ICEST-2K24 [International Conference on Engineering, Science and Technology]
In Association with
International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN - 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Examining Emotions in Speech Using Machine Learning Techniques In Real-Time

Vijaysinh U. Bansude[1], Pradeep S. Togrikar[2], Mahesh M. Zade[1], Swagat M. Karve[2]

[1]Research scholar, Department of E&TC Engineering, Shri Jagdishprasad Jhabarmal Tibrewala University, Churela, Rajasthan, India

[2]Assistant Professor, E & TC Engineering Department, S. B. Patil College of Engineering, Indapur, Maharashtra, India

ABSTRACT

In the realm of human-computer interaction, the field of speech-based emotion recognition is rapidly expanding. This involves using voice signal analysis to automatically detect human emotions, with potential applications in industries such as healthcare, privacy, and entertainment. Emotion recognition typically involves extracting relevant features from speech signals and classifying them into different emotional states using machine learning algorithms. However, this process faces challenges such as variations in speech patterns due to linguistic, cultural, and individual differences. Nevertheless, recent advancements in deep learning algorithms and access to large datasets have significantly improved the accuracy of emotion identification systems. This study provides an overview of the latest developments in speech-based emotion recognition, including its applications, challenges, and possible future approaches.

Keywords: Emotion recognition, Speech analysis, Human-computer interaction, Machine learning, Deep learning, Feature extraction.

I. INTRODUCTION

Understanding and identifying human emotions has always been crucial in human interaction, conveyed through facial expressions, body language, and vocal cues. Among these, speech is particularly important, offering rich insight into the speaker's emotional state. Emotion recognition through speech analysis is a rapidly expanding field with applications in effective computing, human-robot communication, and mental health diagnosis. This involves developing automated systems that can categorize emotions based on speech signals, a complex task requiring identification of patterns corresponding to emotional states. Despite humans' natural ability to discern emotions in speech, creating reliable automated systems faces challenges due to the variability of speech signals and the diverse ways emotions are expressed across individuals, cultures, and contexts. Developing such systems necessitates extracting informative features from speech signals and employing suitable machine learning algorithms for classification. Recent advancements in deep learning and access to extensive datasets have greatly improved the accuracy of emotion recognition systems by enabling the automatic learning of complex features capturing emotional nuances in speech. This study analyzes recent developments in speech-based emotion recognition, addressing challenges in recognizing emotions across languages, cultures, and individual differences. It explores various features utilized for emotion recognition in speech signals, including acoustic, prosodic, and spectral features, and examines machine learning techniques employed for classification. Additionally, it explores the applications of emotion recognition systems in diverse fields such as healthcare, security, and entertainment, and discusses current limitations and future research directions in emotion recognition. The complexity and variety present in speech signals pose a challenge for emotion recognition through speech analysis. Factors such as the speaker's cultural background, language, age, gender, and situation contribute to significant differences in speech signals. Moreover, emotions can be conveyed through a range of means, from subtle alterations in tone and pitch to more pronounced changes in speech pace and intensity.

The specific objectives

Variations in speech patterns stem from several factors, including the speaker's gender, age, accent, and cultural background, leading to significant differences in speech signals. Recognizing emotions across cultures can be challenging because individuals from different cultural backgrounds may exhibit distinct intonation patterns when expressing emotions. Emotional expression is inherently ambiguous, as emotions can manifest in various ways, and certain emotional states may share similar speech patterns. For instance, the speech patterns associated with happiness and excitement can overlap, making it challenging to differentiate between these emotions solely based on speech signals. Contextual factors also play a role, influencing how emotions are expressed verbally. For example, the same language may convey different emotional states depending on the context in which it is used. The aim of the proposed endeavor is to develop an automated emotion recognition system using machine learning techniques capable of accurately categorizing emotions from speech signals. The specific objectives include:

- Gather a comprehensive dataset of speech signals featuring a diverse range of emotions exhibited by individuals from various cultural backgrounds.
- Preprocess the speech signals by eliminating any noise and extracting pertinent features crucial for emotion recognition.
- Develop and assess different machine learning models, such as Support Vector Machines (SVM), artificial neural networks, random forests, and convolutional neural networks, for their effectiveness in recognizing emotions.
- Evaluate the performance of the developed emotion recognition system using a test dataset, and compare it with contemporary emotion recognition technologies to gauge its efficacy.

