Orclever Proceedings of Research and Development
https://doi.org/10.56038/oprd.v3i1.382
OR CLEVER Science & Research Group

Conference Article

Enhancing IVR Systems in Mobile Banking with Emotion Analysis for Adaptive Dialogue Flows and Seamless Transition to Human Assistance

Alper Ozpinar[1*], Ersin Alpan[2], Taner Çelik[3]

[1] Istanbul Commerce University, https://orcid.org/0000-0003-1250-5949, alper.ozpinar@ticaret.edu.tr,
[2] Softtech, https://orcid.org/0009-0007-9798-1253, ersin.alpan@softtech.com.tr,
[3] Softtech, https://orcid.org/0009-0005-1978-0822, taner.celik@softtech.com.tr,
* Correspondence: alper.ozpinar@ticaret.edu.tr; +90 552 336 46 24

(First received October 23, 2023 and in final form December 26, 2023)

3rd International Conference on Design, Research and Development
(RDCONF 2023)
December 13 - 15, 2023

Reference: Özpinar, A., Alpan, E., Çelik, T. Enhancing IVR Systems in Mobile Banking with Emotion Analysis for Adaptive Dialogue Flows and Seamless Transition to Human Assistance**.** Orclever Proceedings of Research and Development,3(1), 592-605.

Abstract

This study introduces an advanced approach to improving Interactive Voice Response (IVR) systems for mobile banking by integrating emotion analysis with a fusion of specialized datasets. Utilizing the RAVDESS, CREMA-D, TESS, and SAVEE datasets, this research exploits a diverse array of emotional speech and song samples to analyze customer sentiment in call center interactions. These datasets provide a multi-modal emotional context that significantly enriches the IVR experience.

The cornerstone of our methodology is the implementation of Mel-Frequency Cepstral Coefficients (MFCC) Extraction. The MFCCs, extracted from audio inputs, form a 2D array where time and cepstral coefficients create a structure that closely resembles an image. This format is particularly suitable for Convolutional Neural Networks (CNNs), which excel in interpreting such 'image-like' data for emotion recognition, hence enhancing the system's responsiveness to emotional cues.

Proposed system's architecture is adeptly designed to modify dialogue flows dynamically, informed by the emotional tone of customer interactions. This innovation not only improves

https://doi.org/10.56038/oprd.v3i1.382

customer engagement but also ensures a seamless handover to human operators when the situation calls for a personal touch, optimizing the balance between automated efficiency and human empathy.

The results of this research demonstrate the potential of emotion-aware IVR systems to anticipate and meet customer needs more effectively, paving the way for a new standard in user-centric banking services.

Keywords: Emotion Analysis, CNN, NLP, Adaptive Dialogs, IVR Systems

1. Introduction

The financial technology (Fintech) sector, particularly within the banking industry, has experienced a significant transformation over the last decade [1]. At the forefront of this evolution is the integration of Interactive Voice Response (IVR) systems into call centers, marking a pivotal shift towards self-service banking [2]. These systems are instrumental in streamlining operations, offering 24/7 customer service, and playing a critical role in fraud detection and warning mechanisms [3], [4]. The adoption of IVR has enabled banks to provide a more robust and secure service, empowering customers to perform a variety of banking functions independently [5].

With the increasing demand for enhanced customer service and efficient call handling, speech emotion detection has emerged as a critical component within the banking sector's IVR systems. The implementation of this technology signifies a novel approach to understanding and responding to the emotional states of customers during interactions. By recognizing subtle nuances in speech, banks can tailor their services to the specific needs of each customer, thereby fostering a more personalized banking experience.

The extraction of Mel-Frequency Cepstral Coefficients (MFCCs) from audio data is a sophisticated process that forms the bedrock of speech emotion detection [6], [7]. The intricate procedure involves segmenting the audio signal into short frames, applying a window function, and subsequently conducting a Fourier transform to transition each frame from the time domain to the frequency domain. Following this, a Mel filter bank is applied to the power spectra, the logarithm is taken, and a Discrete Cosine Transform (DCT) is performed to yield the MFCCs [8], [9].

