DOI. No. 10.13052/2794-7254.002

# Emotion Detection from Speech: A Comprehensive Approach Using Speech-to-Text Transcription and Ensemble Learning

*Fatima M Inamdar[1,*], Sateesh Ambesange[2], Parikshit Mahalle[1], Nilesh P. Sable[1], Ritesh Bachhav[1], Chaitanya Ganjiwale[1], Shantanu Badmanji[1] and Sarthak Agase[1]*

**Abstract:** The field of text-to-emotion analysis is investigated in this study, which uses an interactive methodology to reveal subtle emotional insights in textual data. The research explores the complex relationship between language and emotion using sophisticated methods without focusing on any particular frontend or backend technology. The research attempts to improve our understanding of how literary information transmits emotional subtleties by emphasizing a broad but methodical examination. The lack of mentions of particular libraries and backends indicates an emphasis on the general ideas and techniques used in text-to-emotion analysis.

The findings demonstrate the possibility of deriving significant emotional context from text, opening doors for applications in a variety of fields where user sentiment analysis is essential. This study adds to the body of knowledge on emotional intelligence in computational linguistics and lays the groundwork for future developments in text analysis techniques.

**Keywords:** Text-to-Emotion analysis, natural language processing, sentiment analysis, emotion recognition, computational linguistics, emotional insights, textual sentiment, emotional nuances, user sentiment, language and emotion.

## 1. Introduction

In the ever-evolving landscape of computational linguistics, the intersection of natural language processing and emotional intelligence has emerged as a focal point of exploration. This research embarks on an insightful journey into the realm of text-to-motion analysis, seeking to unravel the intricate connections between language and sentiment. As the digital age transforms communication paradigms, understanding and interpreting the emotional nuances embedded within textual data becomes paramount. Our investigation delves into the methods and approaches employed in real-time emotion analysis, employing a lens that avoids strict delineation of backend and frontend technologies, fostering a holistic exploration of the field.

The surge in digital content consumption, especially in the context of news and information dissemination, underscores the need for personalized and engaging experiences. Text-to-Emotion Analysis offers a gateway to comprehend not only the explicit meaning but also the underlying sentiments woven into language. This research aims to contribute to the broader discourse on emotional intelligence in computational linguistics by examining how textual content resonates emotionally with users. By acknowledging the dynamic nature of emotional responses, we aspire to pave the way for enhanced user experiences and applications across diverse domains. In the following sections, we delve into the methodologies, insights, and implications drawn from our exploration of text-to-motion analysis.

## 2. Literature Survey

This research introduces the Avatar Affordances Framework, a sophisticated tool designed to map design trends and affordances in character creation interfaces. The framework aims to enhance the understanding of interactive character design across diverse applications. It meticulously examines the evolving landscape of character design interfaces, providing valuable insights for designers, developers, and researchers. By mapping affordances, the authors contribute to the field's theoretical foundations, offering a systematic approach for analyzing and optimizing character creation interfaces. The paper emphasizes the importance of character design in interactive systems and lays the groundwork for future advancements in this dynamic and evolving field [1].

The paper, "Face Detection and Recognition Using OpenCV" by Ramadan TH. Hasan and Amira Bibo Sallow aim to investigate the crucial role of OpenCV in computer vision, specifically focusing on face detection and recognition. The research explores various algorithms, modules, and applications of OpenCV, offering a comprehensive assessment of recent literature. The objective is to emphasize OpenCV's diverse applications and its statistical effectiveness in enhancing human life through image and video processing. The discussed technologies include VGGFace, Attitude Tracking Algorithm (EATA), YCbCr Color Model, and LBPH (Local Binary Pattern Histogram). The paper provides a thorough examination of OpenCV's applications,

[1]Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
[2]Pragyan Smartai Technology Llp, Bangalore, Karnataka, India
E-mail: fatima.inamdar@viit.ac.in; sateesh.ambesange@gmail.com; parikshit.mahalle@viit.ac.in; riteshbachhav251@gmail.com; chaitanyaganjiwale812@gmail.com; shantanubadmanji1912@gmail.com; sarthak.agase@gmail.com
*Corresponding Author
Manuscript received 07 November 2023, accepted 02 February 2024, and ready for publication 21 March 2024.

shedding light on face detection, recognition, and real-world implementations. However, it falls short of providing an in-depth exploration of certain algorithms, and the literature reviews are confined to specific timeframes, potentially missing recent advancements in the field [2].

