IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AUDIENCE INTERACTION IN PUBLIC SPEECHES FOR FUTURE IMPROVISATION

[1]V. Simhadri, [2]Y. Pushpa Keerthi, [3]N. Lavanya, [4]D. Sai Bharath [5]P. Venkat Sai

[1]Asst Prof, [2]Student(20551A04C5), [3]Student(20551A0499), [4]Student(20551A0476), [5]Student(20551A04A3)

ELECTRONICS AND COMMUNICATION ENGINEERING,

GODAVARI INSTITUTE OF ENGINEERING AND TECHNOLOGY(AUTONOMOUS), RAJAHMAHENDRAVARAM

Abstract: This paper presents a method for decoding audience interaction in public speeches using advanced signal processing techniques implemented in MATLAB. By analyzing audience responses such as speech, applause, and laughter in real-time based on spectral and temporal features, the system provides immediate feedback to speakers. Through simultaneous recording of the speaker and audience, nuanced delivery and reactions are captured, enhancing analysis. The process involves noise removal and foreground extraction to isolate significant audience reactions. Thresholding methods identify key responses like applause and laughter, aiding speakers in targeted improvements. The study contributes to refining communication accuracy and event management, benefiting both speakers and researchers in communication disciplines.

Keywords: Public Speaking, Machine Learning, Signal Processing, Quality-Aware Techniques, Audience Interaction.

I. INTRODUCTION

Public speaking is a vital form of communication, enabling individuals to convey ideas, inspire change, and engage audiences. However, understanding how an audience reacts to a speech remains a challenging aspect of the speaking process. The ability to decode audience interaction can significantly impact the effectiveness of speeches, providing speakers with valuable insights into audience engagement and reactions. Additionally, a high-fidelity speaker recorder captures the nuances of the delivery, including pitch, tone, and clarity, facilitating targeted improvements for the speaker.

In this paper, we present a quality-aware Speech detection and classification system, aimed at personal growth in public speaking skills. Our approach integrates machine learning algorithms with advanced signal processing techniques to extract informative features from Audio Signals while addressing common challenges such as noise and artifacts. By incorporating quality-aware mechanisms, our system can adaptively adjust its performance based on signal quality, enhancing overall reliability in audio signals.

II. Literature Survey:

Yasuyuki Nakajima et al. (1999) proposed an approach for quickly and precisely classifying audio MPEG-coded data, having particular focus on categorizing the audio clips into four group: speech, music, applause, and silence. The system is made to be flexible enough to work with a variety of sound sources and show a good level of accuracy when it comes to identifying quiet and speech parts. The importance of proficient analysis of audio-visual information concerning the organization, retrieval, and study of multimedia content.

Min Xu et al. (2003) proposed a framework to study the temporal structure of live broadcast sports videos. This presented analysis mainly focuses on detecting events in sports videos using a fusion scheme that combines visual and auditory information. The primary objective of the paper is to bridge the gap between the simplicity of available visual and auditory features and the complexity of user semantics in sports video indexing. This methodology involves Semantic shot classification, Auditory Signal Segment, and Event Detection Process.

Hugo Bohy et. al. (2022) proposed a classification of smiles and laughter, presenting a deep learning-based system that incorporates both audio and visual. The authors also explore the fusion of these modalities to enhance the overall accuracy of classification. Significantly, they underscore the impact of intensity levels on the behavior of classification models, including that the connection between smiles and laughter may not be straightforward. The study highlights the necessity for more refined methods in dealing with these expressions.

Li Lu et.al. (2010) proposed a method for audio event detection, for cheering and applause specifically. This system was given 8 hours of TV programming to see how well it would work, the average F value came out 79.71%. They says that SVM stands out even among the Gaussian Mixture Model (GSM) but their emphasis on cheering and applause could that judgement. The author also noted that it's hard to tell the difference between the two given someones subjective opinion, making them overlap. The algorithm was tested using 8 hours of TV shows, consisting mostly of things like presents an approach using SVM's for audio event detection in TV programs.

Min Xu et. al. (2005) proposed a content analysis of comedy and horror videos was created, concentrating on parts that's cause the viewer to experience strong emotions like fear or laughter. A direct representation of the emotions of the audience may be obtained by employing various audio processing techniques to identify these AEEs. The article' experimental the effectiveness of fusing visual and recall and accuracy rate of over 90%, indicating the effectiveness of fusing visual and aural cues for emotive content analysis. This paper presents an efficacious way to distinguish affective content found in horror and comedy movies. Analyzing the affective content us distinct from semantic analysis as it offers a one-of-a-kind wat for viewers to access multimedia databases.

