International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)

ISSN: 2456-236X

Vol. 10 Issue 01 | 2025

A Survey of Deep Learning Techniques for Speech Emotion Recognition

Mr. Prasad K. Kajale[1], Prof. Manjiri Karande[2]

[1,2] *Department of Computer Engineering, DR. V.B. COE, Maharashtra, India*

DOI: 10.5281/zenodo.15751350

ABSTRACT

Speech emotion recognition (SER) is vital for interpreting a speaker's emotional state through audio analysis. By categorizing emotions like anger, happiness, and sadness, SER unlocks insights into thoughts and well-being. Leveraging speech as a fundamental and efficient mode of communication, extensive research has focused on developing automated voice-based emotion recognition systems to enhance human-machine interaction. Voice is increasingly central to Human-Machine Interfaces, a multidisciplinary field integrating computer science, signal processing, and psychology. Advancements in this area enable more seamless communication, with speech recognition capturing both linguistic content and emotional nuances, establishing its significance in human-machine communication systems.

Keywords: - SER, SVM, ML, speech, deep learning, CNN.

1. INTRODUCTION

Emotions are deeply woven into our lives, influencing everything we do and how we see the world. We express these feelings in many ways – through our words, the look on our faces, and even our gestures. Understanding these emotions often starts with listening to the way someone speaks.

There's strong evidence that when we feel different emotions, our bodies react in ways that change our voice. For instance, if someone is angry, their breathing might change, their muscles might tense up, and even the shape of their throat can shift. These physical changes then affect how their vocal cords vibrate and the actual sound of their speech. So, the way we speak is closely tied to our emotions, and by recognizing these emotional cues in our voices, we can learn a lot about human communication and behavior.

Speech is our go-to way of connecting, and it's packed with more than just words. Things like body language, gestures, facial expressions, and our tone of voice all help us quickly share information. They give us instant clues about someone's age, gender, and more. Speech and emotion recognition (SER) is a fascinating area of computer science that's taken off in recent years. It has tons of potential in areas like smart homes, social media, education, healthcare, and lots of other AI applications. It gives us a way to automatically understand the emotions in speech. Exciting new approaches in SER research, like using clever training methods to create diverse and realistic speech data, are helping to make these systems even better. This progress promises more accurate and dependable emotion recognition, leading to much better interactions between humans and machines.

One of the trickiest parts of SER is figuring out the best things to listen for in speech to accurately identify emotions. Raw speech signals don't always clearly distinguish between different feelings. SER is a tech-heavy process that uses advanced techniques to spot human emotions in recordings or live situations. This not only highlights the amazing progress in machine learning but also has big implications for industries like entertainment, healthcare, and how we interact with customers.

Consider the strong link between our emotional state and our physiology. When we feel an emotion, like anger, it often triggers physical changes – our breathing might alter, our muscles could tense, and even the shape of our vocal tract can change. These physiological shifts directly impact how our vocal cords vibrate and the acoustic properties of our speech. Essentially, our emotions are embedded in how we speak, and deciphering these vocal cues offers valuable insights into human communication and behavior.

Speech stands out as our most prevalent form of communication, a rich medium carrying a wealth of information beyond just the words themselves. Elements like body language, gestures, facial expressions, and crucially, our voice and tone, all contribute to quickly conveying information. They allow us to instantly glean details about others, such as their age or perceived gender. Speech and emotion recognition (SER) has become a dynamic and rapidly growing area within computer science research over the past few decades. Its applications are diverse, spanning smart homes, social media, education, healthcare, and numerous AI-driven applications, providing a

framework to automatically identify emotions in spoken language. Innovative approaches in SER, such as using advanced training techniques to generate varied and realistic speech data, are expanding training datasets and boosting system performance. These advancements promise to further refine the accuracy and reliability of automated emotion recognition systems, paving the way for more intuitive human-machine interactions.