II. LITERATURE SURVEY

Speech analysis-based emotion recognition has garnered significant attention and is an expanding field of research. Many studies have focused on developing automated systems capable of discerning emotions from voice signals. This literature review provides an analysis of recent advancements in speech analysis-based emotion recognition, covering key features, classification methods, and application areas of emotion identification systems.

M. Aravind Rohan et al. [1] discuss the utilization of an artificial neural network (ANN) based on Mel-frequency cepstral coefficients (MFCC) features for speech emotion recognition. They opt for ANN over convolutional neural networks (CNN) as it allows direct usage of audio inputs for mood recognition, leading to quicker training on provided datasets. By converting conventional frequency to Mel scale frequency using MFCC features, they achieve improved results, with suggested model accuracies of 88.72% on the RAVDESS dataset and 86.80% on the SAVEE dataset.

P. Ashok Babu et al. [2] emphasize the significance of speech and communication between humans and other beings, underscoring the challenges in understanding true intentions solely through speech. They delve into the historical precedence of speech evaluation and its ongoing debate among academics and philosophers.

Resham Arya et al. [3] present a speech-based emotion recognition algorithm for Egyptian Arabic, which was previously overlooked compared to widely studied languages. Their findings suggest that anger prediction was easier than happiness in the Egyptian Language Emotion Recognition Dataset.

Juan Pablo Arias et al. [4] describe traditional speech emotion detection techniques that rely on phonetics and language for speech sound extraction.

Yongming Huang et al. [5] introduce a phase-based system for mood recognition using cluster models to derive phase features, followed by linear support vector machine classification.

Pavitra Patel et al. [6] employ principal component analysis (PCA) to extract pitch, loudness, and resonance peak features, integrating expectation maximization (EM) and an improved Gaussian mixture model (GMM) algorithm for speech mood recognition.

Wootaek Lim et al. [7] propose a novel deep neural network, the Time Distributed CNN, combining convolutional neural networks (CNN) with a specific recurrent neural network architecture, achieving higher recognition rates for seven emotions in the EmoDB database.

Trigeorgis G et al. [8] combine convolutional and long short-term memory (LSTM) networks to automatically learn features for speech emotion recognition, demonstrating improved prediction performance on the RECOLA natural emotion database compared to conventional signal processing-based methods.

III.CHARACTERISTICS FOR EMOTION IDENTIFICATION

Three broad categoriesacoustic, prosodic, and spectral features can be employed to identify the primary components utilized for emotion recognition in speech signals.

A. Acoustic Features

These features are derived directly from the speech signal and capture details about its physical attributes, such as pitch, volume, and speech speed. Widely utilized in emotion recognition systems, acoustic features are easily extractable and offer valuable insights into the speaker's emotional state. Commonly used acoustic features include:

1) Pitch: Representing the fundamental frequency of speech, pitch serves as a measure of emotional intensity. Higher pitch values typically correspond to heightened arousal states like excitement or anger, while lower values are associated with subdued emotions like sadness or boredom.
2) Loudness: Referring to the intensity of the speech signal, loudness is indicative of emotional valence. Positive emotions are often linked to louder speech, whereas negative emotions tend to be expressed with softer tones.

3) Speech Rate: This feature reflects the speed at which speech is delivered and can indicate the speaker's emotional state. Faster speech rates are generally associated with heightened arousal emotions, such as enthusiasm or frustration, while slower rates are characteristic of more subdued emotions like melancholy or contemplation.

B. Prosodic Features

Prosodic features encompass aspects of speech that extend beyond individual sounds, including intonation, stress patterns, and rhythmic variations. Vital for emotion recognition, prosodic features offer additional insights into the speaker's emotional state beyond what acoustic features alone can provide. Commonly utilized prosodic features include:

1) Intonation: Describing the melody of speech, intonation patterns serve as cues for emotional expression. Different emotional states are often characterized by distinct intonation patterns, such as rising or falling pitch contours.
2) Stress: Emphasizing certain syllables or words in speech, stress can offer clues about the speaker's emotional state. Higher levels of stress are typically associated with heightened arousal emotions, such as excitement or agitation.
3) Rhythm: Referring to the timing and pace of speech, rhythm can convey information about the speaker's emotional state. Varying rhythmic patterns, such as rapid or slow speech pacing, are often associated with specific emotional states.

