfaces. Research suggests that comprehending a speaker's intent is crucial for grasping sarcasm, and context plays a vital role [6], [11], [12]. However, specific prosodic cues, independent of context, may betray underlying sarcasm [13], [14]. Studies have identified slower tempo, lower pitch, and variations in intensity as potential indicators [13], [14]. Cheang and Pell [15] further confirm a reduction in both mean and variation of fundamental frequency (F0) during sarcasm. Analysing spontaneous sarcasm, Caucci, Kreuz, and Buder [16] confirmed a lower average amplitude and a slower tempo, reinforcing the idea that there exists a recognizable ironic or sarcastic tone of voice. However, this notion is challenged by Bryant and Fox Tree [17]. Building upon the assumption that sarcasm detection in speech is possible without context, and inspired by existing work by Bharti, Gupta, Shukla, *et al.* [18], this project aims to reproduce the experiment by Gao, Nayak, and Coler [1] with a focus on enhanced pre-processing. Our goal is to confirm the feasibility of a convolutional neural network (CNN) to detect sarcastic utterances in voice solely through their vocal features.

2. METHODOLOGY

2.1. Dataset

The MUStARD dataset, a multimodal compilation comprising 690 audiovisual excerpts from various TV shows including ”Friends,” ”The Golden Girls,” and ”The Big Bang Theory,” was utilized. This dataset is evenly distributed, with 345 utterances marked as sarcastic and another 345 as non-sarcastic, enabling a detailed exploration of sarcasm in spoken language. [19]

2.2. Pre-Processing

As the audio data was embedded within the MP4 files, we first extracted it using the ffmpeg command-line tool. This initial pre-processing step was followed by applying a basic audio normalization filter (speechnorm) with ffmpeg and converting the audio format to a standard configuration: 48 kHz sampling rate, PCM (Pulse-Code Modulation) codec, 32-bit floating-point representation (little-endian), and saved as WAV files.

These manipulations were necessary to prepare the audio data for subsequent denoising and speech enhancement. We employed one of three tested deep learning models [20]–[22] specifically designed for voice extraction. Ultimately, we chose the ready-to-use DeepFilterNet for automation. Its core component, an adaptive filter, reconstructs the periodic component of the voice and performs well under low SNR (Signal-to-Noise Ratio) conditions [22].

After extracting audio, we augmented the data through shifting, pitch changing, and time-stretching the original samples using a Python script, effectively quadrupling our dataset's size. Each audio file underwent a shift by rolling the waveform by a fraction of its sampling rate, a random pitch change within a range of -3 to +3 semitones, and a stretch by a factor of 0.8. We then resampled the augmented audio to a sampling rate of 22,050 Hz to ensure compatibility with our neural network architecture.

Feature extraction involved pitch and spectrograms, utilizing Librosa and Matplotlib libraries. Features were extracted with a size of 64x96, focusing on 64 Mel Bins and applying FFT window techniques.

2.3. Pre-trained Model

In this study the VGGish model is used in a transfer learning context, utilized to extract relevant audio features for sarcasm detection. This pre-trained model is trained on the AudioSet dataset.

Notably, all layers up to the embedding layer are frozen, meaning their weights are not updated during the training process. This ensures that the model retains the generalized feature representation learned from its extensive training on the AudioSet.

Balancing the audio samples involved adjusting the lengths of sarcastic (2021 seconds) and non-sarcastic (1581 seconds) samples through oversampling or undersampling, ensuring a balanced dataset for training the model.

This study utilized transfer learning with feature representation for the VGGish model. A targeted modification was applied to only the top layer, which can be understood as a removal of the classifier, leaving VGGish as a feature extractor. With the feature-respresentation of the pretrained model, a new classifier can be added as a top layer to deliver the predictions desired.

The VGGish model, an adaptation of the VGG network for audio, was trained and evaluated on AudioSet. AudioSet being an expansive dataset using labelled YouTube Videos. With over 2 million audio samples annotated accross 632 classes, this dataset provides a foundational base for training networks to detect audio events under real-world conditions efficiently.