A significant challenge in SER lies in identifying effective features within speech that reliably indicate different emotional states. Raw speech signals don't always clearly differentiate between emotions. SER is a technology-driven process that employs sophisticated techniques to discern human emotions from recorded speech or in real-time scenarios. This not only showcases the remarkable progress in machine learning but also carries significant implications across various industries, including entertainment, healthcare, and customer service. By enabling machines to understand and respond to human emotions, we are laying the groundwork for more empathetic and contextually aware AI systems. This capability enhances our ability to engage with machines on a deeper emotional level, unlocking new possibilities for human-computer interaction. This comprehensive survey will delve into the key methodologies, datasets, and cutting-edge techniques employed in speech emotion recognition. Our primary goal is to explore how machines can interpret the intricate web of human emotions embedded in spoken words. Along this path, we will examine everything from the fundamental principles of extracting acoustic features to the latest advancements in deep learning architectures. Teaching machines to recognize and react to human emotions is paving the way for a future characterized by more understanding and context-sensitive AI.

Emotions manifest in various ways, including facial expressions, tone of voice, and body language. This understanding has implications not only for emotion recognition but also for related fields like emotional computing and human-computer interaction. Developing an effective SER framework requires precise calculations and the use of specific datasets to train machines to identify and categorize emotions based on word choice and vocal tone. Bridging the gap between acoustic features (related to sound intensity and frequency patterns) and human emotions (e.g., happiness, sadness) makes automated SER a complex undertaking, heavily dependent on capturing discernible acoustic features relevant to the task. As we navigate this field, we will also address the ethical considerations surrounding emotion recognition technology, its potential societal impacts, and the critical need for responsible development and deployment.

Furthermore, we will explore real-world applications where SER is already making a positive impact, ranging from mental health support systems to virtual assistants that adapt to users' emotional states. This research not only deepens our understanding of the field but also illuminates its broader implications for society and human interaction.

Moving on, Section II lays the groundwork with a Background Study, providing essential context for the research. Following this, Section III delves into Related Work, exploring what others have already done in this area. Then, Section IV offers an Analysis and Discussion of the Related Work presented in the previous section, examining its strengths, weaknesses, and relevance. Section V then turns its attention to the Challenges involved in encrypting an image, highlighting the complexities of this task. Finally, Section VI wraps everything up with the Conclusion of the paper, summarizing the key findings and insights.

2. BACKGROUND STUDY

Speech Emotion Recognition (SER) has taken off as an exciting area of research and development. This is largely because we're all looking for ways to make our interactions with computers and technology feel more natural and human-like. At the heart of this is the idea that if machines can understand our emotions when we speak, they can become much more empathetic and responsive to our needs. Think of SER systems as being designed to automatically listen to our voices, figure out the emotions we're expressing, and then categorize those feelings. This technology isn't just a cool idea; it's finding its way into all sorts of places, like making customer service interactions smoother, helping with mental health assessments, improving how we interact with robots, and so much more.

The whole field of SER is built on a strong foundation of knowledge from different areas, like how we process speech, analyze signals, use machine learning, and even understand psychology. Researchers have been digging deep into the connection between the way our voices sound, the rhythm and melody of our speech, and the emotions we're feeling. This has led to the creation of some pretty sophisticated algorithms and models that can pick up on emotional cues in our speech. As SER technology gets better, it's not only benefiting from having more and more diverse data to learn from, but also from using powerful deep learning techniques, which allow for more accurate and nuanced emotion recognition. The importance of SER goes way beyond just making our gadgets friendlier. It has some really meaningful applications, especially in healthcare. For example, it could help doctors diagnose and keep an eye on mental health conditions by analyzing subtle changes in someone's speech that might indicate depression, anxiety, or other emotional states. In education, SER could be used to see how engaged students are and help teachers tailor their lessons accordingly. Plus, as we have more and more smart devices and connect everything

through the Internet of Things, SER can make these systems more adaptable and responsive to our emotional needs, making them feel more like helpful companions.