This research introduces a groundbreaking model of a humanoid robot equipped with facial expression capabilities and natural language processing (NLP) skills. The authors showcase the practical implementation of this novel model, emphasizing the robot's ability to recognize objects and convey emotions through facial expressions. The integration of NLP suggests a sophisticated level of interaction, allowing the robot to understand and respond to human language. The implementation of such a robot holds promise for advancements in human-robot interaction, particularly in scenarios where robots are expected to comprehend and express emotions, enhancing their utility in various contexts, from customer service to educational settings [3].

Focusing on human-computer interactions, this research presents an emotionally responsive avatar with dynamic facial expressions. The authors delve into the intricacies of designing an avatar capable of conveying emotions responsively. By incorporating dynamic facial expressions, the avatar enhances the interactive experience, allowing for more nuanced and emotionally resonant digital interactions. This work contributes to the growing field of affective computing, where the goal is to imbue computers and digital interfaces with the ability to recognize and respond to human emotions. The advancements presented in this paper have implications for the design of emotionally engaging avatars in applications ranging from virtual assistants to gaming [4].

Addressing challenges in 3D avatar generation, this research proposes a novel method guided by text and image prompts. Leveraging the Contrastive Language Image Pre-training (CLIP) model and a pre-trained 3D Generative Adversarial Network (GAN), the approach enables efficient and high-level manipulations of shape and texture in 3D avatars. The use of CLIP allows for contextual understanding from textual and visual prompts, facilitating more intuitive and user-friendly avatar manipulation. This work contributes to the field of computer graphics and virtual environments, providing a pathway for creating personalized and expressive 3D avatars for various applications, including virtual worlds, gaming, and digital communication [5].

This research focuses on Persian texts, presenting a speech act classifier and its application in identifying rumors. The authors address the challenge of classifying speech acts in Persian language texts, contributing to natural language processing tasks specific to this linguistic context. The application in rumor identification highlights the broader societal implications of the research, as it provides a tool for automatically discerning linguistic expressions associated with rumors. This work contributes not only to the field of natural language processing but also to misinformation detection, offering a valuable resource for analyzing and mitigating the spread of rumors in Persian-language digital communication [6].

The authors systematically review emotion recognition and detection methods in this comprehensive survey. Covering a wide range of approaches and techniques, the survey serves as a valuable resource for researchers in the field of affective computing. The paper categorizes and analyzes various methods employed for recognizing and detecting emotions, considering modalities such as facial expressions, speech, and physiological signals. This work not only provides an overview of the current state of the art but also identifies trends and challenges in emotion recognition, paving the way for future research and development in this dynamic and evolving field [7].

This research delves into linguistic features to explore the automatic identification of mental states from language text. The authors propose a method to read "mindprints" by analyzing deeper linguistic features, providing insights into the cognitive and emotional states reflected in language. The work contributes to understanding the intricate relationship between language and cognition, offering a nuanced perspective on how linguistic expressions can serve as indicators of mental states. By leveraging deeper linguistic analysis, the research expands the toolkit for studying and interpreting mental states through textual communication, opening avenues for further exploration in cognitive science and natural language processing [8].

This research presents a framework for designing emotionally realistic chatbots, aiming to enhance their believability. Leveraging AIML (Artificial Intelligence Markup Language) and information states, the framework focuses on imbuing chatbots with emotional expression and responsiveness. By incorporating elements of emotion into chatbot interactions, the authors seek to create a more engaging and believable conversational experience. The framework's potential applications range from customer service bots to virtual companions, where emotional intelligence is increasingly recognized as a valuable aspect of human-computer interaction. The work contributes to the development of more sophisticated and emotionally resonant conversational agents in the field of artificial intelligence [9].