Once extracted, MFCCs manifest as a 2D array, embodying a temporal and coefficient-based structure akin to an image. This format is exceptionally conducive to the

https://doi.org/10.56038/oprd.v3i1.382

application of Convolutional Neural Networks (CNNs), which excel in image recognition. CNNs can adeptly navigate the 'image-like' data of MFCCs, enabling the accurate classification or regression of emotional states [10], [11]. The design of a CNN architecture necessitates a thoughtful assembly of convolutional, pooling, and fully connected layers, with an emphasis on convolutional layers along the time dimension for sequential data such as MFCCs [12], [13].

The intersection of Deep Learning and machine learning techniques, including CNNs and Long Short-Term Memory (LSTM) networks, presents a robust framework for the analysis of complex patterns within speech data. These methodologies are not only critical in emotion detection but also serve as the backbone for a myriad of other applications within the financial sector [14], [15].

The aim of integrating these technologies into banking call centers is to develop hybrid AI-IVR systems that revolutionize customer service. These hybrid systems marry Artificial Intelligence (AI) with conventional IVR, enabling the analysis of customer speech, understanding of requests, and provision of automated responses. A sophisticated mechanism governs the transition from AI to human assistance: upon detecting complex queries or emotional distress, the system seamlessly escalates the call to human agents [14], [16].

The benefits of deploying such advanced AI-IVR systems in banking call centers are manifold. Notably, there is a marked reduction in wait times, an increase in the accuracy of query handling, and an overall enhancement in customer satisfaction. This approach promises a new era of customer service in banking—one that is more responsive, intuitive, and aligned with the emotional and practical needs of customers.

In summary, the integration of AI with IVR systems in the banking industry is not merely an incremental upgrade but a significant leap towards a future where technology and human expertise converge to deliver unprecedented levels of service excellence. The following chapters will delve deeper into the components of this integration, the methodologies employed, and the profound impact it is poised to have on the financial industry.

2. Materials and Methods

https://doi.org/10.56038/oprd.v3i1.382

2.1. Data Privacy and Compliance with GDPR

In the pursuit of technological advancements in the field of Interactive Voice Response (IVR) systems, particularly within the banking sector, it is imperative to address the concerns surrounding data privacy and compliance with legal frameworks [17]. The General Data Protection Regulation (GDPR) presents a comprehensive set of guidelines that enshrine the privacy and protection of personal data within the European Union [18]. Adherence to GDPR is non-negotiable; it shapes the landscape in which customer data is utilized for any form of research and development, including the training of machine learning models [19].

The GDPR enforces stringent conditions on the usage of personal data, ensuring that the privacy of individuals is not compromised. Within the context of IVR systems, original customer records are laden with personal identifiers and sensitive information that fall under the protective umbrella of GDPR. Consequently, using such records for research or for the training of AI systems without explicit consent and proper anonymization would constitute a breach of these regulations. This legal and ethical landscape necessitates a cautious approach to data handling, one that respects the boundaries of privacy while still enabling innovation.

The diversity and quality of these datasets facilitate a more comprehensive and robust analysis than what would typically be possible with a singular, proprietary dataset. They enable researchers to train algorithms that are more generalizable and resilient to overfitting, resulting in models that are better equipped to handle real-world variability in customer interactions. Moreover, the usage of these datasets ensures that the privacy of individuals is maintained, as no direct or indirect identifiers are present that could link the data back to the original contributors.

2.2. Datasets

The constraints imposed by GDPR on the use of customer IVR records have led researchers to seek alternative avenues. Public datasets, such as RAVDESS, CREMA-D, TESS, and SAVEE, have emerged as invaluable assets in this regard. These datasets are meticulously curated to comply with GDPR, offering anonymized and consented data that can be freely used for academic and commercial research. The data contained within these repositories have been provided by individuals who have given their informed consent for the use of their voice recordings, with all personal identifiers removed to ensure anonymity.

https://doi.org/10.56038/oprd.v3i1.382

These public datasets not only align with GDPR compliance but also offer several advantages for research in emotion recognition and speech processing. They encompass a wide spectrum of emotional expressions, diverse accents, and various modalities of speech and song, providing a rich testing ground for the development of sophisticated algorithms. Their extensive use in the realms of machine learning and artificial intelligence has been pivotal in advancing the fields of human-computer interaction and affective computing. There are various studies about these datasets and implications with CNN and MFCC [20]–[24].

- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): This dataset contains audio and video recordings of actors performing emotional expressions in speech and song. The dataset includes 24 professional actors (12 male, 12 female), vocalizing two lexically-matched statements in a neutral North American accent. Emotions expressed in the dataset include calm, happy, sad, angry, fearful, surprise, and disgust, along with a neutral expression. Each expression is produced at two levels of emotional intensity (normal and strong), along with a neutral expression. The dataset is widely used for training models to recognize emotional cues in speech [25].
- CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset): The CREMA-D dataset consists of 7,442 clips of 91 actors (48 male, 43 female) speaking 12 sentences, each conveying one of six different emotions (anger, disgust, fear, happy, neutral, sad) and four levels of emotional intensity (low, medium, high, unspecified). It's a multimodal dataset, including audio, video, and physiological signals, making it useful for studies and applications that involve multi-channel data processing [26].
- TESS (Toronto Emotional Speech Set): TESS is a dataset of emotional speech, which was collected from two actresses with a total of 2,800 utterances, covering seven different emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). The actresses read a set of target sentences to portray each emotion, providing a rich source of data for emotional speech processing tasks. This dataset is particularly noted for its high-quality recordings and consistent emotion portrayals [27].
- SAVEE (Surrey Audio-Visual Expressed Emotion): The SAVEE database contains audio-visual recordings from 4 male actors, each expressing 7 different emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral). Each actor speaks 15 sentences per emotion, leading to a total of 480 utterances. This dataset is known

https://doi.org/10.56038/oprd.v3i1.382

for its inclusion of British English speakers, providing a different accent and set of expressions compared to other datasets [28].

2.3. CNN Architecture

The CNN Architecture implemented in the paper is composed of the following sub structures

- Convolutional Layer: This is the core building block of a CNN. The convolutional layer applies a number of filters to the input. Each filter scans the input image like a sliding window and produces a feature map. This process captures the local dependencies in the input data, such as edges, corners, or textures in images.
- Activation Function: Typically, after each convolution operation, an activation function is applied to introduce non-linear properties to the system. The most common activation function is the Rectified Linear Unit (ReLU).
- Pooling Layer: Pooling (also known as subsampling or downsampling) reduces the dimensionality of each feature map but retains the most important information. Max pooling, for example, takes the maximum value in each window of the feature map.
- Fully Connected Layer: After several convolutional and pooling layers, the high-level reasoning in the neural network is done via fully connected layers. Neurons in a fully connected layer have full connections to all activations in the previous layer.
- Output Layer: The final layer is a fully connected layer with an activation function such as the softmax function for classification tasks that outputs the probabilities of the classes.

2.4. Mel-Frequency Cepstral Coefficients (MFCC) Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) are a feature widely used in speech and audio processing, particularly in fields like speech recognition, speaker identification, and music genre classification. The extraction of MFCCs from an audio signal involves several steps, each designed to mimic the human auditory system's response and to capture the relevant characteristics of the speech. Here's a breakdown of the MFCC extraction process.

- Frame the Signal into Short Windows: The audio signal is divided into short frames (typically 20-40 ms long) with overlap (often 50%). This is done because the frequency content of the entire signal varies over time, so analyzing short

https://doi.org/10.56038/oprd.v3i1.382

segments provides a good approximation of the frequency contours of the signal over time.

- Apply the Fourier Transform to Each Frame: A Fast Fourier Transform (FFT) is applied to each frame to convert it from the time domain into the frequency domain. This step generates a spectrum for each frame.
- Map the Powers of the Spectrum onto the Mel Scale: The Mel scale is a perceptual scale that better represents human hearing than the linearly-spaced frequency bands. It scales the frequency in a way that approximates the human ear's response more closely than the linear scales. The frequencies are converted to the Mel scale using a set of triangular filters, each covering a specific range of frequencies.
- Take the Logarithm of the Mel Spectrum: Human perception of sound intensity also follows a logarithmic scale rather than a linear one. Therefore, the logarithm of the Mel spectrum is taken to capture this aspect of human auditory perception.
- Apply the Discrete Cosine Transform (DCT): The DCT is used to convert the log Mel spectrum into time coefficients. The DCT helps to de-correlate the feature set (which represents the rate of change in the different spectrum bands) and compresses the spectrum, emphasizing the more important coefficients. Typically, the first 12-13 coefficients are used as they contain the most significant information about the signal.
- Delta and Delta-Delta Coefficients (Optional): Sometimes, the first and second derivatives of the MFCCs (called delta and delta-delta coefficients) are computed and appended to the feature set. These derivatives provide information about the rate and acceleration of the spectral variation, adding more detail to the representation of the speech signal.

