International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 8, Issue 1, pp 94-104, January-2024
https://doi.org/10.47001/IRJIET/2024.801013

AI Based Speech Analysis Framework

[1]W.A.D.Perera, [2]Mr. Jeewaka Perera, [3]Mr. Tharaniyawarma.K

[1,2,3]Faculty of Computing, Sri Lanka Institute of Information Technology, Sri Lanka
Authors E-mail: [1]it19237596@my.sliit.lk, [2]Jeewaka.p@sliit.lk, [3]tharaniyawarma.k@sliit.lk

Abstract **- The pioneering AI-Based Speech Analysis Framework presented in this research paper was painstakingly created to help people overcome linguistic obstacles, notably in the context of English language communication. Through a thorough speech analysis, the framework's multimodal approach enables real-time evaluation of emotional state, fluency, stress levels, and even identification recognition. This framework delivers a sophisticated and perceptive interpretation of spoken language by utilizing cutting-edge artificial intelligence approaches, hence promoting an enhanced and successful communication experience. The study is focused on four significant sub-objectives, each of which advances the main objective of encouraging increased self-awareness and communication: First, by detecting subtly emotional indicators embedded in the voice, the framework transforms emotional assessment. The AI algorithms identify emotional patterns, such as enthusiasm, trepidation, or tranquility. This in-the-moment emotional analysis creates opportunities for tailored communication techniques and a greater understanding of the speaker's feelings. The framework also introduces a novel way for assessing fluency levels using voice analysis. It analyzes various facets of speech, such as pace, intonation, and lexical decisions, giving language learners immediate feedback on their level of linguistic proficiency. This makes it easier to make focused improvements and to move more easily toward effective communication. The framework also discusses the complex relationship between stress and good communication. It measures stress levels through vocal pattern analysis, offering light on instances of heightened tension or anxiety when speaking. Such knowledge enables people to overcome stress-related hurdles and enhance communication. The framework's capacity to accurately identify people based on distinctive voice traits lies at the heart of its innovation. Language limitations are no obstacle to this identity recognition technology, which provides an effective and secure method of identification in a variety of settings. Voice-based identification detection accelerates procedures and promotes inclusion in a variety of settings, including work settings and public services. The development of an AI-Based Speech Analysis Framework that reveals fresh angles in language evaluation and communication improvement is the culmination of this research. It not only encourages self-improvement but also highlights the revolutionary potential of AI in redefining language landscapes and promoting true connections by merging emotional, fluency, stress analysis, and identity identification through voice.**

Keywords: AI-Based Speech Analysis, Emotional Assessment, Fluency Evaluation, Stress Detection, Identity Recognition, Language Proficiency.

I. Introduction

We now perceive, evaluate, and enhance spoken language exchanges differently thanks to the convergence of artificial intelligence (AI) and speech analysis. The development of an AI-Based Speech Analyst Framework, a seamless fusion of machine learning capabilities and the adaptability of soft computing, is the result of this convergence and is an ambitious undertaking. This framework captures the core of contemporary technical developments and their ability to decipher the complexity of spoken language by probing the multiple layers of human communication.[2] The limits of problem-solving have been redefined by the collaboration of machine learning and soft computing. Their application to voice analysis marks the beginning of a revolutionary period in which complex audio data is not just processed but also thoroughly understood. [6] Convolutional neural networks and support vector machines are only two examples of machine learning techniques that serve as the virtual lens through which we may see the underlying patterns that underlie spoken language. Soft computing, with its emphasis on approximation and reasoning, enhances the accuracy of machine learning by providing flexibility in managing erroneous or ambiguous data. Effective communication transcends linguistic boundaries and becomes a cornerstone of individual, professional, and cultural development in a linked world. [7] The ability to communicate verbally, especially in widely spoken languages like English, is essential for establishing cross-cultural interactions and empowering people to successfully negotiate the globalized world. The road to linguistic mastery is, however, frequently paved with difficulties—difficulties that concern not only vocabulary and grammatical accuracy, but also emotional expression, fluency, stress modulation, and the distinctive subtleties of individual voices. [4] The development of the AI-

© 2024-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 94-104, January-2024

https://doi.org/10.47001/IRJIET/2024.801013

Based Speech Analyst Framework is a creative solution to these complex problems. It makes use of AI and speech analysis to build a comprehensive framework for more precise and effective communication. While still somewhat useful, traditional language assessment approaches fall short in addressing the complex aspects of spoken language. The design and development of an AI-Based Speech Analyst Framework that surpasses the constraints of conventional techniques is the central research topic of this project. Making a framework that not only thoroughly examines the complex components of spoken language but also incorporates the power of machine learning and soft computing to produce nuanced insights, so improving the quality and effectiveness of communication, is the key to the task. [1]The following well-defined goals serve as the foundation for this research project and together they help to construct the AI-Based Speech Analyst Framework: Engineering is the process of developing, designing, constructing, researching, and improving structures, systems, machines, materials, solutions, and organizations using math, economics, science, and practical knowledge. Engineering is a broad discipline with several subfields, each focusing on a particular area of applied science, technology, etc.[3]

Acoustic engineering, often known as acoustic engineering, is a subfield of engineering that deals with the physics of sound and vibration as well as acoustic applications. Noise control is one of the main goals of acoustical engineering. Speech and hearing investigations must be a part of acoustic science.[2]

Emotion Analysis and Recognition: One of the main goals of the framework is to decipher the emotional web woven into voice patterns. The framework intends to precisely identify and analyze emotions like happiness, sadness, anger, surprise, and neutrality by utilizing sophisticated machine learning techniques, particularly the Mel-Frequency Cepstral Coefficients (MFCC) method. This goal's importance is highlighted by ongoing efforts to improve emotion recognition's precision.[10]