VGGish is trained to process Mel-scale spectrograms. Benefiting from the depth of the VGG network and the diversity of the AudioSet, VGGish demonstrates enhanced general-

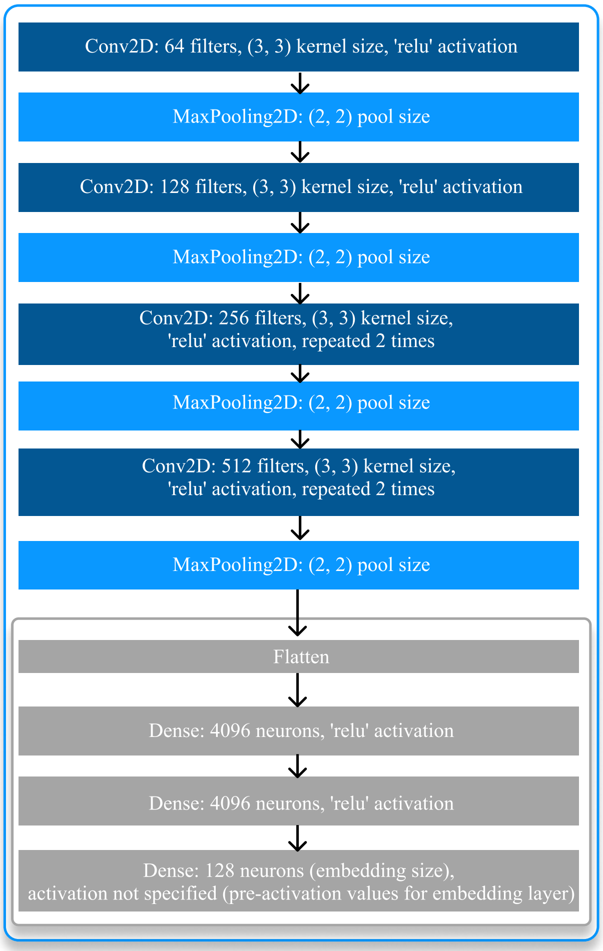
Conv2D: 64 filters, (3, 3) kernel size, 'relu' activation
MaxPooling2D: (2, 2) pool size
Conv2D: 128 filters, (3, 3) kernel size, 'relu' activation
MaxPooling2D: (2, 2) pool size
Conv2D: 256 filters, (3, 3) kernel size, 'relu' activation, repeated 2 times
MaxPooling2D: (2, 2) pool size
Conv2D: 512 filters, (3, 3) kernel size, 'relu' activation, repeated 2 times
MaxPooling2D: (2, 2) pool size
Flatten
Dense: 4096 neurons, 'relu' activation
Dense: 4096 neurons, 'relu' activation
Dense: 128 neurons (embedding size), activation not specified (pre-activation values for embedding layer)

Fig. 1. Transferred layers. Convolutional network architecture provided by VGGish. The convolutional and top layers have been frozen and used for transfer learning.

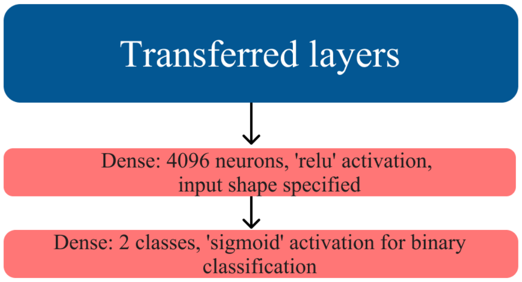
Transferred layers
Dense: 4096 neurons, 'relu' activation, input shape specified
Dense: 2 classes, 'sigmoid' activation for binary classification

Fig. 2. Fine-tuning layers. After the embedding and postprocessing steps, two layers have been added to adapt the pretrained model to the specific classification task.

Optimizer	**Batch size**	**Epoch**	**Learning rate**
Adam	8, 16, 32, 64	12, 16, 50	0.01, 1e-4, 1e-5

Table 1. Hyperparameters for fine tuning.

ization capabilities. In this research, VGGish's performance in personalized Speech Emotion Recognition (SER) is assessed, leveraging its 128-dimensional output embeddings as inputs for subsequent layers designed specifically for the target dataset. During training, the pre-trained layers of VGGish remain fixed, ensuring the retention of generalized features learned from AudioSet, while the weights of the added layers are updated with the target data. The pre-trained models architecture is detailed in figure 1.
The entire architecture of VGGish is employed for feature extraction, including all its convolutional and pooling layers, culminating in a fully connected layer (fc2) that outputs 128-dimensional embeddings. These embeddings serve as a compact representation of the audio's characteristics, capturing essential features necessary for the classification task at hand. Following extraction, the embeddings undergo a post-processing step using the VGGish postprocessor, which applies Principal Component Analysis (PCA) for dimensionality reduction and quantization.