Kornel Laskowski et, al. (2009) proposed multiple-party meetings and how they can be automatically detected. The analysis demonstrates that the content of speaker contributions is only one element to understanding humorous discourses; speakers employ explicit downgrades to their statements or propositions deemed serious. Humor qualifies information; thus, its presence influences conversational settings like automatic summarization. The proposed system uses contextual features that describe laughter's spread across time and among more than two participants as a signal for a joke.

Rui Cui et, al(2003) proposed an important task of recognizing sound effects highlights in audio streams, for video summarization. The authors introduce a flexible framework that employs hidden Markov models (HMMs) to model specific sound effects like applause, laughter, and cheer. This study does not aim at classifying audio segments into pre-defined classes but rather identifies these highlights's sound effects. Thus, the proposed framework has been designed specifically not to categorize audio segments into predefined classes. Specifically, the proposed framework is designed to highlight recall and precision by identifying the desired sound effect while disregarding others.

III.EXISTING SYSTEM:

Existing systems in audio signals classification typically, a combination of signal processing techniques and machine learning algorithms are used. These systems initiate the acquisition of audio signals from speakers, which are preprocessed to remove noise and artifacts. Features such as Noise Reduction, Segmentation, Normalization, Filtering, and Feature Extraction Calculating statistical or spectral features from the audio segments, such as mean, variance, or Mel-frequency cepstral coefficients (MFCCs), are extracted from the preprocessed signals. Machine learning models, including traditional classifiers such as support vector machines or more advanced methods such as deep mechanism networks, have been trained on these features extracted from audio data to classify it into different categories or classes, such as distinguishing between stand-up comedy performances and stadium speeches.

IV.PROPOSED SYSTEM AND WORKING METHODOLOGY:

Our proposed system consists of three main stages: preprocessing, feature extraction, and classification to develop a comprehensive audio processing framework tailored specifically for analyzing stand-up comedy performances and stadium speeches. The quality and relevance of the audio data will be improved by integrating advanced audio preprocessing techniques into this system. The key features of the MFCCs will be used for capturing the characteristics of these signals. These extracted features are then used as input to machine learning classifiers like Support Vector Machines (SVM) or Random Forests for classification according to different categories such as identifying stand-up comedy performances versus stadium speeches. Furthermore, real-time analysis capabilities will be included to allow on-the-fly processing and classification of audio streams. This proposed system holds great potential for transforming performance art analysis and public speaking dynamics thereby providing insights on how audiences interact with each other during such presentations.

A) AUDIO DATABASE:

Audio database creation is gathering different kinds of audio, for example, stand-up comedy sessions, stadium speeches and other relevant sources. Additionally, this consists of CSV files that indicate that each mat file is classified as time domain features, Statistical and informative Features, MFCC Features computation. Such all Audio recordings metadata like labels or tags are carefully annotated to provide useful information about what they contain and in which contexts they were recorded. About 90% of the audios are used as training data. This part of the database is a major resource for developing and refining models/ algorithms / analytical techniques.

B) BAND PASS FILTER:

A band-pass filter is a circuit or device designed to allow a specific range of frequencies to pass through, while blocking out all others. Imagine it like a gatekeeper for sound or radio waves, only letting certain tones through based on their pitch. This is useful in many applications, like extracting a radio station's signal from all the surrounding frequencies or focusing on a particular instrument in a musical recording.

C) NORMALIZATION:

Normalization is a process in database design that organizes data into tables to minimize redundancy and improve data integrity. This involves structuring tables to avoid storing the same information in multiple places, reducing wasted space and the risk of errors when data needs updating. By following specific normal forms (increasing levels of organization), normalization ensures data is efficiently stored, retrieved, and manipulated.

D) MEAN REMOVAL:

Mean removal is a data pre-processing technique used in machine learning. It centers each feature in a dataset around zero by subtracting the mean value of that feature from each data point. This helps to remove bias from the features and improve the performance of some machine learning algorithms, especially those that rely on distance calculations or assume a normal distribution in the data.

E) SEGMENTATION:

Segmentation is the process of dividing a large market into smaller groups of people with similar characteristics. This is done in marketing to better understand customer needs and preferences. By segmenting the market, businesses can tailor their messages and products to each group, making their marketing more effective and efficient. There are many ways to segment a market, such as by demographics, interests, or behavior.