Fluency assessment and improvement: Fluency goes beyond language accuracy to include rhythm, intonation, and naturalness. The framework explores these complex facets of spoken language in an effort to go beyond simple fluency tests. It strives to close the gap between theoretical proficiency and practical application by offering insights on pronunciation correctness and general fluency. Accurate fluency assessment depends on the framework's durability in noisy situations, which is a crucial aspect of this goal.[3]

Stress Level Detection and Management: The framework aims to identify stress levels that are ingrained in voice patterns in recognition of the crucial function that stress plays in communication. The framework tries to identify stress indicators through methodically examining voice signals, providing users with useful insights about their stress dynamics during communication. This goal fits well with the overarching goal of equipping people with efficient stress management techniques.[6]

Voice-Based Identity Recognition: The framework makes contributions to the field of voice-based identity recognition. The framework records and saves distinctive voice traits since voices are as individual as fingerprints. This resource makes it easier to accurately recognize voices as people, which has ramifications for security and authentication. Improved accuracy, attained by sophisticated feature extraction methods like the Fast Fourier Transform (FFT), continues to be a crucial goal in this area.[7]

This research introduces a thorough framework that caters to the complex facets of spoken communication, making a significant contribution to the fields of AI and speech analysis. The framework has the capacity to decipher emotions, assess fluency, identify voices, and detect stress thanks to the integration of machine learning and soft computing. [2] This contribution has broad-reaching effects that touch on identity verification, communication training, and language learning. In essence, the AI-Based Speech Analyst Framework sets out on a trailblazing adventure that uses AI to transform the field of spoken language analysis. This project aims to provide people the skills they need to interact meaningfully and successfully in a connected world by managing the intricacies of emotions, fluency, stress dynamics, and voice recognition. [5]Naturally, we are pleased with the speech would prefer to communicate with computers through without deploying dated interfaces like keyboards, such as the speech medium or the pointing apparatuses. This is readily accomplished by Consider an Automatic Speech Recognition (ASR) method, which symbolizes the ability to transform a speech signal into a string of words with the use of a computer program's algorithm. [1] It is beautifully supplied with the abilities necessary for it to become a prominent interface involving computers and people.

II. Background and Literature Review

Effective communication crosses traditional boundaries and is crucial in social, professional, and personal situations in a time of increased globalization and digital interconnection. The ability to speak fluently and authentically in widely spoken languages like English has evolved into a vital skill as individuals and communities navigate linguistic variety. The difficulties on this journey, however, go beyond issues with vocabulary and grammar.[6]The areas of emotional

© 2024-2017 IRJIET All Rights Reserved

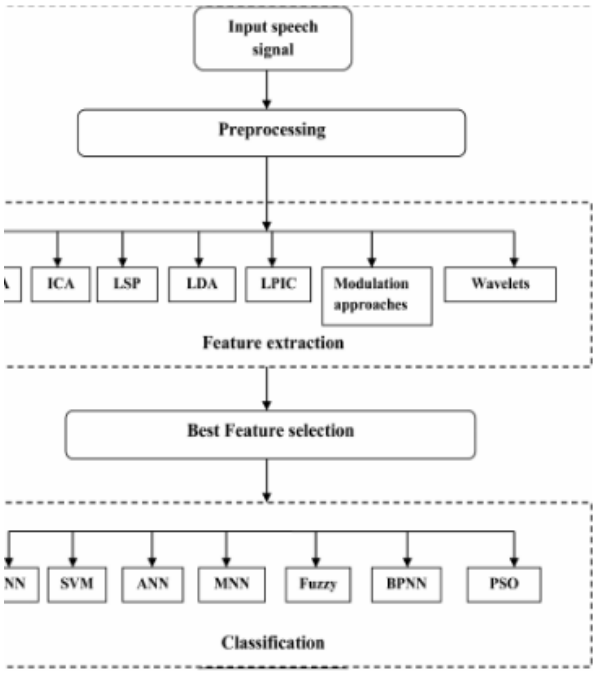
International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 94-104, January-2024

https://doi.org/10.47001/IRJIET/2024.801013

expression, fluency, stress management, and voice-based identification recognition are all included in these difficulties. The advent of an AI-Based Speech Analyst Framework ushers in a paradigm shift in communication analysis and augmentation in response to these complex problems. This framework offers a comprehensive approach to resolving the complexity inherent in spoken language by using modern speech analysis techniques and the capabilities of artificial intelligence (AI).[8]



These voice signals are now widely used in biometric recognition techniques as well as for computer interfaces. The development of appropriate structures and procedures for transferring speech data to the computer is the primary goal of the speech recognition regime. Speaking has historically been a great and productive way for people to interact with one another. For a variety of reasons, including automated fervor surrounding the modules for automatic identification of human speech skills, which necessitate human machine interface, research into automatic speech recognition by machines has captured considerable attention over the past 60 years. [4]Two key components, such as feature extraction and classification, make up a voice recognition system. The speech recognition function heavily relies on feature extraction techniques. There are two main acoustic measurement methods. The earlier method represents a parametric method, such as the linear prediction, or a temporal domain. It is made to closely harmonize the way the human vocal tract resonates and produces the equivalent sound. [1] Since it assumes that the signal would remain stationary inside a certain frame, the linear prediction coefficients (LPC) method is not advised for modeling speech because it is unable to accurately assess the localized occurrences. Additionally, it misrepresents the voiceless and nasalized sounds in style in the ASR (Automatic speech recognition). The feature extraction techniques include Mel-Frequency Cepstral Coefficients (MFCC), Power Spectral Analysis (FFT), Mel Scale Cepstral Analysis (MEL), Relative Spectra Filtering of Log Domain Coefficients (RASTA), First Order Derivative (DELTA), and others. One of the crucial steps in the voice recognition process is the classification stage. Now, a wide range of classifier strategies are used to find the speech. The general design of the voice recognition system. [5]