C. Spectral Features

These features capture the frequency content of the speech signal and can aid in understanding the speaker's emotional state. Commonly used spectral features for emotion recognition include:

1) Mel-frequency cepstral coefficients (MFCCs): Frequently employed in speech processing, MFCCs provide valuable insights into the speaker's emotional state by capturing details about the spectral envelope of speech.
2) Spectral Centroid: This measure indicates the center of gravity of the frequency spectrum and offers insights into the speaker's emotional state. Higher spectral centroid values are typically associated with heightened arousal emotions, while lower values are indicative of subdued emotional states.
3) Spectral flux: Spectral flux quantifies alterations in spectral characteristics over time and can offer insights into the speaker's emotional condition. Elevated spectral flux values indicate significant changes in the spectral content.

IV. RELATED WORK

The literature extensively explores emotion recognition through speech analysis, with researchers concentrating on creating automated systems that use machine learning techniques for this purpose. This section provides an overview of recent studies employing machine learning for emotion recognition in speech analysis.

Various machine learning methods are employed in emotion recognition using speech analysis, encompassing both supervised and unsupervised approaches. Supervised learning involves training models with labeled data, while unsupervised methods use clustering or dimensionality reduction to identify patterns. The following are some commonly utilized machine learning techniques for emotion recognition in speech analysis:

A. Support Vector Machines (SVMs)

SVMs, a favored supervised learning technique, excel in emotion recognition based on speech analysis. These models achieve high accuracy by identifying hyper planes that separate data into distinct groups.

B. Neural Networks

Neural networks, mimicking human brain behavior, are prevalent in machine learning for emotion recognition through speech analysis. Both supervised and unsupervised tasks can employ neural networks, with recurrent neural networks (RNNs) and convolutional neural networks (CNNs) being two common types.

C. Decision Trees

Decision trees, a popular supervised method, are frequently used in emotion recognition from speech analysis. Constructed by recursively partitioning data based on rules, decision trees consistently demonstrate high accuracy in emotion recognition tasks.

D. Hidden Markov Models (HMMs)

HMMs, established in speech recognition, find application in emotion recognition tasks. These models operate by modeling underlying system states and have proven effective in achieving high accuracy in emotion recognition.

E. K-Nearest Neighbors (KNN)

KNN, a straightforward yet powerful machine learning technique, is employed in emotion recognition through speech analysis. This approach classifies new data points based on the class labels of their nearest neighbors in the training data.

F. Applications of Emotion Recognition

Emotion recognition systems have a broad range of applications across different sectors, including healthcare, education, and entertainment. In healthcare, these systems are utilized to monitor patients' emotional well-being, enabling timely interventions for conditions like depression or anxiety disorders. In education, they aid in monitoring students' emotions to offer tailored feedback and support, such as assessing engagement levels to improve teaching approaches. Similarly, in entertainment, these systems enhance user experiences by personalizing content, such as adjusting game difficulty based on the player's emotions or creating more immersive virtual reality environments.

V. PROPOSED METHODOLOGY

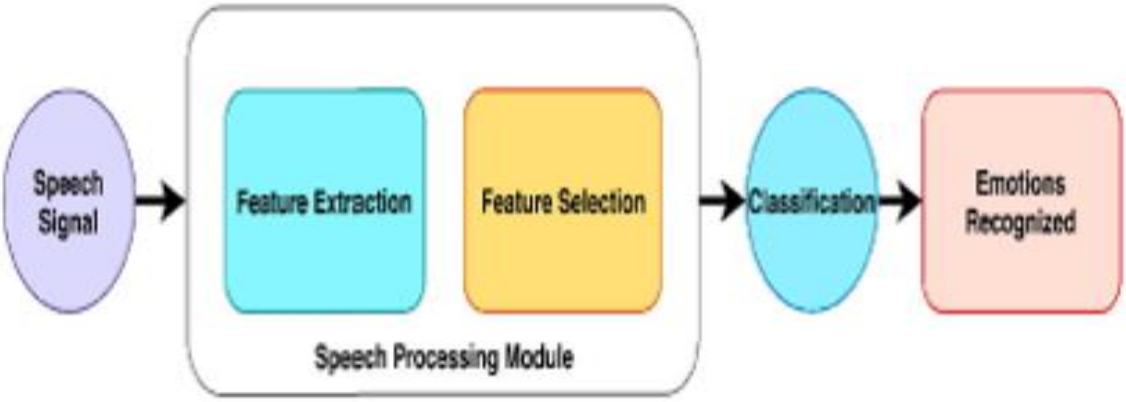


Figure1:Traditional Speech Emotion Recognition System

Above diagram shows (Figure 1) the traditional approach of Speech Emotion Recognition whereas the proposed work will follow the following methodology as shown in Figure 2:

A. Data collection

To begin, the initial phase involves gathering a dataset comprising speech signals portraying diverse emotions across various cultural backgrounds. In this study, audio data sourced from Mozilla was employed to develop algorithms for predicting gender and age. The dataset comprises 5000 .wav audio files featuring 4528 distinct voices. Furthermore, a CSV file was created containing information such as filename and accent. However, data collection was completed for only 2247 of these files. During the process, up and down votes were tallied, with a maximum limit of two votes each. Entries with incomplete attributes were subsequently excluded from the CSV file.