Of course, there are still some hurdles to overcome in SER research. One challenge is that people from different cultures might express the same emotion in slightly different ways. We also need to make sure these systems can recognize emotions in real-time and that we're being ethical about privacy and getting people's consent when we analyze their speech. As SER technology matures and becomes more common, tackling these challenges will be key. Ultimately, with speech and emotion recognition becoming more integrated into our daily lives, SER is set to play a crucial role in shaping the future of how we interact with technology, making it more intuitive, empathetic, and in tune with our human emotional experiences.

3. LITERATURE SURVEY

At its core, a typical Speech Emotion Recognition (SER) system often relies on a Convolutional Neural Network (CNN) algorithm. Think of this as having two main jobs: first, figuring out *if* there's an emotion in the speech, and second, classifying *what* that emotion is, like happiness, surprise, anger, neutrality, or sadness. These systems usually learn from datasets full of speech samples.

For instance, Apoorv Singh and his team [9] proposed a model that uses CNNs for SER and thought about putting it into robots and music apps. The idea is that if a robot can understand your mood, it can have a much better conversation with you. And for music apps, the system could recommend songs that match how you're feeling. To figure out the best way to distinguish between emotions, they built and tested several CNN models and found one that was 71% accurate. They pointed out that it could be even better if it could tell the difference between male and female voices. Overall, they showed how adding CNN-based SER to robots and music apps could make the user experience much more engaging by making the technology emotionally aware and able to respond appropriately.

Another researcher, Huihui and colleagues [10], came up with a clever way to boost the performance of regular CNNs to better pick out emotion-related features. They called their model ICNN. The main innovation was using "interactive convolution" on feature maps of different scales, which helped them achieve a pretty good accuracy of 76% in recognizing emotions. Specifically, they split the Mel-frequency cepstral coefficients (MFCs) – a way to represent audio – into two parallel streams using their ICNN process. The general idea here is that combining the strengths of traditional audio features like MFCCs with the power of deep learning CNNs can lead to better and more reliable performance in understanding audio and speech.

Then, Taiba and her team [11] focused on really nailing down how to identify emotions and build strong methods for detecting them. They introduced a new approach called Deep Stride Convolutional Neural Networks (DSCNN). This was a twist on regular CNNs, where they got rid of the "pooling" layers and instead used "strides" to shrink down the feature maps more effectively. To see how well their DSCNN model worked compared to a standard CNN, they ran two experiments. In both, they found that both models got better as they trained for longer. However, the DSCNN model outperformed the regular CNN, reaching an accuracy of 87.8% compared to the CNN's 79.4%. This suggests that for really good emotion detection, regular CNNs might need some architectural improvements.

Finally, Ming-Hsiang Su and his colleagues [12] presented a novel way to recognize emotions in conversations by considering both the words spoken and the nonverbal sounds. They achieved an accuracy of 68.87%. Their approach used a support vector machine-based detector along with deep residual networks and a special type of recurrent neural network called a mindful long short-term memory network to get these results on a specific dataset. They found that both the type of sound and nonverbal vocalizations helped recognize emotions and improved the accuracy of their method. This method is also good at handling large amounts of data and helps prevent the model from overfitting by using "skip connections."

Chenghao Zhang and his team [13] came up with a clever way to automatically pull out the most important features related to emotions in speech. They called their method "Autoencoder with Emotion Embedding." Think of an autoencoder as a way for a computer to learn the most crucial aspects of something. By adding "emotion embedding," they could also teach it to understand how the emotion labels relate to these features. Their goal was to improve how well SER systems work (they got an accuracy of 65.76%) by being smarter about which parts of the speech data are most important for recognizing emotions. An autoencoder with emotion embedding is a neural network that's not just good at learning compressed representations of data but can also understand and represent emotional information.