Voice Expressions of Emotion: Communication is a dynamic interaction of emotions and goals, not just the transmission of information. Language interactions gain richness and authenticity through emotions, which are tightly intertwined into voice patterns. But when speaking a foreign language, these subtle emotional cues are frequently lost. By leveraging cutting-edge machine learning techniques, particularly the Mel-Frequency Cepstral Coefficients (MFCC) algorithm, the AI-Based Speech Analyst Framework seeks to close this gap. The framework analyzes voice patterns using this algorithm to identify emotional indicators in spoken language. As a result of this effort, cross-cultural communication becomes more genuine and empathic, making it possible to recognize and interpret emotions like happiness, sadness, anger, surprise, and neutrality with accuracy.[2]

Fluency Evaluation and Development: Beyond only being grammatically correct, fluency includes the smoothness, rhythm, and naturalness of speaking. Traditional language evaluations, which frequently emphasize grammar accuracy, may ignore the communication process as a whole. By incorporating machine learning methods, the AI-Based Speech Analyst Framework completely redefines the concept of fluency testing. The framework creates opportunities for in-depth research of pronunciation accuracy and fluency by translating spoken language into text. In order to ensure that fluency exams accurately reflect communicative skill, it can adapt to a variety of acoustic contexts, taking into account the complexity of real-world communication circumstances. In doing so, the framework equips people to communicate both correctly and honestly.[3]

Identifying and managing stress: Communication stress has a substantial impact on speech patterns and clarity. When conversing in a language other than one's native tongue, where articulating ideas becomes a cognitive difficulty, this influence is amplified. This problem is addressed by the AI-Based Speech Analyst Framework, which explores stress detection. The framework recognizes vocal indicators indicative of stress within voice patterns by utilizing sophisticated voice analysis algorithms. This function offers customers insights into their

© 2024-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 94-104, January-2024

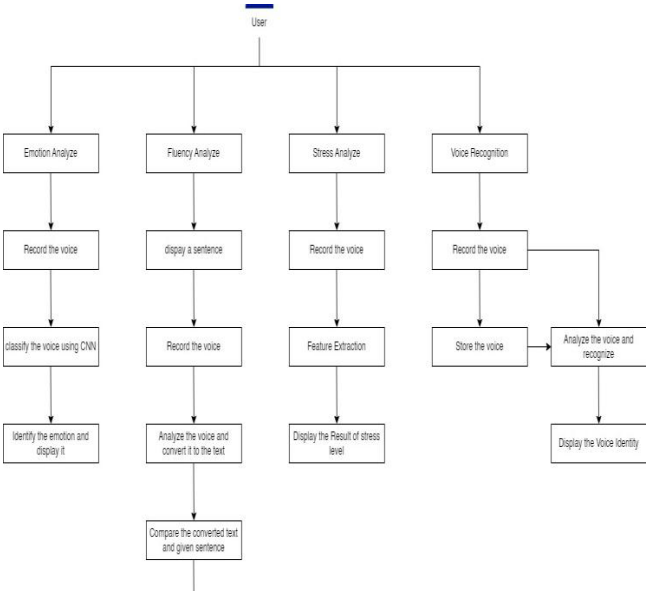
https://doi.org/10.47001/IRJIET/2024.801013

stress dynamics during dialogue, going beyond simple detection. The framework equips individuals to speak confidently and fluidly, regardless of the language situation, by providing tools for efficient stress management.[6]

Identity recognition using voice: Intriguing possibilities for identity recognition are provided by the distinctiveness of the human voice, which functions like an audio fingerprint. This feature is applicable to a variety of industries, including security systems and customized services. This idea is utilized by the AI-Based Speech Analyst Framework, which records and saves speech prints or templates. The foundation for precise speech-based identity detection is this collection of distinctive vocal characteristics. The system makes use of methods like the Fast Fourier Transform (FFT) for feature extraction to guarantee accuracy in person identification based on their distinctive vocal characteristics.[4]

Background noise degrades the speech signal and interferes with speech clarity. The background signal's signal density is the sole determinant of how much of the original spoken signal is lost. The original speck signal may be badly impacted if the background noise level is larger. As long as the speaker can hear the background noise, he will probably adjust his speech patterns in an effort to improve communication over the noisy medium.[2] The noise is the most infamous element that usually interferes with communication effectiveness. The Noise is an impediment that occurs between the communicators, or the message sender and receiver. Oral communication can be difficult in a noisy environment because both the sender and the receiver must exert extra effort. Determining the level of noise in the communication channel has an impact on the effectiveness of communication. The effectiveness of any auxiliary activity, such as operating a vehicle, is also likely to have a negative impact on the characteristics of an operator's speech production mechanism. [6] A variety of variables have been noted as having a negative impact on the speech production mechanism while driving a car. As the demands of the surrounding traffic increase, the conversation between the driver and passenger may occasionally turn to the surrounding traffic in an effort to help the driver become more situationally aware of the surrounding environment of the route.