2.4. Training

Our experimental methodology encompassed various analyses and evaluations:

- Analysis of raw data, both with augmented (RawA) and not augmentated data (RawNA).
- Examination of denoised data, considering cases with augmented (DenA) and not augmented (DenNA) data.

The hyperparameters used for the training have been detailed in Table 1.

3. RESULTS

From all the tested hyperparameters in table 1, the combination of 50 epochs, a batch size of 32, and a learning rate of 1e-5 using the Adam optimizer emerged as the best for each training setting. As demonstrated in Fig. 3, there is convergence observed for both the loss and accuracy metrics of the test set. Sequentially, the loss and accuracy exhibit improved outcomes for the DenA, followed by RawA, DenNA and finally RawNA. The use of a test set for the demonstration of these metrics diverges from conventional practices. These aspects will be elaborated upon in the discussion. The findings provided in Table 2 indicate that denoised and augmented data (DenA) yielded the highest F1-score of 0.64, showcasing the effectiveness of noise reduction and data augmentation in enhancing model sensitivity to sarcasm. Interestingly, both raw and denoised data without augmentation (RawNA and DenNA) achieved identical precision scores of 0.60. However, denoised data without augmentation (DenNA) outperformed its raw counterpart in recall (0.65 vs. 0.59) and F1-score (0.63 vs. 0.60), underscoring the value of denoising in capturing relevant features for sarcasm detection. Augmentation of raw data (RawA) did not significantly enhance recall, which was observed at 0.55, the lowest among the scenarios, yet it improved precision to 0.64, equal to that of the best-performing scenario (DenA).

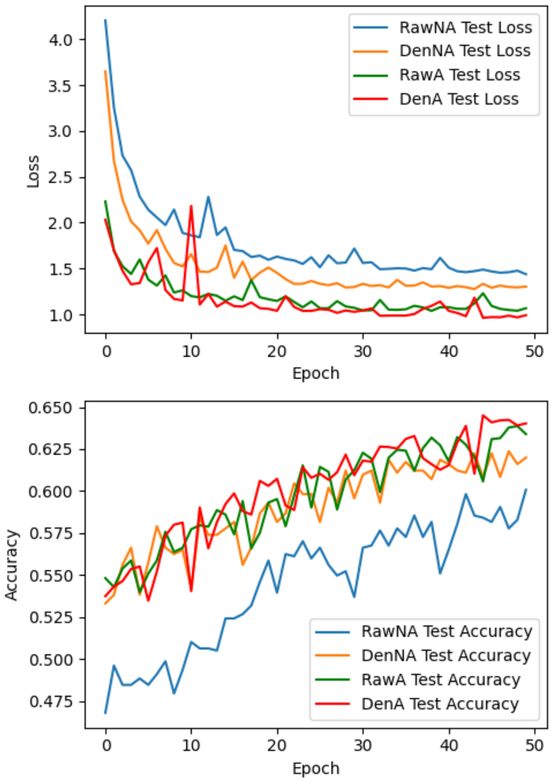

Fig. 3. Accuracy and loss comparison of the test set for all models for 50 epochs, Adam optimizer with Learning rate of 1-e5 and a batch size of 32.

	RawNA	DenNA	RawA	DenA
Precision	0.60	0.60	0.64	0.64
Recall	0.59	0.65	0.55	0.64
F1-score	0.60	0.63	0.60	0.64

Table 2. Performance Comparison for Sarcasm Recognition: Precision, Recall, and F1 Score across Raw and Denoised Data, Both Augmented and Non-Augmented (RawNA, RawA, DenNA, DenA) for 50 epochs, Adam optimizer with Learning rate of 1-e5 and a batch size of 32.