Researchers Jorge Oliveira and his colleagues [14] explored whether we could use the knowledge that pre-trained speech recognition systems already have to help identify emotions in the workplace. They aimed to make manufacturing processes better and boost efficiency by reducing negative emotions. They achieved a pretty impressive accuracy of 89.43%. Their work involved training and testing their method on a dataset called Merge, which contains recordings from the TV show "Friends." They used a "weighted strategy" and found that these pre-

trained speech recognition layers are a really good starting point for building accurate and efficient speech processing systems for all sorts of applications. This approach, called transfer learning, can also help when you don't have a huge amount of data specifically for emotion recognition.

Julia Sidorova and her team [15] focused on creating a system that could automatically assess the emotional competence of patients with neurological conditions, particularly those with Foreign Accent Syndrome (FAS). They reached an accuracy of 87.89% using what they called an "Aggregated Ear model" to understand the patients' emotional abilities. They also provided detailed information about the different therapies used and their expected effects on the brain. "Aggregated Ear Models," or "Ensembled Ear Models," is like getting a second opinion (or even more!) from multiple machine learning models to get a better overall result, especially in tasks related to the ear, like recognizing a person by their ear shape. It's worth noting that how well these combined models work depends on the specific task, the quality and variety of the individual models, and how they're combined (like through voting or averaging). This approach can also be helpful when dealing with uneven datasets and trying to make the system work well in real-world situations.

Finally, S. Hamsa and colleagues [16] set out to build an AI system that could accurately identify a speaker's emotional state even when there's a lot of noise and interference. They achieved an impressive accuracy of 90.76%. Their innovative method uses a new framework for emotion recognition that looks at important aspects of speech like energy, how it changes over time, and its spectral features (the different frequencies present). To do this, they used a "random forest classifier" along with a "wavelet packet transform-based cochlear filter bank" (WPT-CFB). Combining WPT-CFB with a Random Forest classifier has several advantages in processing audio and speech. However, it's important to consider the type of data you have and how much labeled data is available for training.

Mauajama Firdaus and her team [29] have been doing some really interesting work on how sentiment and emotion together influence how computers generate dialogue in situations where there's more than just text involved (like video and audio too). 1 They looked at how both the general feeling (sentiment) and specific emotions play a role in shaping conversations. To make their system better, they used all sorts of data, including text, video, audio, reinforcement learning (where the system learns from rewards), and even user feedback. Impressively, their model achieved an accuracy of 89.76%. This kind of approach has the potential to be useful in a wide range of applications.

Mohammad Ariff Rashidan and his colleagues [30] proposed a model for understanding emotional states through voice analysis in consumer electronics. They did point out that their work had some limitations in terms of what they looked at, the specific datasets they used, the period of their research, and their criteria for including or excluding information. Their main focus was on recognizing emotions in technology-assisted communication with children who have autism.

4. ANALYSIS AND DISCUSSIONS

It's clear that Speech Emotion Recognition (SER) has come a long way in the last ten years. With more and more devices listening to our voices and a growing interest in understanding people's feelings, the need for good SER systems has taken off.

Speech Emotion Recognition has become a hot topic lately because it can be used in so many different areas. Think about how we interact with computers, virtual assistants like Siri or Alexa, trying to understand the sentiment behind social media posts, and even helping to assess someone's mental well-being. This technology is all about automatically figuring out and categorizing the emotions we express when we speak.