F) PEAK DETECTION:

Peak detection for audio signals is a crucial step in various audio processing tasks such as audio compression, normalization, dynamic range compression, and music analysis. Peaks represent the highest points or maximum values in an audio signal. Detecting peaks helps in identifying significant events or moments within the audio.

G) TIME DOMAIN FEATURES:

This time-domain feature analysis used in our project aims at capturing important temporal attributes of audio signals. They have such measures as the mean, standard deviation, root mean square amplitude, zero crossing rate and energy entropy. Our aim is to understand the spread, dynamics and complexity of audio signal by studying these features which will enable us perform speech recognition as well as sound event detection easily. This method has improved our knowledge of audio signal processing and its use in a number of areas.

H) FREQUENCY DOMAIN FEATURES:

These features aim to analyze how the energy of the signal is distributed across different frequency bands, providing insights into the spectral characteristics of the audio. Specifically, frequency domain features may focus on measures such as spectral centroid, spectral bandwidth, spectral flatness, and spectral rolloff, which describe the central frequency, spread, flatness, and decay of the signal's frequency spectrum. By examining these features, your project aims to understand the spectral properties of audio signals, aiding in tasks such as music genre classification, sound event detection, and speech recognition.

I) STATISTICAL DOMAIN FEATURES:

statistical and informative features refer to various characteristics derived from audio segments. SDF calculates the standard deviation of features within each segment, offering insights into their variability This feature can include standard deviation, mean, variance, root mean square of successive differences (power spectral density), we extract essential statistical metrics to characterize the audio data, enhancing our understanding of its properties and facilitating further analysis.

J) MEL – FREQUENCY CEPSTRUM COEFFICIENT:

Mel-frequency cepstral coefficients (MFCC) from audio segments using MATLAB functions. Specifically, we utilize the vs_mfcc1 and vs_mfcc2 functions to extract MFCC features with 12 coefficients each. MFCCs provide a compact representation of the spectral envelope of the signal, making them useful for tasks such as speech recognition, audio classification, and speaker identification.

K) MULTI LAYER PERCEPTION:

The neurons in the MLP are trained with the back propagation learning algorithm. MLPs are designed to approximate any continuous function and can solve problems which are not linearly separable. The major use cases of MLP are pattern classification, recognition, prediction and approximation.

L) K-NEAREST NEIGHBORS (KNN):

K-Nearest Neighbors (KNN) is a simple but powerful algorithm used in machine learning for classification and regression tasks. It works on the principle of proximity, where the unlabeled sample is classified based on the class of its nearest neighbors in the feature space. In KNN, 'K' represents the number of neighbors considered for classification, which is usually determined through cross-validation.

M) RANDOM FOREST:

Decision trees are the fundamental building blocks of Random Forest, where each tree is trained on a random subset of the audio data and a random subset of features. By aggregating the predictions from multiple decision trees, Random Forest achieves robust classification performance, effectively capturing the complex relationships present in audio signals. This ensemble learning approach enhances the accuracy and reliability of our classification model, enabling precise classification of audio signals into different categories or classes.