4. DISCUSSION

In this paper, we aimed to replicate and extend the findings of Gao, Nayak, and Coler [1], focusing on the impact of preprocessing on the detection of sarcasm in audio data. Despite our efforts in improving preprocessing techniques, the results did not yield significant enhancements beyond those reported by Gao et al. Specifically, our experiments, as outlined in the results section, confirmed the convergence of loss and accuracy metrics for the test set (Fig. 3), with the denoised and augmented data (DenA) achieving the highest F1-score of 0.64. This marginal improvement underscores the complexity of sarcasm detection and suggests that the preprocessing of audio data, while beneficial, is not the sole determinant of model performance.

Limitations of our study include the homogeneity of our dataset, predominantly sourced from TV series with acted voices and addition, which might limit the generalizability of our findings. The inherent inconsistency in the original audio, including missing segments, variable audio lengths (27.8% more of sarcastic audio), and non-standardized audio levels, presented additional challenges for our deepfilter denoising efforts. Moreover, the exclusive use of this dataset for testing precludes a comprehensive evaluation of our model's generalizability across diverse audio sources.

Our analysis prominently highlighted that overfitting became a significant concern, as demonstrated by loss values exceeding 1.0, indicating a considerable margin of error. This issue was initially thought to stem from the imbalance in audio length between sarcastic and non-sarcastic samples. Efforts to rectify this imbalance by reducing the volume of the more populous sarcastic samples led to consistent underfitting, manifesting as a lack of coherence across various data states, whether non-augmented/augmented or denoised/undenoised. In response, we chose to oversample the less represented non-sarcastic samples, cognizant of the increased risk of overfitting this strategy posed. Nevertheless, this approach resulted in outcomes that were markedly more coherent and consistent. Attempting to bolster these findings, we incorporated a validation split, which inadvertently reduced the volume of data available for both training and testing, subsequently reigniting issues of underfitting.

In our pursuit to both replicate and extend the findings of Gao's study through an evolved preprocessing approach and a novel denoising system, we encountered challenges directly linked to our model's complexity. Specifically, our strategy to refine the model's capability by introducing additional dense layers—a departure from merely unfreezing the top layer for fine-tuning—precipitated overfitting. This outcome suggests that our architectural modifications, aimed at enhancing the model's learning capacity, inadvertently increased its complexity to a level where it could not generalize well to unseen data. This realization underscores the delicate balance required in model architecture adjustments; enhancing model complexity through the addition of dense layers, rather than focusing on strategic fine-tuning within vggish neuronal architecture, may contribute significantly to overfitting.

Despite these challenges, our findings suggest avenues for future research, particularly in exploring the contextual elements of sarcasm. As Bryant and Fox Tree [17] argued, the detection of sarcastic irony may indeed require more comprehensive context information. This insight points towards the potential value of incorporating a broader range of prosodic cues and linguistic context into sarcasm detection models. For future enhancements in the field of sarcasm detection in speech, two pivotal adjustments are proposed. First, prioritizing the analysis of tempo over pitch could unveil more reliable indicators of sarcasm, given that speech tempo changes—like variations in speaking rate or strategic pauses—may signal sarcasm more clearly than pitch alterations. Second, broadening the scope of data collection beyond sitcom audio, which often features biased elements such as recurring laughter, to include a variety of speech contexts will enrich the dataset.

References

[1] X. Gao, S. Nayak, and M. Coler, "Deep CNN-based Inductive Transfer Learning for Sarcasm Detection in Speech," in *Proc. Interspeech 2022*, 2022, pp. 2323–2327. DOI: 10.21437/Interspeech.2022-11323.

[2] R. Gibbs, "Irony in talk among friends," *Metaphor and Symbol - METAPHOR SYMB*, vol. 15, pp. 5–27, Apr. 2000. DOI: 10.1207/s15327868ms151&2_2.

[3] R. J. Kreuz and S. Glucksberg, "How to be sarcastic: The echoic reminder theory of verbal irony.," *Journal of Experimental Psychology: General*, vol. 118, no. 4, pp. 374–386, Dec. 1989. DOI: 10.1037/0096-3445.118.4.374.