Table: Summary of Emotion/SER Algorithms and Their Performance

Authors	Model/Algorithm	Accuracy (%)
Apoorv et al. (2020)	CNN algorithm for SER with integration into robots and music apps	71.00
Huihui et al. (2020)	ICNN with MFCC and LMFCC	76.00
Taiba et al. (2020)	Modified CNN model	87.80
Ming et al. (2021)	SVM detector, deep residual networks, and attentive LSTM sequence-to-sequence model	68.87
C. Zhang et al. (2021)	Autoencoder with Emotion Embedding	65.76
Julia et al. (2021)	Aggregated EAR model	87.89
S. Hamsa et al. (2020)	Random forest classifier	90.76
Parthasarathy et al. (2020)	Semi-supervised ladder networks	65.67
Sudarsana et al. (2020)	Excitation features around glottal closure instants	86.96

Siddique et al. (2020)	Multi-task learning framework	65.43
Chang et al. (2021)	Two-level feature fusion	69.30
Peng et al. (2021)	Front-end auditory perception + attention-based back-end	85.46
Zhao et al. (2020)	SSGAN and VSSSGAN methods	61.25
Awotunde et al. (2023)	CNN-based method for speech segregation	78.52
Wei Sun et al. (2020)	Deep learning with speaker gender	95.65
Guang et al. (2020)	Non-contact emotion recognition using video-based features	67.43
Qiugang et al. (2022)	Neural networks on large-scale datasets	65.48
Zhen et al. (2021)	Speech personality recognition with audio features	76.43
Mauajjama et al. (2022)	Sentiment and emotion-controlled dialogue	89.76
Mohammad et al. (2021)	Affective states in speech analysis for consumer electronics	97.50

This table provides a comparative overview of various speech emotion recognition (SER) and sentiment analysis techniques, along with their performance in terms of accuracy. From the data, a few key trends and insights emerge:

1. **Deep Learning Dominance:** Approaches that utilize deep learning models, particularly convolutional neural networks (CNNs) and their variants, tend to achieve higher accuracy. For instance, the models by Taiba et al. (87.8%), Julia et al. (87.89%), and Mohammad et al. (97.5%) highlight the effectiveness of CNN-based architectures in emotion recognition tasks.
2. **Multimodal and Ensemble Approaches:** The highest accuracy was recorded by Mohammad et al. (2021) at 97.5%, using affective state analysis likely combined with multimodal data. Similarly, high-performance models often involve hybrid or ensemble learning strategies, such as aggregated models and attention mechanisms.
3. **Classical Machine Learning vs. Deep Learning:** Traditional machine learning techniques like Random Forest (S. Hamsa et al.) surprisingly outperform some deep learning methods, achieving 90.76% accuracy. This suggests that for certain datasets or features, classical models can still be competitive.
4. **Feature Engineering Matters:** Models using specialized features, like excitation around glottal closure instants (Sudarsana et al.), also perform well (86.96%), emphasizing the importance of carefully crafted input features.
5. **Lower Performance in GAN-Based and Unsupervised Approaches:** Methods like SSGAN and VSSSGAN (Zhao et al.) and semi-supervised ladder networks (Parthasarathy et al.) show lower accuracy (61.25% and 65.67%, respectively), indicating the challenges in training generative or unsupervised models effectively for SER without large annotated datasets.
6. **Incorporation of Metadata:** Models considering speaker attributes (e.g., Wei Sun et al. with gender info) demonstrate high performance (95.65%), suggesting the benefit of contextual metadata.

Overall, this comparison shows that the choice of model architecture, feature extraction techniques, and training strategies significantly influences SER performance. Models that combine deep learning with domain-specific features and metadata tend to achieve the best results.

5. CHALLENGES AND GAPS

Even though Speech Emotion Recognition has made huge strides, there are still some significant hurdles to overcome:

- **Emotions are Personal:** The way one person shows they're happy might be different from how someone else expresses the same feeling. Things like culture, where you're from, and just individual quirks can all play a role. The challenge is to build systems that can accurately understand and categorize a wide range of emotions, taking all these differences into account.
- **Everyone Sounds Different:** Think about it – everyone has a unique voice, with different pitches, accents, and ways of speaking. These differences can affect how emotions come across in speech, making it tricky to build models that work well for everyone.
- **Privacy Matters:** As SER technology gets more advanced, we need to be careful about privacy and how this technology is used ethically. We need better guidelines and rules to make sure it's deployed responsibly and ethically.
- **Real-World Noise is a Problem:** Building SER systems that can still work well even when there's background noise or other environmental distractions is a tough nut to crack. Getting reliable performance in everyday, noisy situations is a big goal.