## 3. Existing System and Algorithm

Different approaches are used in the field of text-to-emotion analysis to interpret the emotional undertones included in textual data. Lexicon-based methods use sentiment lexicons or dictionaries to identify the overall emotional tone of a text by giving words sentiment ratings and adding them together. Though basic, these approaches could have trouble with subtleties and expressions that depend on context. However, machine learning models which include deep learning architectures and conventional classifiers use labeled datasets to identify complex patterns and forecast the emotional content of text. Although they need a large amount of training data, these models are excellent at capturing context-dependent emotions and intricate interactions.

Natural Language Processing (NLP) techniques contribute significantly to text-to-emotion analysis by extracting meaningful features from the text. Techniques such as part-of-speech tagging and syntactic analysis aid in understanding the structural and grammatical elements that contribute to the emotional context of the text. Additionally, word embeddings, such as Word2Vec and GloVe, provide continuous vector representations of words, capturing semantic relationships and enriching the understanding of emotional nuances.

Furthermore, rule-based systems explicitly define linguistic rules for identifying emotional content, offering transparency and interpretability. However, these systems may face challenges in handling diverse expressions and linguistic variability. Hybrid approaches, combining rule-based components with machine learning models, aim to harness the interpretability of rule-based

systems and the predictive power of machine learning, offering a potential middle ground for accurate and robust emotion analysis. The choice of methodology depends on the specific requirements, nuances, and context of the text data under consideration.

## 4. Proposed Methodology

This research paper proposes a comprehensive methodology for emotion detection from speech, leveraging the capabilities of speech-to-text transcription and ensemble learning techniques. The initial phase involves the use of open-source speech-to-text models, particularly the OpenAI Whisper model, to accurately transcribe spoken words into textual representations. These transcriptions undergo preprocessing to address challenges such as noise and accents. Subsequently, the text is processed using natural language processing techniques, including tokenization, normalization, and the implementation of a bag-of-words model to capture emotional content effectively.

In the next step, machine learning classification algorithms are employed to categorize emotions within the transcriptions. Selected algorithms, such as Support Vector Machines, Random Forest, and Naive Bayes, are trained and evaluated using a divided dataset for training and testing purposes. Hyperparameter tuning and optimization are performed through cross-validation to enhance model performance. To further improve accuracy, Ensemble Methods for Classification are explored, combining predictions from multiple base models using techniques like Voting, Bagging, and Boosting.

The evaluation of the proposed methodology relies on standard metrics such as accuracy, precision, recall, and F1 score. Comparative analyses are conducted between individual machine learning models and the ensemble model to discern the impact of ensemble methods on overall performance. The results obtained from these experiments, along with discussions on the strengths and limitations of the methodology, form the core of the findings. Practical implications and potential applications of the proposed approach are deliberated, concluding with suggestions for future research directions in the domain of emotion detection from speech. The paper is substantiated by references to relevant literature, previous studies, and resources that informed the research methodology and analysis.

## 5. Results

In the pursuit of emotion detection from speech, the research employed a methodology encompassing speech-to-text transcription using the OpenAI Whisper model, subsequent text preprocessing with Natural Language Processing (NLP) techniques facilitated by NLTK tools, and the transformation of text into a bag-of-words model for training machine learning classification models. The dataset chosen for evaluation was the "Emotions dataset for NLP," consisting of 23,000 sentences, each associated with one of the six emotions: anger, fear, happiness, love, sadness, and surprise.

To assess the individual performance of various machine learning models, accuracy scores were computed for each emotion detection task. The results are detailed as follows:

- Ensemble Classifier: 85.89%
- Neural Network: 80.82%
- GradientBoostingClassifier: 78.68%
- LGBM_classifier: 84.05%
- CatBoost: 84.87%
- XGB: 84.89%
- ExtraTreesClassifier: 85.06%
- Random Forest: 83.69%
- DecisionTree: 79.75%
- LogisticRegression: 84.61%
- Kernel SVM: 81.76%
- Linear SVM: 83.39%
- SVM: 81.76%
- KNeighbours: 53.85%
- Naive Bayes: 37.71%
- AdaBoostClassifier: 39.25%