2.5. Methodology

Detailed steps of the methodology explained as follows;

- Extract MFCCs: Firstly, MFCCs extracted from audio data. This process involves taking the audio signal, framing it into short segments, applying a window function, and then using the Fourier transform to convert each frame from time to frequency domain. After this, apply the Mel filter bank to the power spectra, take the logarithm, and then the Discrete Cosine Transform (DCT) to get the MFCCs.
- Preprocess MFCCs: The MFCCs extracted from an audio file result in a 2D array where one dimension is time, and the other is the MFCC coefficients. This 2D array can be treated similarly to an image, which is an input that CNNs can handle well.

https://doi.org/10.56038/oprd.v3i1.382

- Design CNN Architecture: Design a CNN architecture suitable for classification or regression task. The architecture typically consists of convolutional layers, pooling layers, and fully connected layers towards the end.
- Normalize Input Data: Normalize the MFCCs for each audio sample to have zero mean and unit variance. This normalization helps with the convergence during the training of the CNN.
- Train the CNN: Input the normalized MFCCs into the CNN. Train the CNN using a labeled dataset that leads to the MFCCs as the input features and the corresponding labels as the targets.
- Evaluate the Model: After training, evaluation of the performance of the CNN on a test set to see how well it generalizes to new, unseen data.
- Fine-tuning: Depending on the initial results, fine-tuning of CNN architecture, hyperparameters, or even the preprocessing steps of the MFCCs.

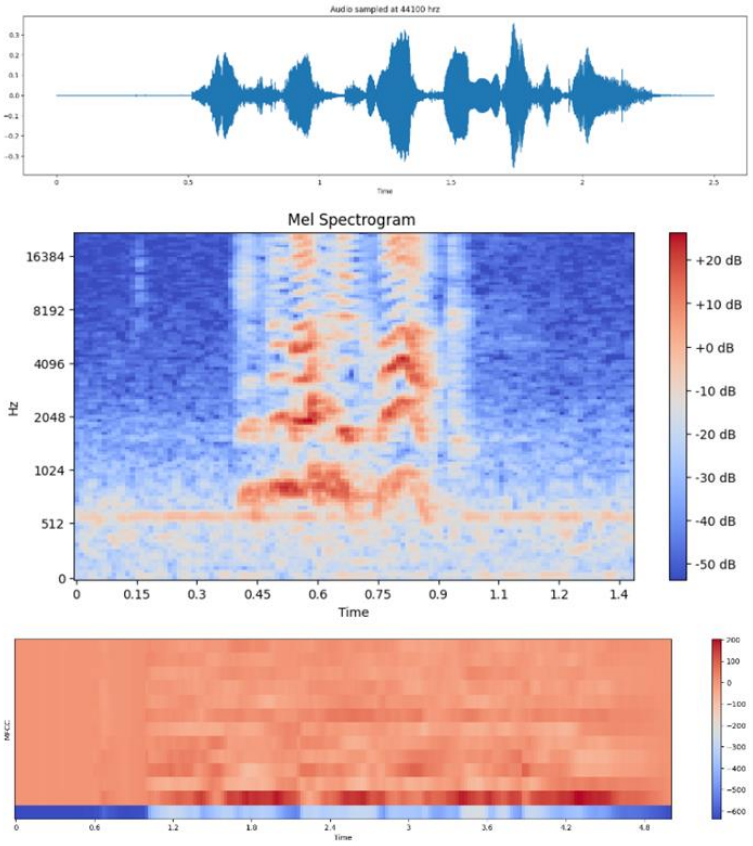

Figure 1: Sample MFCC Extraction for a Audio File

https://doi.org/10.56038/oprd.v3i1.382

Table 1: CNN Architecture

Layer Type	Filters	Kernel Size	Activation	Dropout Rate	Pool Size
Input	N/A	N/A	N/A	N/A	N/A
Conv1D	256	8	relu	N/A	N/A
Conv1D	256	8	relu	N/A	N/A
BatchNormalization	N/A	N/A	N/A	N/A	N/A
Dropout	N/A	N/A	N/A	0.25	N/A
MaxPooling1D	N/A	N/A	N/A	N/A	8
Conv1D	128	8	relu	N/A	N/A
Conv1D	128	8	relu	N/A	N/A
Conv1D	128	8	relu	N/A	N/A
Conv1D	128	8	relu	N/A	N/A
BatchNormalization	N/A	N/A	N/A	N/A	N/A
Dropout	N/A	N/A	N/A	0.25	N/A
MaxPooling1D	N/A	N/A	N/A	N/A	8
Conv1D	64	8	relu	N/A	N/A
Conv1D	64	8	relu	N/A	N/A
Flatten	N/A	N/A	N/A	N/A	N/A
Dense	N/A	N/A	softmax	N/A	N/A