- **It's Not Just About the Voice:** Sometimes, to understand an emotion, you need more than just the audio. Combining the words being said or even visual cues like facial expressions could make things more accurate, but it also makes the technology more complex because you're dealing with different types of data at once.
- **We Need to Talk to the Emotion Experts:** To get to the bottom of how emotions are expressed in speech, it's crucial for people who understand speech processing to work with people who study the psychology of emotions. This teamwork can lead to a much deeper understanding of the emotional signals in our voices.

The way we design and build SER algorithms is key to tackling these challenges. Researchers are constantly trying to create new and improved algorithms that can handle the variety in emotional expression, deal with noise, and accurately recognize emotions across different speakers and languages. These algorithms also aim to be adaptable to new situations through robust generalization and efficient transfer learning.

6. CONCLUSION

This paper has primarily focused on how advanced technologies can be used to detect emotions in spoken language. After looking at several research papers, we can see some clear trends in the artificial intelligence and machine learning algorithms being used. In the world of speech emotion detection, researchers have explored various methods, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Deep Convolutional Neural Networks (DCNN), Vector Space Models (VSM), and combinations of these. Notably, when it comes to understanding emotions in conversations and extracting the key features, the Convolutional Neural Network (CNN) algorithm and its more advanced versions, like DCNN have been shown to be more accurate than other techniques. These classifiers are good at picking out distinct human emotions like happiness, anger, neutrality, and more by analyzing the unique characteristics of our voices and speech.

The success of these approaches has been demonstrated across a wide range of datasets containing diverse speech samples. Because of this, they have a lot of potential in many real-world applications, from improving education and healthcare to making business process outsourcing (BPO) more effective and even helping with crime detection.

Recognizing how versatile and impactful Speech Emotion Recognition (SER) systems are, there's a growing interest in developing conversational AI models that can be tailored to the specific needs and preferences of individual users. These models aim to create more personalized and adaptive interactions between people and machines, leading to better user experiences in various areas. By using the power of SER and related technologies, these models can enable more empathetic and responsive human-machine interactions, ultimately improving communication and user satisfaction. This ongoing research and development in SER and conversational AI represent a really exciting area with significant implications for the future of technology and how we interact with computers.

REFERENCES

[1] R. R. Sehgal, S. Agarwal, and G. Raj, "Intelligent Voice Reaction involving Opinion Examination in Programmed Discourse Acknowledgment Frameworks," in *2018 Global Gathering on Advances in Processing and Correspondence Designing (ICACCE)*, 2018, pp. 213-218, doi: 10.1109/ICACCE.2018.8441741.

[2] K. Huang, C. Wu, Q. Hong, M. Su, and Y. Zeng, "Discourse Feeling Acknowledgment utilizing Convolutional Brain Organization with Sound Word-based Installing," in *2018 eleventh Worldwide Conference on Chinese Communicated in Language Handling (ISCSLP)*, 2018, pp. 265-269, doi: 10.1109/ISCSLP.2018.8706610.

[3] J. Cornejo and H. Pedrini, "Bimodal Feeling Acknowledgment In light of Sound and Facial Parts Utilizing Profound Convolutional Brain Organizations," in *2019 eighteenth IEEE Global Meeting On AI And Applications (ICMLA)*, 2019, pp. 111-117, doi: 10.1109/ICMLA.2019.00026.

[4] G. Deshmukh, A. Gaonkar, G. Golwalkar, and S. Kulkarni, "Discourse based Feeling Acknowledgment utilizing AI," in *2019 third Global Meeting on Processing Systems and Correspondence (ICCMC)*.