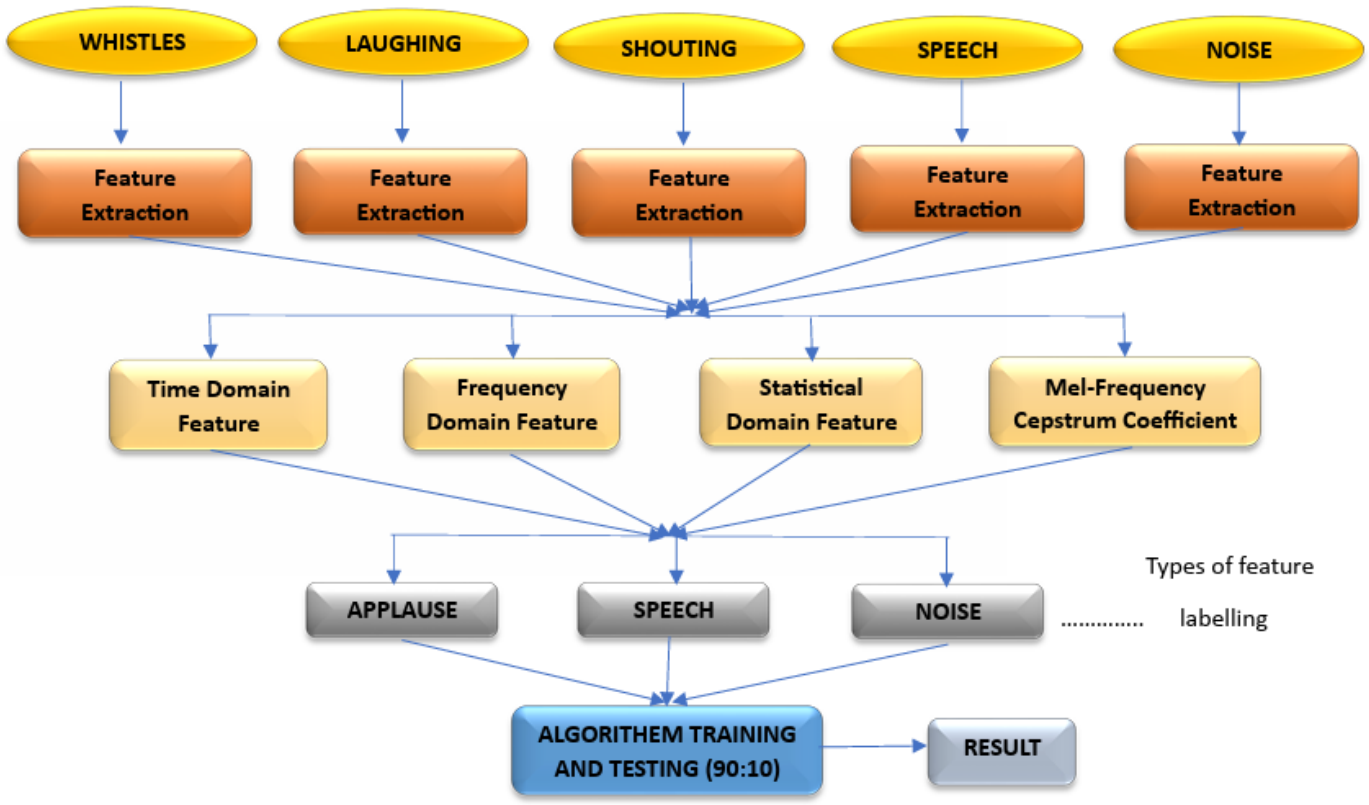
N) RANDOM TREE:

random tree algorithm with Weka for audio signal data training and testing, several key steps are essential. These include feature extraction to capture relevant audio characteristics, preprocessing for data cleaning and normalization, model training using random subsets of features and instances, and finally, evaluation to gauge model performance. Through these steps, the algorithm efficiently processes audio data for classification or regression tasks, offering insights into its practical application in machine learning workflows.

O) NAÏVE BAYES:

It is based on Bayes' theorem and the "naïve" assumption of feature independence, where features are assumed to be conditionally independent given the class label. We evaluate the model performance using error metrics including Absolute Error, Root Mean Squared Error (RMSE), Relative Absolute Error, and Root Relative Squared Error. Naive Bayes is known for its simplicity, robustness, and scalability, making it a popular choice for classification tasks in diverse fields.

V.BLOCK DIAGRAM FOR THE PROPOSED MODEL:



VI. RESULT:

A) TIME DOMAIN FEATURES:

Feature Number	Features name	Accuracy	Algorithm
F20	Kurtosis and derivative of the signal	80.2%	Random Forest
F20	Kurtosis and derivative of the signal	80.2%	Random Tree
F20	Kurtosis and derivative of igna signal	80.2%	IBk(KNN)
F18	Max. value of derivative l	77.8%	Random Tree
F6	Max. absolute of segment	77.8%	Random Forest
F21	Pitch period of segment using Autocorrelation	67.2%	Multi Layer Perception

B) FREQUENCY DOMAIN FEATURES:

Feature Number	Features name	Accuracy	Algorithm
F35	Spectral Centroid of a derivative segment	79%	Random Forest
F35	Spectral Centroid of a derivative segment	79%	Random Tree
F26	Spectral bandwidth of a segment	78.1%	IBk(KNN)
F46	Relative entropy	76.9%	Random Forest
F44	Spectral Kurtosis of a derivative segment	75.9%	Random Tree
F25	Spectral centroid of a segment	65.1%	Multi Layer Perception

C) STATISTICAL DOMAIN FEATURES:

Feature Number	Features names	Accuracy	Algorithm
F58	Max value of mode centre frequencies	79%	Random Forest
F60	Standard deviation of mode centre frequencies	78.2%	Random Tree
F67	Energy computed for bands in range (300-4000HZ)	77.7%	Random Tree
F64	Energy computed for bands in range (500-2000HZ)	76.7%	IBk(KNN)
F60	Standard deviation of mode centre frequencies	64.4%	Multi Layer Perception
F51	Max value of normalized mode energies	60.4%	Naïve Bayes

D) MEL – FREQUENCY CEPSTRUM COEFFICIENT:

Feature Number	Features names	Accuracy	Algorithm
F148	Modified Delta Delta-2 Feature	82%	IBk(KNN)
F148	Modified Delta Delta-2 Feature	82%	Random Forest
F80	MFCC-12 Feature	81.7%	Random Tree
F82	Delta-2 Feature	70.4%	Multi Layer Perception
F82	Delta-2 Feature	70.3%	Naïve Bayes
F137	Modified Delta-2 Feature	69.5%	Multi Layer Perception

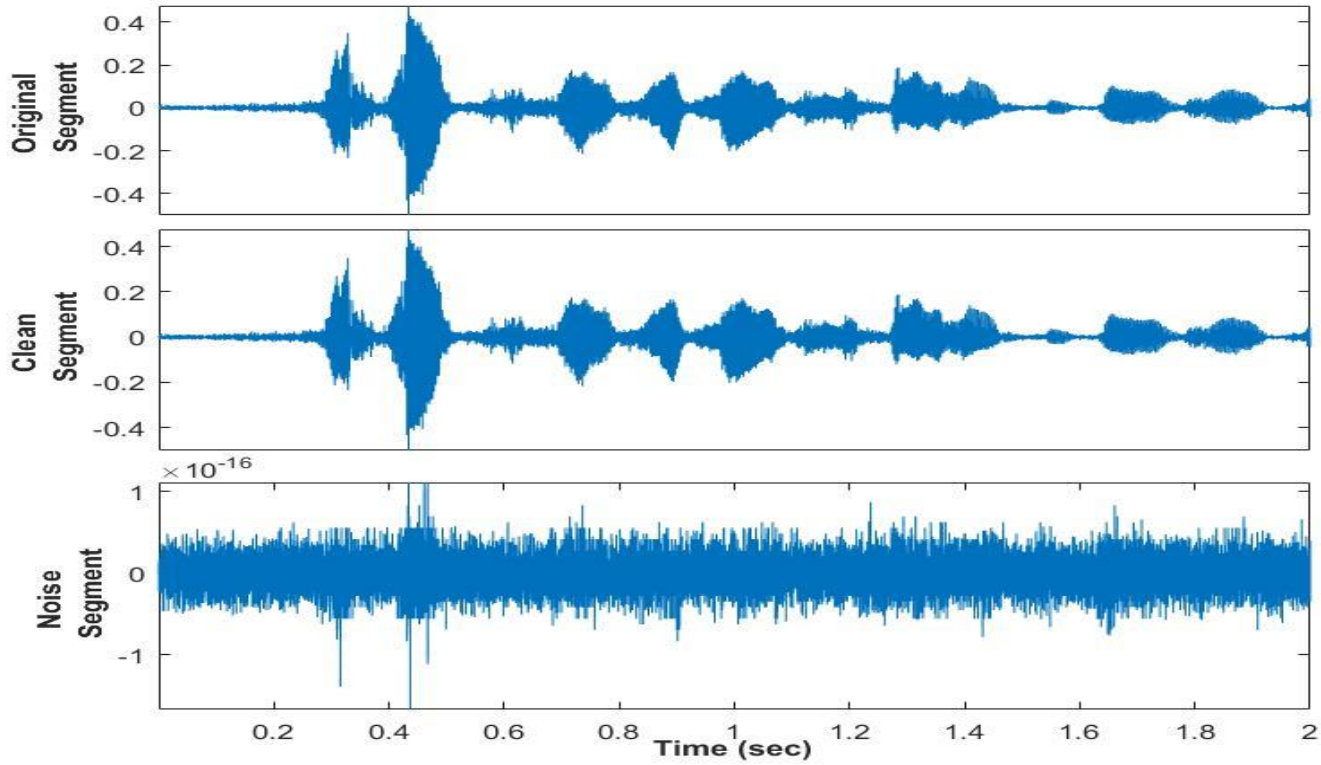


FIGURE - 1

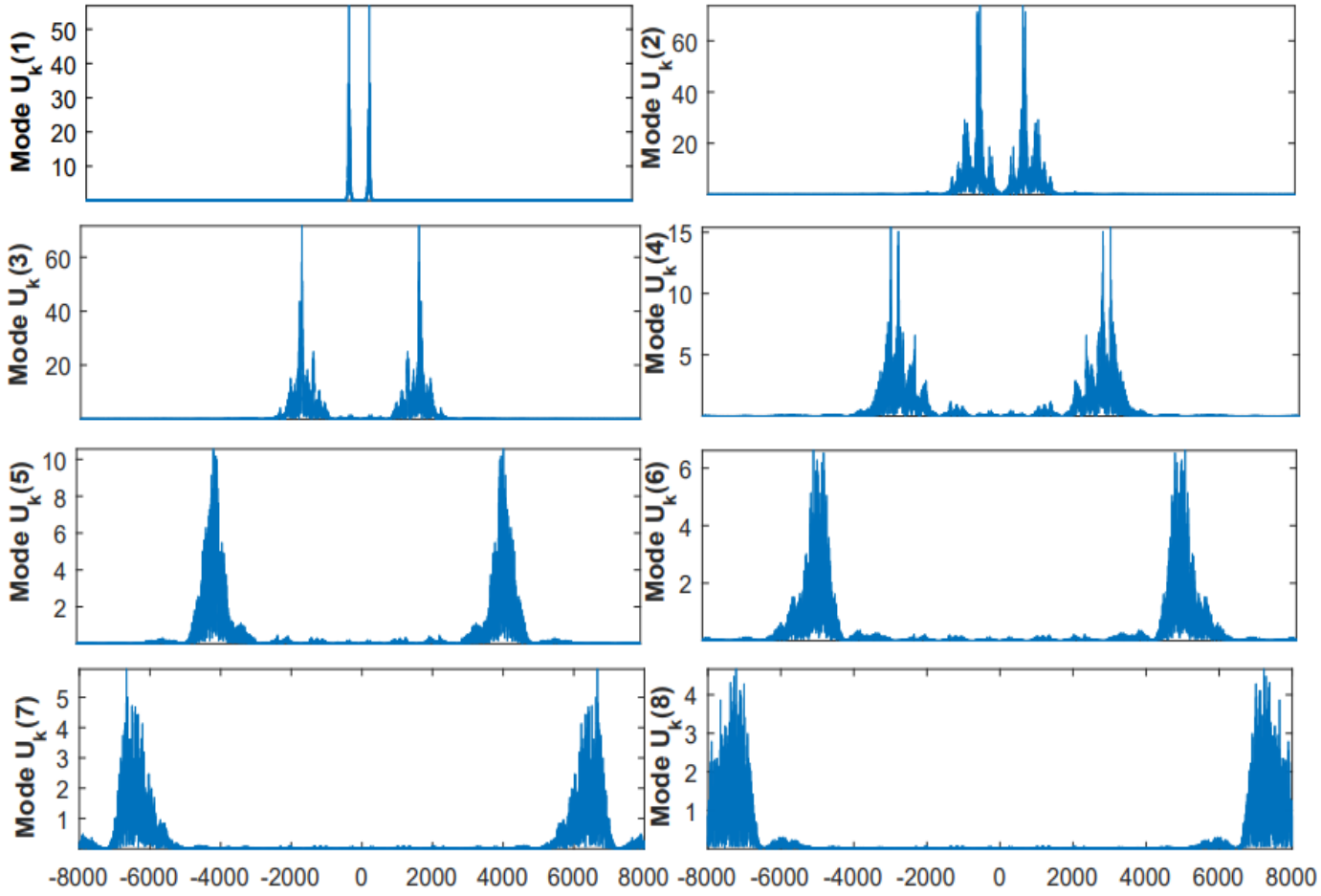


FIGURE - 2: DECOMPOSED FREQUENCY MODES (MODE 1 - MODE 8)