Figure 2: Seamless Denim Product

https://doi.org/10.56038/oprd.v3i1.382

3. Result

The evaluation of the model's performance through the provided metrics indicates a commendable level of success in classifying emotions from speech. Notably, emotions such as 'angry' and 'surprise' are identified with a high degree of accuracy, which suggests that the model is particularly adept at detecting the distinct features associated with these emotional states.

The overall accuracy, as reflected by the precision, recall, and F1-score metrics, demonstrates that the model is generally effective in its classifications. The training and testing loss curves suggest that the model is learning well and is able to generalize from the training data to unseen data, which is a positive indicator of its robustness.

Furthermore, the success of the model across different emotions, while varied, shows that it can capture a broad range of emotional expressions, a crucial aspect of effective IVR systems. There is, however, an observed variation in performance between genders, indicating potential areas for model refinement.

In summary, the model exhibits a solid foundation for accurately interpreting emotional content in speech, with certain emotions being recognized with greater precision. This establishes a promising basis for future enhancements and applications in real-world scenarios.

4. Discussion and Conclusion

This paper has explored the innovative integration of emotion analysis within Interactive Voice Response (IVR) systems in mobile banking, leveraging public datasets to ensure GDPR compliance. The discussion has spanned the implications of such integrations for customer service, the technical methodologies involved in processing and interpreting emotional speech, and the broader ramifications for privacy and data protection.

The novel approach of employing Mel-Frequency Cepstral Coefficients (MFCC) extraction and Convolutional Neural Network (CNN) architectures presented in this paper underscores the potential for AI to revolutionize customer interaction within the banking sector. By analyzing the emotional content of customer speech through these advanced techniques, banks can anticipate customer needs more effectively and tailor

https://doi.org/10.56038/oprd.v3i1.382

their services accordingly. The research has highlighted the significant benefits of this approach, such as reduced wait times and enhanced customer satisfaction, which are paramount in today's competitive market.

Moreover, the study has illuminated the ethical landscape of using customer data, stressing the importance of GDPR compliance. The adherence to GDPR principles in the use of public datasets has been emphasized not only as a legal obligation but as a commitment to ethical research practices. By utilizing datasets such as RAVDESS, CREMA-D, TESS, and SAVEE, the research maintains the highest standards of data privacy while contributing valuable insights into the fields of emotion recognition, speech processing, and human-computer interaction.

The implications of this research are far-reaching, extending beyond technological advancements to propose a new paradigm in customer-bank interactions. The transition to hybrid AI-IVR systems, as discussed, represents a fusion of technology and human expertise, offering a glimpse into the future of banking services that are more responsive and attuned to the emotional states of customers.

In conclusion, the integration of emotion analysis into IVR systems presents a compelling case for the harmonious coexistence of technology and personalized customer service. As the banking industry continues to evolve, the insights derived from this research will likely inform future developments in the sector. The discussion posits that the continual refinement of these systems, with an emphasis on ethical data usage and customer-centric innovation, will be critical to achieving excellence in banking services. The path forward, as this research suggests, is one that embraces the transformative power of AI while upholding the sanctity of individual privacy—a balance that will define the trajectory of customer service in the years to come.

5. Acknowledge

The research presented in this paper is the culmination of extensive investigative efforts undertaken within the realm of two distinct research projects at Softtech R&D Center. The authors would like to express their profound gratitude to the Softtech R&D Center for providing the fertile ground for intellectual inquiry and for the steadfast support that has been pivotal to the progression and fruition of this scholarly work.

References

https://doi.org/10.56038/oprd.v3i1.382

[1] R. Alt, R. Beck, and M. T. Smits, "FinTech and the transformation of the financial industry," *Electronic markets*, vol. 28. Springer, pp. 235–243, 2018.

[2] P. Manatsa, "An analysis of the impact of implementing a new interactive voice response system (IVR) on client experience in the Canadian Banking Industry," 2019.