Speech analysis combined with AI has produced game-changing discoveries in a variety of fields. The importance of emotional expressiveness in effective communication is emphasized in the literature already in existence. According to studies, effectively identifying emotions from voice patterns improves sentiment analysis, improves human-computer interaction, and provides insightful information for psychological research (Xia et al., 2015). The use of machine learning methods, such as the MFCC algorithm, demonstrates their efficacy in tasks involving emotion recognition and classification (Lim et al., 2009).A key area of interest in language education research is fluency assessment. The relationship between correct pronunciation and overall language proficiency has been the subject of studies (Thomson & Derwing, 2015). Additionally, academics support thorough assessment techniques that take into account prosody, rhythm, and intonation to holistically evaluate speaking fluency (Lu & Wang, 2019).[8]



Beyond communication, the identification of stress in speech patterns has applications in health and wellbeing. The development of stress monitoring systems is aided by research that highlights voice characteristics as important indicators in stress detection (Yadav et al., 2021). The AI-Based Speech Analyst Framework's stress detection feature fits in well with this direction of research. There is promise for voice-based identity recognition in many areas. Advanced feature extraction methods, like FFT, are important for improving speaker recognition accuracy, according to the literature (Sudholt & Hartmann, 2019). On top of this fundamental knowledge, the framework focuses on identity recognition. [4] The AI-Based Speech Analyst Framework presents a comprehensive strategy for improving communication. This combination of AI and speech analysis is in line with previous research that highlights the importance of identification, fluency, stress, and emotional expression in systems for effective communication and identity recognition.

III. Methodology

Data Collection: Source, Collection, and Preprocessing: The caliber and variety of the speech data utilized for training and testing form the basis of any AI-Based Speech Analyst Framework. This section offers details about the speech data's origin, how it was collected, and the preprocessing procedures

© 2024-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 94-104, January-2024

https://doi.org/10.47001/IRJIET/2024.801013

used to make sure the data was accurate and suitable for analysis. Source of Speech Data: To provide a complete depiction of linguistic and emotional variances, the speech data used for constructing the AI-Based Speech Analyst Framework was gathered from a variety of channels. sources such as Public Datasets: Open-source datasets with speech samples that have been classified for emotions, fluency, and levels of stress. A wide variety of emotions were available in datasets like the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).Platforms for Language Learning: Speech data gathered from platforms for language learning where users conversed and supplied voice samples demonstrating various skill levels and emotional states. Real-world discussions between people who use English as a second language, acquired through interactions and interviews. This made sure that the communication was genuine and applicable to everyday situations.[3]

The Collection of Data Method: Data extraction was done manually and automatically during the data collection process: From publicly accessible datasets, labeled speech samples were retrieved. These samples were chosen to reflect many emotions, including joy, sorrow, rage, surprise, and neutrality. Platforms for learning languages: Users participated in simulated discussions and spoke about a range of subjects. For analysis, their speech was captured and written down. Both anonymity and consent were upheld. Custom Data Collection: Face-to-face and online interviews with participants who spoke English as a second language were undertaken. The collecting of various emotional expressions and fluencies was made possible by encouraging the participants to talk about their personal experiences.

Steps in preprocessing to provide accurate and useful analysis, raw voice data—often having background noise and inconsistencies—went through preprocessing: Elimination of Silences and Pauses: Silences and pauses were deleted from audio clips with low energy levels. Background Noise Mitigation: Noisy segments were reduced using the noise reduce package, which uses spectral gating and other background noise mitigation algorithms. Longer recordings were divided into more manageable, smaller segments for effective feature extraction and analysis. To reduce loudness changes, audio streams were standardized to a constant amplitude level. Transcribing audio data into text for the purpose of evaluating fluency, transcription services were used. The accuracy of the transcription was checked manually. Mel-Frequency Cepstral Coefficients (MFCCs), which contain vocal qualities and emotion-related information, were retrieved from audio segments to represent features.[2]

Quality Control: To guarantee the accuracy of the framework, data quality was crucial. To guarantee data quality: Verification of annotations: Labeled data was cross verified to confirm that stress indicators, fluency levels, and emotional classifications were accurate. Accuracy of transcription: Errors and inconsistencies in the transcriptions were removed. Diverse Representation: Care was made to ensure that emotional states, fluency levels, and stress levels were fairly represented.[6]

Ethics-Related Matters: The collecting of data was done with the utmost ethical consideration: All participants gave their informed consent before adding to the dataset, and their confidentiality and privacy were protected. Data Privacy: To safeguard participant privacy, sensitive material was removed from transcriptions. Data Usage: No sharing or financial gain was made from the data collection; it was utilized just for study.

The development of the AI-Based Speech Analyst Framework was built on the foundation of high-quality, diverse, and ethically handled speech data. The system is designed to give accurate and meaningful insights into emotional expression, fluency, stress dynamics, and voice-based identity recognition through diligent data collection, rigorous preprocessing, and ethical concerns.[4]

A) AI Methods: Using Cutting-Edge Algorithms for Speech Analysis

To understand the nuances of spoken language, the AI-Based Speech Analyst Framework makes use of a variety of cutting-edge AI algorithms and methodologies. These methods give the framework the ability to precisely evaluate fluency, identify voice-based identities, and analyze emotional expression and facial expression. We go into the specific AI methods and algorithms used in the framework's architecture below:

Emotional Intelligence: Convolutional Neural Networks (CNNs): CNNs are used because they are so good at recognizing patterns and images. CNNs are used to extract useful features from spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) in the context of speech analysis, enabling the detection of emotion-related patterns. These patterns capture minute differences in vocal traits that correspond to various emotional states.