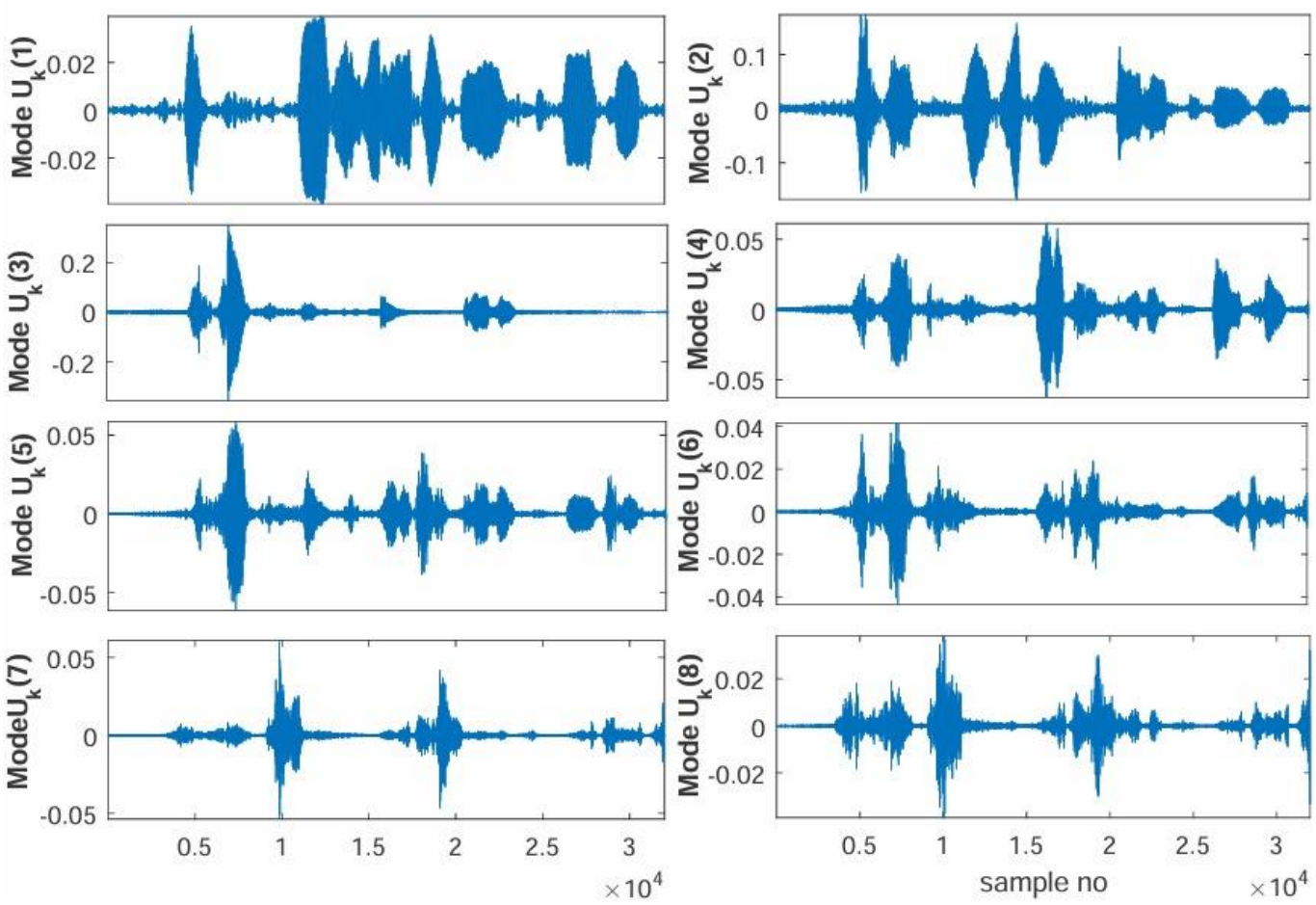


FIGURE-3: DECOMPOSED TIME MODES (MODE1 - MODE8)

VII. CONCLUSION:

In this paper, our project successfully developed a sophisticated detection and classification method to accurately identify audience interaction during public speeches. By analyzing audio input, we were able to detect various audience reactions such as applause, clapping, shouting, and other vocal responses with high accuracy. This method provides valuable insights into audience engagement dynamics, enabling speakers to tailor their presentations for better interaction and impact. Our approach contributes to enhancing public speaking strategies by providing real-time feedback on audience reactions, ultimately leading to more engaging and effective communication."

VIII. FUTURE SCOPE:

In the future iterations, the project aims to extend its capabilities to run on Raspberry Pi devices. This will enable the creation of portable, real-time feedback systems for public speakers, enhancing public speaking training, event management, and market research. Additionally, further enhancements may include real-time visualization, integration with IoT devices, and continuous machine learning optimization for improved accuracy and adaptability."