[3] R. A. Feinberg, L. Hokama, R. Kadam, and I. Kim, "Operational determinants of caller satisfaction in the banking/financial services call center," *International Journal of Bank Marketing*, vol. 20, no. 4, pp. 174–180, 2002.

[4] S. M. Yacoub, S. J. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems.," in *Interspeech*, 2003.

[5] L. E. Rocha, D. M. R. Glina, M. de Fatimá Marinho, and D. Nakasato, "Risk factors for musculoskeletal symptoms among call center operators of a bank in Sao Paulo, Brazil," *Ind Health*, vol. 43, no. 4, pp. 637–646, 2005.

[6] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[7] B. Logan, "Mel frequency cepstral coefficients for music modeling.," in *Ismir*, Plymouth, MA, 2000, p. 11.

[8] S. A. Khayam, "The discrete cosine transform (DCT): theory and application," *Michigan State University*, vol. 114, no. 1, p. 31, 2003.

[9] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *2010 4th International Conference on Signal Processing and Communication Systems*, IEEE, 2010, pp. 1–5.

[10] F. Jiang, H. Li, Z. Zhang, and X. Zhang, "An event recognition method for fiber distributed acoustic sensing systems based on the combination of MFCC and CNN," in *2017 International Conference on Optical Instruments and Technology: Advanced Optical Sensors and Applications*, SPIE, 2018, pp. 15–21.

[11] K. Mridha, S. Sarkar, and D. Kumar, "Respiratory disease classification by CNN using MFCC," in *2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, 2021, pp. 517–523.

[12] S. Jin, X. Wang, L. Du, and D. He, "Evaluation and modeling of automotive transmission whine noise quality based on MFCC and CNN," *Applied Acoustics*, vol. 172, p. 107562, 2021.

[13] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," *IEEE transactions on information forensics and security*, vol. 15, pp. 1616–1629, 2019.

[14] G. Petmezas *et al.*, "Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function," *Sensors*, vol. 22, no. 3, p. 1232, 2022.

[15] H.-A. Rashid, A. N. Mazumder, U. P. K. Niyogi, and T. Mohsenin, "CoughNet: A flexible low power CNN-LSTM processor for cough sound detection," in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, IEEE, 2021, pp. 1–4.

[16] E. Åberg and Y. Khati, "Artificial Intelligence in Customer Service: A Study on Customers' Perceptions regarding IVR Services in the Banking Industry." 2018.

[17] G. D. P. Regulation, "General data protection regulation (GDPR)," *Intersoft Consulting, Accessed in October*, vol. 24, no. 1, 2018.

[18] E. Arfelt, D. Basin, and S. Debois, "Monitoring the GDPR," in *Computer Security–ESORICS 2019: 24th European Symposium on Research in Computer Security, Luxembourg, September 23–27, 2019, Proceedings, Part I 24*, Springer, 2019, pp. 681–699.

[19] C. Tankard, "What the GDPR means for businesses," *Network Security*, vol. 2016, no. 6, pp. 5–8, 2016.

[20] M. R. Ahmed, S. Islam, A. K. M. M. Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," *Expert Syst Appl*, vol. 218, p. 119633, 2023.

[21] S. Ullah, Q. A. Sahib, S. Ullahh, I. U. Haq, and I. Ullah, "Speech Emotion Recognition Using Deep Neural Networks," in *2022 International Conference on IT and Industrial Technologies (ICIT)*, IEEE, 2022, pp. 1–6.

[22] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, "Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets," *Electronics (Basel)*, vol. 11, no. 22, p. 3831, 2022.

[23] H. Dolka, A. X. VM, and S. Juliet, "Speech emotion recognition using ANN on MFCC features," in *2021 3rd international conference on signal processing and communication (ICPSC)*, IEEE, 2021, pp. 431–435.

[24] N. Chitre, N. Bhorade, P. Topale, J. Ramteke, and C. R. Gajbhiye, "Speech Emotion Recognition to assist Autistic Children," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, 2022, pp. 983–990.

https://doi.org/10.56038/oprd.v3i1.382

[25] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS One*, vol. 13, no. 5, p. e0196391, 2018.

[26] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans Affect Comput*, vol. 5, no. 4, pp. 377–390, 2014.

[27] K. Dupuis and M. K. Pichora-Fuller, "Toronto emotional speech set (tess)-younger talker_happy," 2010.

[28] P. Jackson and Sju. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.