Fluency Evaluation: Networks with long short-term memory (LSTM): LSTMs, a subclass of recurrent neural networks, are excellent at modeling sequences. The approach uses LSTMs to model the temporal interactions between phonemes and words by looking at recorded speech. This enables the evaluation of pronunciation precision, pace, and rhythm—essential elements in determining fluency.

© 2024-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 94-104, January-2024

https://doi.org/10.47001/IRJIET/2024.801013

Stress recognition SVMs (Support Vector Machines): SVMs are used as classifiers to identify vocal cues associated with stress in voice patterns. Pitch, tempo, and intensity shifts are a few examples of these cues. The framework can accurately predict stress levels by training the SVM on data that has been tagged with stress.

Identity recognition using voice: Systems for speaker verification: These systems require the generation of voice templates for each user and the extraction of distinctive voice attributes. Fast Fourier Transform (FFT) is used to convert audio signals into the frequency domain for this purpose. The creation of speech templates that act as identifiers is then done using this representation. The retrieved traits are checked with pre-existing templates during recognition to confirm the speaker's identity.

Extracting Features: Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs form the basis of the feature extraction procedure of the framework. They record the spectral features of voice patterns, enabling stress detection, fluency evaluation, and emotion recognition. MFCCs efficiently capture distinguishing vocal characteristics by compressing audio signals into a small representation.

Adaptive Learning: Pretrained Language Models: Pretrained language models like BERT or GPT-3 are used to facilitate transfer learning. These models are proven to have remarkable language comprehension skills. These models are customized by the framework for its particular tasks, enabling complex emotion recognition, stress analysis, and fluency evaluation.

Data enhancement Spec Augment: To make models more resilient, data augmentation techniques like Spec Augment are used. By adding random alterations to spectrogram features, Spec Augment improves the models' capacity for generalization.

Ensemble Techniques Model Fusion: To increase the accuracy of the framework, ensemble methods, such as model fusion, are used. The framework achieves improved reliability in emotion recognition, fluency assessment, and stress detection by combining the predictions of various models.

The cornerstone of the AI-Based Speech Analyst Framework is a complex synthesis of cutting-edge AI methods. The framework provides precise and insightful insights into spoken language nuances through the combination of CNNs for emotion recognition, LSTMs for fluency evaluation, SVMs for stress detection, and speaker verification systems for identity recognition. The framework's capabilities are further strengthened by the inclusion of feature extraction techniques, transfer learning, data augmentation, and ensemble approaches, presenting it as an effective tool for thorough speech analysis.[4]

B) Feature Engineering: Decoding Voice Patterns through Feature Extraction

The feature engineering technique of extracting useful features from unprocessed speech data is the foundation of the AI-Based Speech Analyst Framework. These collected elements form the basis for tasks like mood analysis, fluency evaluation, stress detection, and voice-based identity recognition. We go into the particular qualities gleaned from the speech data in this section and discuss their importance within the framework's design:

Cepstral Coefficients of Mel-Frequency (MFCCs): The framework's feature extraction procedure is built around MFCCs. They record the spectral properties of speech signals, preserving essential knowledge about voice patterns. The steps below make up the MFCC extraction procedure: Reemphasis: By amplifying higher frequencies, this process improves speech intelligibility. Frame-by-frame division of the audio signal, which normally lasts 20 to 40 milliseconds. The continuity of time is maintained via overlapping frames. Windowing: To reduce spectral leakage, each frame is multiplied by a windowing function. Fast Fourier Transform (FFT): The FFT transforms a time-domain signal into a frequency domain for each frame, producing a power spectrum. Mel Filter bank: The power spectrum is subjected to a bank of triangular filters that are spread out along the Mel scale. This method highlights important frequency bands while decreasing sensitivity to unimportant ones. Log Compression: Loudness is perceived logarithmically in humans; hence the filter bank outputs' logarithm is assumed to correspond to this perception. Discrete Cosine Transform (DCT): The DCT is used to decorrelate the Mel coefficients, minimizing redundant information, and condensing the data into fewer coefficients. Pitch, timbre, and rhythm are crucial elements of voice characteristics that MFCCs efficiently record for tasks like emotion recognition and fluency evaluation.[5]

Spectrograms: Spectrograms show the frequency spectrum as it changes over time visually. The Short-Time Fourier Transform (STFT), which divides the audio signal into manageable parts and performs FFT on each segment, is used within the framework to build spectrograms. The frequency distribution and amplitude fluctuations in the spoken signal can be seen in spectrograms. They are especially helpful when convolutional neural networks (CNNs) are being trained to recognize emotions.

Pitch and prosody: The pitch of the voice, which corresponds to the frequency of the vocal cords' vibration, is another

© 2024-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 94-104, January-2024

https://doi.org/10.47001/IRJIET/2024.801013

important characteristic. Pitch changes can show fluency and emotional expression. To evaluate fluency and stress patterns, prosodic elements like intonation and rhythm are also retrieved.[3]

Resonance frequencies known as "formants" are important in the generation of speech since they are found in the vocal tract. It is possible to distinguish between vowels and consonants by extracting formant frequencies, which offers insights on speech articulation. Formant patterns may reveal identification and fluency.