Accuracy = 85.8947%
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| anger | 0.82 | 0.80 | 0.81 | 683 |
| fear | 0.85 | 0.83 | 0.84 | 520 |
| happy | 0.85 | 0.91 | 0.88 | 1579 |
| love | 0.83 | 0.74 | 0.78 | 303 |
| sadness | 0.90 | 0.88 | 0.89 | 1403 |
| surprise | 0.85 | 0.74 | 0.79 | 184 |
| accuracy | | | 0.86 | 4672 |
| macro avg | 0.85 | 0.82 | 0.83 | 4672 |
| weighted avg | 0.86 | 0.86 | 0.86 | 4672 |

**Figure 1.**
Ensemble model classification report.

Furthermore, the Ensemble Model for Classification, designed to aggregate predictions from multiple base models, demonstrated an elevated accuracy score of 0.859. This result underscores the effectiveness of the ensemble approach in surpassing the performance of individual models, showcasing its potential for accurate and robust emotion detection from speech. The ensemble model's ability to harmonize diverse classifiers and leverage their collective insights contributes significantly to the overall success of the proposed methodology.

These outcomes underscore the potential practical applications of the research in real-world contexts where nuanced emotion analysis from spoken language is essential. The detailed accuracy scores offer valuable insights into the comparative performance of various machine learning models, affirming the efficacy of the proposed methodology in enhancing emotion detection accuracy.

## 6. Future Work

The present research on emotion detection from speech lays the foundation for compelling avenues of future exploration. One prominent trajectory involves extending our current methodology to encompass AI-driven video emotion enhancement. This endeavor seeks to leverage identified emotions from speech to dynamically enhance emotional content within corresponding

video footage. Emphasis will be placed on mapping detected emotions to appropriate facial expressions and gestures, thereby elevating the emotional expressiveness of video content. Additionally, we envision integrating the developed emotion detection model with AI News Anchors, aiming to generate or edit videos dynamically. This endeavor involves rendering video content with appropriately matched emotional expressions and gestures based on the emotions detected from the speech or transcript of the news anchor.

Further directions include the exploration of real-time emotion detection and mapping for live speech scenarios, with potential applications in live broadcasts, interviews, and interactive video content. To enhance accuracy, we propose investigating advanced ensemble methods and their combination with deep learning approaches. This includes the exploration of neural network-based models within ensemble frameworks to optimize performance. The extension of our methodology to include multimodal features, incorporating both speech and facial expressions for a more comprehensive understanding of emotional states, is another avenue of interest.

Assessing the resilience of our approach in various linguistic and cultural situations is crucial, as we acknowledge. Increasing the dataset's diversity in terms of language and culture is essential to improving the model's capacity for generalization. To guide changes based on real-world user experiences, user feedback, and experience studies will be carried out to evaluate the perceived efficacy and accuracy of the emotion detection model in practical applications.

Furthermore, ethical considerations associated with emotion detection technology, such as privacy concerns and potential biases, will be addressed. Measures will be implemented to ensure fair and unbiased emotion analysis across various demographic groups. This collective future work aspires to extend the impact of our research into practical applications, revolutionizing the portrayal and perception of emotional content in multimedia formats, particularly in the emerging domain of AI News Anchors and video emotion enhancement.

## 7. Conclusion

In conclusion, this research endeavors to advance the field of emotion detection from speech by leveraging cutting-edge technologies and methodologies. The utilization of open-source models, particularly the Open AI Whisper model, for speech-to-text transcription, has proven to be a foundational step in transforming spoken words into a textual format. The subsequent processing of this text involves sophisticated Natural Language Processing (NLP) techniques, including the implementation of a bag-of-words model, to extract and categorize emotional states.

The integration of machine learning classification algorithms further enhances the accuracy of emotion detection. Our evaluation encompassed a diverse set of classifiers, demonstrating the efficacy of models such as CatBoost, XGBoost, and ExtraTreesClassifier in accurately discerning emotions from speech. To refine our approach and achieve even greater precision, Ensemble Methods for Classification were employed, showcasing the potential of combining multiple algorithms for improved performance.