IX. REFERENCES:

1. Nakajima, Yasuyuki, et al. "A fast audio classification from MPEG coded data." In 1999 IEEE Int Con on Aco, Spe, and Sig Pro. Pro. ICASSP99 (Cat. No. 99CH36258). Vol. 6. 1999.
2. M. Xu et.al., "A fusion scheme of visual and auditory modalities for event detection in sports video". In IEEE Int. Con. on Mult, and Expo. ICME'03. Proc, Cat. No. 03TH8698, Vol. 1, pp. I-333, July. 2003.
3. H. Bohy, K. El Haddad, & T. Dutoit. "A new perspective on smiling and laughter detection: Intensity levels matter". In IEEE Int. Con. on Aff, Com, and Intelligent Int. (ACII) pp. 1-8, October.2022.
4. Li. Lu et.al., "An SVM-based audio event detection system." In IEEE Int. Conf. on Ele, and Con, Eng, pp. 292-295,2010.
5. M. Xu et.al., "Affective content analysis in comedy and horror videos by audio emotional event detection". In IEEE Int. Conf. on Multi and Expo pp. 4-pp. 2005, July.
6. Zaid. Harchaoui, et.al., "A regularized kernel-based approach to unsupervised audio segmentation." In IEEE int. con. on aco, spe, and sign processing, pp. 1665-1668, 2009.
7. R. Cai, et.al., ``Highlight sound effects detection in audio stream''. In Proc. IEEE *Int. Conf. on Multimedia and Expo,* pp. III-37, Vol. 3, Jul. 2003.
8. E. Lieskovska, M. Jakubec, & R. Jarina., "Acoustic surveillance system for children's emotion detection". In IEEE Int. Conf. on Tele and Sig Pro (TSP) pp. 525-528., July.2019.
9. H. Inaguma et.al.,. "An end-to-end approach to joint social signal detection and automatic speech recognition". In IEEE Int. Conf. on Aco, Spe, and Sig Proc. (ICASSP), pp. 6214-6218. April. 2018.
10. N. Samadiani et.al., "Happy emotion recognition from unconstrained videos using 3D hybrid deep features". IEEE Access, Vol. 9, pp. 35524-35538, Feb. 2021.
11. M. Zhao, J. Bu, and C. Chen., "Audio and video combined for home video abstraction". In IEEE Int. Conf. on Aco, Spe, and Sig Pro, Proc, (ICASSP'03). Vol. 5, pp. V-620, April. 2003.
12. H. Bohy, K. El Haddad, & T. Dutoit. "A new perspective on smiling and laughter detection: Intensity levels matter". In IEEE Int. Con. on Aff, Com, and Intelligent Int. (ACII) pp. 1-8, October.2022.
13. M. Fukumto, R. Nagamatsu, "Feedback of Laughter by Detecting Variation in Respiration Amplitude for Augmenting Laughter". In IEEE 10th Int. Conf. on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), pp. 139-142, Jul. 2016.
14. M. Xu et.al., "A fusion scheme of visual and auditory modalities for event detection in sports video". In IEEE Int. Con. on Mult, and Expo. ICME'03. Proc, Cat. No. 03TH8698, Vol. 1, pp. I-333, July. 2003.
15. S. Cosentino, S. Sessa, and A. Takanishi. "Quantitative laughter detection, measurement, and classification—A critical survey". In IEEE Reviews in Biomedical Engineering, vol.9, pp.148 162.2016.
16. Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, & O. Cappé. "A regularized kernel-based approach to unsupervised audio segmentation". In IEEE int. con. on ac, spe and sig proc (pp. 1665-1668) 2009, April.
17. L. Lu, F. Ge, Q. Zhao, & Y. Yan," An SVM-based audio event detection system". In IEEE Int. Con on Elec and Con Eng, pp. 292-295. (2010, June).
18. E. Lieskovska, M. Jakubec, & R. Jarina, "Acoustic surveillance system for children's emotion detection". In IEEE 42nd Int. Con. on Tele and Sigl Proc (TSP). pp.525-528. 2019, July.
19. M. Xu, L. T. Chia, & J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection". In IEEE Int. Conf. on Mult and Expo, pp. 4-pp. 2005, July.
20. H. Inaguma, M. Mimura, K. Inoue, K. Yoshii, &, T. Kawahara. "An end-to-end approach to joint social signal detection and automatic speech recognition". In IEEE Int. Con. on Aco, Spe, and Sig Proc (ICASSP). pp. 6214-6218. 2018, April.