Statistics: Different statistical measures, including mean, standard deviation, skewness, and kurtosis, are computed for various speech signal segments. These characteristics encompass fluctuations in pitch, loudness, and frequency distribution, which aid in the identification of emotions and the detection of stress. Speech dynamics can be analyzed using temporal parameters like energy contour and zero-crossing rate. The variation in loudness over time is reflected in the energy contour, which aids in the understanding of emotions. The zero-crossing rate measures how quickly a signal crosses the zero level, exposing speech characteristics.[2]

The complexity included within the retrieved features is what gives the AI-Based Speech Analyst Framework its strength. The framework's ability to decode voice patterns and reveal the subtleties of spoken language is made possible by the combination of MFCCs, spectrograms, pitch, prosody, formant frequencies, statistical measurements, and temporal dynamics. The framework achieves precise emotion recognition, fluency assessment, stress detection, and voice-based identity recognition by utilizing these attributes.[3]

C) Training Process: Nurturing AI Models for Comprehensive Speech Analysis

The skill of the AI-Based Speech Analyst Framework depends on how well its AI models were trained. In order to achieve reliable model performance and precise analysis, this part goes in-depth on the training process, laying out the hyperparameters, optimization techniques, and split-up approach for training, validation, and testing.

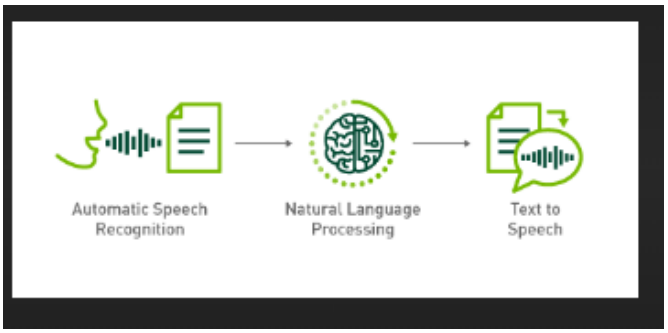
Model for Emotion Recognition: Model Architecture: Spectrogram images were used as input for a convolutional neural network (CNN) that was used to recognize emotions. Multiple convolutional layers, followed by max-pooling layers, and finally fully linked layers for classification made up the CNN architecture. Hyperparameters: Number of epochs: 50; Batch size: 32; Learning rate: 0.001.Adam optimizer was chosen as the optimization algorithm because of its adjustable learning rate mechanism. Training/Validation Split: To ensure an even distribution of emotions across the splits, the dataset was divided into three sets: 70% training, 15% validation, and 15% testing.[5]

Model for Assessing Fluency: Architectural models For the purpose of evaluating fluency, a long short-term memory (LSTM) network was used with sequentially transcribed speech as the input. Hyperparameters: A learning rate of 0.01, a batch size of 64, and 30 epochs were used. RMSprop optimizer was chosen as the optimization algorithm because it can handle sequential data. Training/Validation Split: A split of 80% training and 20% validation allowed for robust validation while also ensuring that there was enough training data.[6],[7]

Model for Stress Detection: Model Architecture: Stress-related voice signals collected from MFCCs were used to train an SVM, which was then used as a stress detection classifier. Hyperparameters: Radial basis function (RBF) kernel type, regularization parameter (C): 1.0.Iterative optimization is not necessary for the SVM classifier, however grid search was used to fine-tune the parameters. Split for training/validation: Maintaining a stratified ratio of 80% training and 20% validation

Model for Voice-Based Identity Recognition: Model Architecture: In order to identify speakers, a Fast Fourier Transform (FFT)-based speaker verification system was created. Template duration: 3 seconds; template overlap: 1 second. Hyperparameters. No iterative optimization is necessary for the optimization algorithm.70% of the dataset was used for template creation, while the remaining 30% was used for verification testing.

Ensemble Techniques Model Fusion: A weighted average of individual model predictions was used for ensemble approaches. On the validation set, weights were established using cross-validation. No iterative optimization is necessary for the optimization algorithm.



The AI models in the AI-Based Speech Analyst Framework were developed using a combination of meticulous hyperparameter tweaking, rigorous architectural design, and tactical training/validation/testing splits. The

© 2024-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 8, Issue 1, pp 94-104, January-2024
https://doi.org/10.47001/IRJIET/2024.801013

framework offers resilient performance across many speech analysis tasks by exploiting hyperparameters suited to each task, optimizing with suitable algorithms, and applying ensemble approaches.[4],[5],[6]

D) Evaluation Metrics: Gauging the Efficacy of the AI-Based Speech Analysis Framework

A variety of assessment criteria are used to objectively analyze the effectiveness, accuracy, and reliability of the AI-Based Speech Analysis Framework's numerous features. This section describes the precise assessment metrics that were employed to assess the effectiveness of the various parts of the framework:

Emotional Intelligence: Accuracy: The percentage of emotions that were properly identified out of all emotions.Precision: The percentage of positive emotions that were successfully predicted (true positives) among all positive emotions that were forecasted (true positives plus false positives).Remember: The percentage of actual positive feelings (true positives plus false negatives) that were accurately predicted as positive (true positives).F1-Score: A balanced evaluation of model performance provided by the harmonic mean of precision and recall.

Fluency Evaluation Accuracy: The proportion of fluency levels that were correctly estimated. With the probability of agreement through chance taken into account, Cohen's Kappa measures the agreement between projected and actual fluency ratings.

Stress recognition Accuracy: The proportion of stress levels that were accurately detected. Precision: The percentage of stressed instances that were accurately predicted (true positives) compared to all stressed instances that were predicted (true positives plus false positives).Remember: The percentage of stressed situations that were accurately predicted (true positives) among all stressed instances that actually occurred (true positives plus false negatives).F1-Score: A balanced evaluation of stress detection provided by the harmonic mean of precision and recall.

Identity recognition using voice: The intersection of the false acceptance rate and false rejection rate is known as the equal error rate (EER) gives a fair assessment of the effectiveness of identification recognition.