Looking forward, the trajectory of our research extends beyond the realms of speech analysis into AI video emotion enhancement. Our vision encompasses the dynamic regeneration of video content based on detected emotions, a significant stride towards more emotionally resonant multimedia experiences. We anticipate transformative applications in content creation, storytelling, and media production by fusing the insights gained from speech analysis with AI-driven tools for video emotion enhancement.

In essence, this research not only contributes to the understanding of emotion detection from speech but also lays the groundwork for innovative applications in the multimedia domain. The fusion of speech-to-text transcription, NLP, machine learning, and the prospect of AI video emotion enhancement opens doors to a future where emotional content is discerned and dynamically translated into rich and expressive visual narratives. As we embark on the journey of implementing our methodology into practical video enhancement tools, we anticipate the continued evolution of emotion detection technologies and their profound impact on the way we experience and interact with multimedia content.

## References


[1] McArthur, V., Teather, R. J., and Jenson, J. (2015). The Avatar Affordances Framework: Mapping Affordances and Design Trends in Character Creation Interfaces. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (pp. 231–240). ACM. doi: 10.1145/2793107.2793121.

[2] Hasan, R. Th., Sallow, A. B., and Hasan, A. B. (2021). Face Detection and Recognition Using OpenCV. *JSCDM, 2*(2). doi: 10.30880/jscdm. 2021.02.02.008.

[3] Budiharto, W., Andreas, V., and Gunawan, A. A. S. (2021). A Novel Model and Implementation of Humanoid Robot with Facial Expression and Natural Language Processing (NLP). *ICIC International 学会*. doi: 10.24507/icicelb.12.03.275.

[4] Wang, H., Gaddy, V., Beveridge, J. R., and Ortega, F. R. (2021). Building an Emotionally Responsive Avatar with Dynamic Facial Expressions in Human–Computer Interactions. *MTI, 5*(3), 13. doi: 10.3390/ mti5030013.

[5] Canfes, Z., Atasoy, M. F., Dirik, A., and Yanardag, P. (2023). Text and Image Guided 3D Avatar Generation and Manipulation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 4410–4420). IEEE. doi: 10.1109/WACV56688.2023.00440.

[6] Jahanbakhsh-Nagadeh, Z., Feizi-Derakhshi, M.-R., and Sharifi, A. (2019). A Speech Act Classifier for Persian Texts and its Application in Identifying Rumors. doi: 10.48550/ARXIV.1901.03904.

[7] Saxena, A., Khanna, A., and Gupta, D. (2020). Emotion Recognition and Detection Methods: A Comprehensive Survey. *AIS, 2*(1), 53–79. doi: 10.33969/AIS.2020.21005.

[8] Pearl, L. S., and Enverga, I. (2014). Can you read my mind print?: Automatically identifying mental states from language text using deeper linguistic features. *IS, 15*(3), 359–387. doi: 10.1075/is.15.3.01pea.

[9] Sutoyo, R., Chowanda, A., Kurniati, A., and Wongso, R. (2019). Designing an Emotionally Realistic Chatbot Framework to Enhance Its Believability with AIML and Information States. *Procedia Computer Science, 157*, 621–628. doi: 10.1016/j.procs.2019.08.226.

[10] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... and Marchi, E. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings of Interspeech* (pp. 148–152). doi: 10.21437/ Interspeech.2013-70.

[11] Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572–587. doi: 10.1016/j.patcog.2010.09 .020.

[12] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... and Tavabi, L. (2015). The Geneva Minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), 190–202. doi: 10.1109/TAFFC.2015.2457417.

[13] Lee, C. M., and Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing, 13*(2), 293–303. doi: 10.1109/TSA.2004.840609.

[14] El Ayadi, M., and Kamel, M. S. (2007). A new framework for audio–visual speech emotion recognition. *Pattern Recognition, 40*(6), 1674–1687. doi: 10.1016/j.patcog.2006.11.029.