[4] H. L. Colston and J. O'Brien, "Contrast and pragmatics in figurative language: Anything understatement can do, irony can do better," *Journal of Pragmatics*, vol. 32, no. 11, pp. 1557–1583, 2000, ISSN: 0378-2166. DOI: 10.1016/s0378-2166(99)00110-1.

[5] P. Brown and S. C. Levinson, *Politeness*. Cambridge University Press, Feb. 1987, ISBN: 9780521313551.

[6] S. Dews, J. Kaplan, and E. Winner, "Why not say it directly? the social functions of irony," *Discourse Processes*, vol. 19, no. 3, pp. 347–367, 1995. DOI: 10.1080/01638539509544922.

[7] S. Dews and E. Winner, "Muting the meaning a social function of irony," *Metaphor and Symbolic Activity*, vol. 10, no. 1, pp. 3–19, 1995. DOI: 10.1207/s15327868ms1001_2.

[8] R. S. Burton, "Principles of pragmatics, geoffrey n. leech. london and new york: Longman, 1983. pp. 250.," *Studies in Second Language Acquisition*, vol. 7, no. 1, pp. 112–113, 1985. DOI: 10.1017/S0272263100005210.

[9] A. Bowes and A. Katz, "When sarcasm stings," *Discourse Processes*, vol. 48, no. 4, pp. 215–236, 2011. DOI: 10.1080/0163853X.2010.532757.

[10] T. Zalla, F. Amsellem, P. Chaste, F. Ervas, M. Leboyer, and M. Champagne-Lavau, "Individuals with autism spectrum disorders do not use social stereotypes in irony comprehension," *PLOS ONE*, vol. 9, no. 4, pp. 1–9, Apr. 2014. DOI: 10.1371/journal.pone.0095568.

[11] E. Winner, *The point of words: Children's understanding of metaphor and irony.* Harvard University Press, Jan. 1988, ISBN: 9780674681262.

[12] E. Winner and H. Gardner, "Metaphor and irony: Two levels of understanding," in *Metaphor and Thought*, A. Ortony, Ed. Cambridge University Press, 1993, pp. 425–444. DOI: 10.1017/cbo9781139173865.021.

[13] P. Rockwell, "Lower, Slower, Louder: Vocal Cues of Sarcasm," *Journal of Psycholinguistic Research*, vol. 29, no. 5, pp. 483–495, Jan. 2000. DOI: 10.1023/a:1005120109296.

[14] D. Voyer and C. Techentin, "Subjective auditory features of sarcasm," *Metaphor and Symbol*, vol. 25, no. 4, pp. 227–242, 2010. DOI: 10.1080/10926488.2010.510927.

[15] H. S. Cheang and M. D. Pell, "The sound of sarcasm," *Speech Communication*, vol. 50, no. 5, pp. 366–381, 2008, ISSN: 0167-6393. DOI: 10.1016/j.specom.2007.11.003.

[16] G. Caucci, R. Kreuz, and E. Buder, "What's a little sarcasm between friends: Exploring the sarcastic tone of voice," *Journal of Language and Social Psychology*, Feb. 2024. DOI: 10.1177/0261927X241233001.

[17] G. Bryant and J. Fox Tree, "Is there an ironic tone of voice?" *Language and speech*, vol. 48, pp. 257–77, Feb. 2005. DOI: 10.1177/00238309050480030101.

[18] S. K. Bharti, R. K. Gupta, P. K. Shukla, W. A. Hatamleh, H. Tarazi, and S. J. Nuagah, "Multimodal Sarcasm Detection: a Deep learning approach," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–10, May 2022. DOI: 10.1155/2022/1653696.

[19] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an $_{o}bviously_{p}erfectpaper$)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, vol. 1, Florence, Italy: Association for Computational Linguistics, 2019. DOI: 10.18653/v1/p19-1455.

[20] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 2018. DOI: 10.1109/icassp.2018.8462417.

[21] F. G. Germain, Q. Chen, and V. Koltun, "Speech Denoising with Deep Feature Losses," in *Proc. Interspeech 2019*, 2019, pp. 2723–2727. DOI: 10.21437/Interspeech.2019-1924.

[22] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, *Deepfilternet: Perceptually motivated real-time speech enhancement*, 2023. DOI: 10.48550/arXiv.2305.08227.