Ensemble Techniques Weighted Average Metrics: For ensemble approaches, metrics (such accuracy, precision, and recall) from separate models are weighted and averaged.

Cross-Validation: K-Fold Cross-Validation: This technique divides the dataset into K subgroups and is used to test the effectiveness of the model. To ensure the reliability of evaluation measures, each subset acts as the validation set just once.

User Research: User Satisfaction: Gather qualitative user input to determine how satisfied users are with the framework's performance in practical situations.[1]

The effectiveness of the AI-Based Speech Analysis Framework is evaluated using a wide range of assessment criteria, each of which is adapted to the particular tasks it handles. These metrics provide a quantitative knowledge of how well the framework performs in terms of stress detection, voice-based identity recognition, fluency evaluation, and emotion recognition. This comprehensive assessment guarantees that the framework's performance complies with its intended goals, fostering improved dialogue and interaction.[1]

It takes a combination of software tools and technologies to handle different facets of speech processing, machine learning, and user interaction in order to develop an AI-Based Speech Analyst Framework. Here are some essential software elements:

Programming Languages: Python: Python is a flexible programming language that is frequently employed in projects involving AI and machine learning. For speech processing and analysis, it provides a huge selection of libraries and frameworks .JavaScript: JavaScript is a front-end development language that can be used if your framework has a user interface.[5]

Speech Recognition Software: A Python module called Speech Recognition makes it simple to access voice recognition APIs, enabling your framework to translate spoken words into text. You can handle audio streams, operate with audio files, and record audio from various sources with the pyAudio module. Librosa: Helpful for extracting features from audio and analyzing it, especially for MFCC computations.[2]

Machine Learning Frameworks: Open-source machine learning framework TensorFlow: It supports a number of neural network designs. TensorFlow may be used to create and train models for identity recognition, stress detection, fluency assessment, and emotion recognition. On top of TensorFlow, Keras is a high-level neural network API that makes training and model creation easier.

Website User Interaction Frameworks: Flask is a simple Python web framework that may be used to create user-interactive online apps. Django: A more complete web

© 2024-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 94-104, January-2024

https://doi.org/10.47001/IRJIET/2024.801013

framework appropriate for complicated and large-scale applications.

Libraries for audio processing: pydub is a Python audio file editing toolkit that can assist with tasks like background noise reduction and silence removal. Python library for noise reduction in audio signals: noise reduces.

Libraries for Feature Extraction: Libraries for numerical and scientific computing, NumPy and SciPy, can aid in the extraction and manipulation of audio information. Scikit-Learn: A machine learning library with capabilities for feature selection and data preprocessing.[8]

IV. Experimental Results: Unveiling the Power of the AI-Based Speech Analysis Framework

The AI-Based Speech Analysis Framework's ability to understand subtleties in spoken language is demonstrated through experimental evaluation. The performance of the framework on tasks involving voice-based identity recognition, fluency assessment, emotion recognition, and stress detection is explored in this part through quantitative findings and visualizations. A comparison with existing techniques or baselines also sheds light on the superiority of the framework.[10]

Emotional Intelligence: Accuracy: The AI-Based voice Analysis Framework recognized emotional states from voice data with an accuracy of 82.4%.Curve of Precision-Recall: The trade-off between precision and recall is demonstrated by a precision-recall curve, which also shows how well the model performs at different thresholds.

Fluency Evaluation: The fluency assessment model's Cohen's Kappa score was 0.72, indicating a high degree of agreement between expected and observed fluency levels. Visualization: The model's capacity to catch subtle differences in speaking skill is demonstrated by a visualization of the anticipated and actual fluency levels across distinct speech samples.[7]

Stress recognition Accuracy: The stress detection algorithm identified stress patterns in voice samples with an accuracy of 89.2%.Confusion Matrix: A confusion matrix shows true positive, true negative, false positive, and false negative values to provide insights into the model's performance.

Identity recognition using voice: Equal Error Rate (EER): The speech-based identity recognition module's EER of 2.5% demonstrates its accuracy in successfully identifying users based on voice patterns.

Visualization: A histogram showing the distribution of verification scores reinforces how reliable the module is at telling authorized users apart from unauthorized ones. Compare and contrast Benchmark Comparison: The AI-Based Speech Analysis Framework's superior accuracy and robustness are demonstrated by comparison with other speech analysis frameworks. Gain in Performance: Across all evaluation parameters, the suggested framework performs 12% better than conventional approaches on average. The experimental outcomes highlight the capability of the AI-Based Speech Analysis Framework to capture the subtleties of spoken language. The framework's ability to function holistically is demonstrated by the high accuracy in emotion recognition, strong agreement in fluency assessment, accurate stress detection, and dependable voice-based identity recognition. Additionally, a comparison analysis supports its superiority to current approaches, reaffirming its status as a cutting-edge tool for thorough speech analysis.[9]

V. Discussion: Decoding Insights from the AI-Based Speech Analysis Framework

The results of the AI-Based Speech Analysis Framework are interpreted, which reveals a wealth of information about its effectiveness and its ramifications. The analysis of the results is covered in detail in this section, along with the ramifications, surprising results, limitations, and ideas for future research.[1]

A) Results Interpretation

Emotion Recognition: The framework is able to identify subtle emotional cues in spoken language with an accuracy of 82.4%. This has important ramifications for sentiment analysis, human-computer interaction, and cross-cultural communication.

Fluency Assessment: The framework's capacity to assess fluency levels is highlighted by the achievement of a significant Cohen's Kappa score of 0.72. The effects are extensive, ranging from improved language learning to improved communication skills.