[15] Busso, C., Lee, S., and Narayanan, S. S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(4), 582–596. doi: 10.1109/TASL.2008.2010426.

[16] Deng, J., Zhang, Z., Marchi, E., Schuller, B., and Wu, D. (2013). Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *Proceedings of Interspeech* (pp. 236–240). doi: 10.21437/Interspeech.2013-79.

[17] Koolagudi, S. G., and Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology, 15*(2), 99–117. doi: 10.1007/s10772-012-9157-x.

[18] Lotfian, R., and Saeidi, R. (2016). Speech emotion recognition using hidden Markov models. *Speech Communication, 77*, 1–17. doi: 10.1016/j.specom.2015.10.010.

[19] Li, C., and Deng, J. (2014). Emotion recognition from speech signals using new harmony features. *IEEE Transactions on Multimedia, 16*(7), 1904–1916. doi: 10.1109/TMM.2014.2323633.

## Biographies

**Fatima M Inamdar**, Assistant Professor at Vishwakarma Institute of Information Technology, holds 18 years of teaching and 7 months of industry expertise. Proficient in Software Engineering, Data Science, AI-ML+DL, Web Technology, she's a Fellow at Eudoxia Research University. Recognized for research excellence, awards from SAW and the Global Research Foundation underscore her impact. The Women Leaders Forum award highlights her commitment to medical sector challenges.

**Sateesh Ambesange**, with 20+ years of industry expertise, is presently linked with PRAGYAN SMARTAI TECHNOLOGY LLP. A committed researcher in Data Science, Machine Learning, Deep Learning, Natural Language Processing (NLP), and Generative AI, he excels in these domains. As a leader, he takes pride in contributing significantly to advancements in artificial intelligence and data-driven solutions.

**Parikshit Mahalle**, a senior IEEE member and Professor at Vishwakarma Institute of Information Technology, Pune, holds 23+ years in teaching and research. With a Ph.D. from Aalborg University, Denmark, and a postdoc at CMI, Copenhagen, he boasts 15 patents, 200+ publications, and 59 books. As an editorial leader at IGI Global's International Journal of Rough Sets and Data Analysis, he significantly influences Machine Learning, Data Science, and IoT, earning accolades like the "Best Faculty Award".

**Nilesh P. Sable** is a senior member IEEE and working as Associate Professor, Head Department of Computer Science & Engineering (Artificial Intelligence) at Vishwakarma Institute of Information Technology, Pune, India. He has 16+ years of teaching and research experience. He is guiding 4 Ph.D. students in the area of Machine Learning, Federated Learning and IoT under his supervision from SPPU. He is working as Research Advisory Committee (RAC) Member for various Research Centres. His research interests are Data Mining, Image Processing,

Machine Learning, Cognitive Computing, Internet of Things, Networking and Security. He has published 75+ papers in National, International conferences and Journals. He had Filed and Published 16 Patents and 18 Copyrights. He has authored books published by National/International publishers.

**Ritesh Bachhav**, currently pursuing a B.Tech in Information Technology at Vishwakarma Institute of Information Technology, seamlessly blends academic excellence with active participation in hackathons and coding competitions. Driven by a profound passion for technology, he specializes in Android development and Blockchain.

**Chaitanya Ganjiwale**, currently pursuing a B.Tech in Information Technology at Vishwakarma Institute of Information Technology, actively engages in competitive coding, hackathons, and coding competitions. Passionate about technology, especially Android development, networking, and machine learning, I possess a quick-learning mindset and an unwavering commitment to excellence, thriving in diverse challenges and opportunities.

**Shantanu Badmanji**, a graduate pursuing a B.Tech in Information Technology from Vishwakarma Institute of Information Technology, seamlessly integrates a robust academic background with active involvement in hackathons and coding competitions. Fueled by a passion for technology, an innovative mindset, and a commitment to excellence, I am poised to be a promising contributor to the IT industry.

**Sarthak Agase**, pursuing B.Tech in IT from VIIT Pune, looking forward to improving my managerial abilities and leadership skills, interested in Machine Learning, Web Development and Blockchain Technologies.