[5] Y. Dong, X. Yang, X. Zhao, and J. Li, "Bidirectional Convolutional Intermittent Inadequate Organization (BCRSN): An Effective Model for Music Feeling Acknowledgment," in *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3150-3163, Dec. 2019, doi: 10.1109/TMM.2019.2918739.

[6] Z. Zhao *et al.*, "Investigating Profound Range Portrayals through Consideration Based Repetitive and Convolutional Brain Organizations for Discourse Feeling Acknowledgment," in *IEEE Access*, vol. 7, pp. 97515-97525, 2019, doi: 10.1109/ACCESS.2019.2928625.

[7] A. Shahin, A. B. Nassif, and S. Hamsa, "Feeling Acknowledgment Utilizing Crossover Gaussian Combination Model and Profound Brain Organization," in *IEEE Access*, vol. 7, pp. 26777-26787, 2019, doi: 10.1109/ACCESS.2019.2901352.

[8] Y. Dong, X. Yang, X. Zhao, and J. Li, "Bidirectional Convolutional Intermittent Inadequate Organization (BCRSN): An Effective Model for Music Feeling Acknowledgment," in *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3150-3163, Dec. 2019, doi: 10.1109/TMM.2019.2918739.
[9] Apoorv Singh, Kshitij Kumar Srivastava, and Harini Murugan, "Discourse Feeling Acknowledgment Utilizing Convolutional Brain Organization (CNN)," *Global Journal of Psychosocial Rehabilitation*, vol. 24, no. 08, 2020.
[10] H. Cheng and X. Tang, "Discourse Feeling Acknowledgment in light of Intuitive Convolutional Brain Organization," in *2020 IEEE third Worldwide Gathering on Data Correspondence and Sign Handling (ICICSP)*, 2020, pp. 163-167, doi: 10.1109/ICICSP50920.2020.9232071.
[11] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, and N. Ismail, "Discourse Feeling Acknowledgment utilizing Convolution Brain Organizations and Profound Step Convolutional Brain Organizations," in *2020 sixth Worldwide Meeting on Remote and Telematics (ICWT)*, 2020, pp. 1-6, doi: 10.1109/ICWT50448.2020.9243622.
[12] "2021 File IEEE/ACM Exchanges on Sound, Discourse, and Language Handling Vol. 29," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3718-3760, 2021, doi: 10.1109/TASLP.2022.3147096.
[13] Zhang and L. Xue, "Autoencoder With Feeling Implanting for Discourse Feeling Acknowledgment," in *IEEE Access*, vol. 9, pp. 51231-51241, 2021, doi: 10.1109/ACCESS.2021.3069818.
[14] J. Oliveira and I. Praça, "On the Utilization of Pre-Prepared Discourse Acknowledgment Profound Layers to Recognize Feelings," in *IEEE Access*, vol. 9, pp. 9699-9705, 2021, doi: 10.1109/ACCESS.2021.3051083.
[15] J. Sidorova, S. Karlsson, O. Rosander, M. L. Berthier, and I. Moreno-Torres, "Towards Issue Free Programmed Evaluation of Profound Skill in Neurological Patients with a Traditional Feeling Acknowledgment Framework: Application in Unfamiliar Articulation Disorder," in *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 962-973, Oct.-Dec. 2021, doi: 10.1109/TAFFC.2019.2908365.
[16] S. Hamsa, I. Shahin, Y. Iraqi, and N. Werghi, "Feeling Acknowledgment From Discourse Utilizing Wavelet Parcel Change Cochlear Channel Bank and Irregular Woodland Classifier," in *IEEE Access*, vol. 8, pp. 96994-97006, 2020, doi: 10.1109/ACCESS.2020.2991811.