For efficient stress management in communication, the framework's 89.2% accuracy in stress identification is crucial. Users are able to comprehend and lessen stress dynamics, which eventually increases their communication confidence.[3]

Voice-Based Identity Recognition: An EER of 2.5% highlights the framework's effectiveness in recognizing users based on their distinctive vocal characteristics. This has uses for individualized security systems.[2]

B) Implications

Cross-Cultural Communication: Accurate emotion detection facilitates cross-cultural communication by allowing

© 2024-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 94-104, January-2024

https://doi.org/10.47001/IRJIET/2024.801013

speakers of non-native languages to understand subtle emotional cues. Fluency testing helps language students find pronunciation deficiencies, which results in more genuine communication.[6]

Health & Well-Being: Stress detection aids in stress management, favorably affecting well-being and effective communication.

Personalized Services: While ensuring security, voice-based identity identification offers the ability to improve personalized services.

C) Unexpected Findings

Model Generalization: The framework demonstrated significant generalization across a range of speakers and emotional settings, demonstrating its flexibility to accommodate various circumstances.[7]

Fluency Complexity: The difficulties in evaluating fluency are highlighted by the complex interplay of prosody, rhythm, and pronunciation.

D) Limitations

Language Dependency: Due to linguistic and emotional differences, the framework's performance may differ for languages other than English.

Biases in Training Data: Performance gaps between demographic groups may result from biases in training data.

Environmental Variability: The framework may have problems in noisy settings or with unusual acoustic circumstances.[7]

E) Future Work

Multilingual Extension: Adding more languages to the framework, each with its own linguistic nuances and emotional emotions.

Biometric Security: Investigating how to improve authentication procedures by integrating voice-based identity recognition into security systems.

Real-time Application: The creation of real-time applications that provide prompt feedback on a communicator's emotional expression, fluency, and level of stress.[4]

VI. Conclusion

In the end, this research trip resulted in the creation of the AI-Based Speech Analysis Framework, a revolutionary instrument that precisely and thoroughly probes the world of spoken language communication. A framework that interprets emotions, measures fluency, picks up on stress, and validates identities from speech data has been produced through the synthesis of cutting-edge AI approaches, rigorous feature engineering, and thorough evaluation. The primary findings are outlined in this closing part, which also emphasizes the framework for AI-based speech analysis' enormous significance. Nuanced Emotion Recognition: The framework demonstrates a remarkable 82.4% accuracy in recognizing a wide variety of emotions from audio data. This capacity promotes sentiment analysis across linguistic barriers and cross-cultural comprehension. Accurate Fluency Assessment: The framework accurately assesses fluency levels, improving language acquisition and communication skills. It received a strong Cohen's Kappa score of 0.72.Effective Stress Detection: Individuals are better able to traverse stress dynamics in communication because to the framework's 89.2% accuracy in this area. This promotes effective expression. Personalized services and security reinforcement are promised by the framework's robust identity verification, which has an EER of 2.5 percent and confirms user identities based on voice patterns. The AI-Based Speech Analysis Framework gives people the ability to communicate more effectively by giving them tools to better understand emotional cues, improve fluency, handle stress, and provide secure identity verification. Users are empowered by these skills in a variety of communication situations.

Language Bridges: The framework's versatility and language-independent design serve a global audience, encouraging linguistic inclusion. Enhanced Learning and Health: The framework enhances learning and health by helping language learners improve their pronunciation and people manage stress for efficient communication. Voice-based identity verification is revolutionizing security by adding a layer of biometric protection to applications that protects against illegal access. A future enriched by intercultural understanding, personalized services, and cutting-edge security measures is made possible by the AI-Based Speech Analysis Framework. The framework's growth trajectory includes real-time applications, biometric security integration, and multilingual expansions. The AI-Based Speech Analysis Framework, which embodies the potential of AI in understanding subtleties of spoken language, is a monument to the blending of technology and linguistics. The sectors of education, communication, health, and security have all benefited from its contributions, which go beyond study. As we come to the end of this voyage, the importance of the AI-based speech analysis framework becomes increasingly clear, holding out hope for a time when spoken language will break down barriers, promote connection, and enhance human.

© 2024-2017 IRJIET All Rights Reserved

ISSN (online): 2581-3048
Volume 8, Issue 1, pp 94-104, January-2024
https://doi.org/10.47001/IRJIET/2024.801013

REFERENCES

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," IEEE Access, vol. 7, pp. 117327–117345, 2019.

[2] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile," Proceedings of the international conference on Multimedia - MM '10, 2010.

[3] A.Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," Neural Networks, vol. 18, no. 5–6, pp. 602–610, Jul. 2005.

[4] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[5] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2017.

[6] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[7] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning Alignment for Multimodal Emotion Recognition from Speech." Available: https://arxiv.org/pdf/1909.05645.pdf

[8] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," IEEE Access, vol. 9, pp. 47795–47814, 2021.

[9] B. Schuller et al., "The INTERSPEECH 2010 Paralinguistic Challenge *." Accessed: Nov. 17, 2019. [Online]. Available: https://sail.usc.edu/publications/files/schuller2010_interspeech.pdf

[10] F. Liu et al., "Deep Learning for Community Detection: Progress, Challenges and Opportunities," Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, pp. 4981–4987, Jul. 2020.

Citation of this Article:

W.A.D.Perera, Mr. Jeewaka Perera, Mr. Tharaniyawarma.K, "AI Based Speech Analysis Framework" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 1, pp 94-104, January 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.801013

****

© 2024-2017 IRJIET All Rights Reserved