[17] S. Parthasarathy and C. Busso, "Semi-Directed Discourse Feeling Acknowledgment With Stepping stool Organizations," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697-2709, 2020, doi: 10.1109/TASLP.2020.3023632.
[18] S. R. Kadiri and P. Alku, "Excitation Elements of Discourse for Speaker-Explicit Feeling Identification," in *IEEE Access*, vol. 8, pp. 60382-60391, 2020, doi: 10.1109/ACCESS.2020.2982954.
[19] S. Latif *et al.*, "Perform various tasks Semi-Managed Ill-disposed Autoencoding for Discourse Feeling Acknowledgment," in *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 992-1004, April-June 2022, doi: 10.1109/TAFFC.2020.2983669.
[20] Li, "Automated Feeling Acknowledgment Involving Two-Level Elements Combination in Sound Signs of Discourse," in *IEEE Sensors Journal*, vol. 22, no. 18, pp. 17447-17454, Sept. 15, 2022, doi: 10.1109/JSEN.2021.3065012.
[21] Z. Peng *et al.*, "Discourse Feeling Acknowledgment Utilizing 3D Convolutions and Consideration Based Sliding Repetitive Organizations With Hear-able Front-Finishes," in *IEEE Access*, vol. 8, pp. 16560-16572, 2020, doi: 10.1109/ACCESS.2020.2967791.
[22] H. Zhao, Y. Xiao, and Z. Zhang, "Powerful Semi-directed Generative Antagonistic Organizations for Discourse Feeling Acknowledgment through Circulation Perfection," in *IEEE Access*, vol. 8, pp. 106889-106900, 2020, doi: 10.1109/ACCESS.2020.3000751.
[23] J. B. Awotunde, R. O. Ogundokun, F. E. Ayo, and O. E. Matiluko, "Discourse Isolation in Foundation Commotion In view of Profound Learning," in *IEEE Access*, vol. 8, pp. 169568-169575, 2020, doi: 10.1109/ACCESS.2020.3024077.
[24] T. -W. Sun, "Start to finish Discourse Feeling Acknowledgment With Orientation Data," in *IEEE Access*, vol. 8, pp. 152423-152438, 2020, doi: 10.1109/ACCESS.2020.3017462.
[25] G. Du, S. Long, and H. Yuan, "Non-Contact Feeling Acknowledgment Consolidating Pulse and Look for Intuitive Gaming Conditions," in *IEEE Access*, vol. 8, pp. 11896-11906, 2020, doi: 10.1109/ACCESS.2020.2964794.
[26] Q. Kong *et al.*, "PANNs: Enormous Scope Pretrained Sound Brain Organizations for Sound Example Acknowledgment," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880-2894, 2020, doi: 10.1109/TASLP.2020.3030497.
[27] Z. -T. Liu *et al.*, "Discourse Character Acknowledgment In view of Explanation Arrangement Utilizing Log-Probability Distance and Extraction of Fundamental Sound Highlights," in *IEEE Transactions on Multimedia*, vol. 23, pp. 3414-3426, 2021, doi: 10.1109/TMM.2020.3025108.

[28] M. Firdaus, H. Chauhan, A. Ekbal, and P. Bhattacharyya, "Emotional Sen: Producing Feeling and Feeling Controlled Reactions in a Multimodal Discourse Framework," in *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1555-1566, July-Sept. 2022, doi: 10.1109/TAFFC.2020.3015491.

[29] M. A. Rashidan *et al.*, "Innovation Helped Feeling Acknowledgment for Mental imbalance Range Problem (ASD) Kids: An Orderly Writing Survey," in *IEEE Access*, vol. 9, pp. 33638-33653, 2021, doi: 10.1109/ACCESS.2021.3060753.

[30] Y. Huang, J. Xiao, K. Tian, A. Wu, and G. Zhang, "Exploration on Vigor of Feeling Acknowledgment Under Ecological Commotion Conditions," in *IEEE Access*, vol. 7, pp. 142009-142021, 2019, doi: 10.1109/ACCESS.2019.2944386.