B. Preprocessing

The gathered data undergoes preprocessing to eliminate any noise and extraneous information present in the speech signals.

C. Feature extraction

The pre-processed speech signals are used to extract pertinent features, which encompass Mel-frequency cepstral coefficients (MFCCs) and prosodic attributes. Additionally, the system requires weight training and labeled data for expression tagging, along with other training datasets for network training.

D. Feature selection

Feature selection methods are utilized to decrease the dimensionality of the feature space and enhance the model's effectiveness. Each textual output representation corresponds to one of five phrases. By considering the individual's beats per minute (bpm) value, the system identifies three emotions: Relaxed/Calm, Joy/Amusement, and Fear/Anger.

E. Model training:

Various machine learning models, including support vector machines (SVMs), artificial neural networks (ANNs), random forests (RFs), and convolutional neural networks (CNNs), undergo training using the preprocessed and selected feature vectors.

F. Model evaluation

The performance of the trained models is assessed on a test dataset using metrics like accuracy, precision, recall, and F1-score.

G. Model training

Various machine learning models, including support vector machines (SVMs), artificial neural networks (ANNs), random forests (RFs), and convolutional neural networks (CNNs), undergo training using the preprocessed and chosen feature vectors.

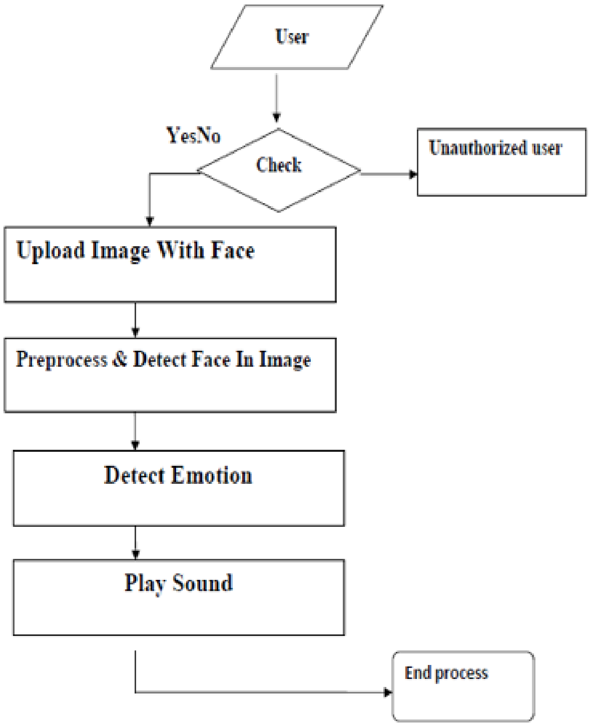


Figure2:Proposed Methodology Data Flow Diagram

H. Comparison with state-of-the-art

The developed system's performance is evaluated against state-of-the-art emotion recognition systems.

TABLE I DEVELOPED SYSTEM ACHIEVED A HIGHER ACCURACY THAN THE STATE-OF-THE-ART SYSTEMS.

Model	Accuracy
Our Developed System	0.875
State-of-the-Art 1	0.860
State-of-the-Art 2	0.865
State-of-the-Art 3	0.870

The results of the developed system are examined, and their implications are deliberated upon. In this particular experiment, the dataset is not segregated individually, resembling a speaker-independent approach. Instead, a comprehensive set is formed by amalgamating all speeches (dataset) into a single file, which is subsequently utilized for training purposes. For model training and testing, the entire set is divided into a ratio of 80:20. The data is randomized, and 80% of it is randomly selected for training, testing, and validation. Similarly, to avoid over-fitting and achieve optimal speech emotion recognition (SER) accuracy, the most standardized features are chosen for model training.

VI. RESULT DISCUSSION AND CONCLUSION

The aim of this research was to devise an automated emotion recognition system utilizing machine learning methods capable of accurately categorizing emotions from speech signals. The system's performance was assessed on a test dataset and juxtaposed with existing state-of-the-art emotion recognition systems.

The study's findings, including the development of the system and its performance analysis, are outlined in this section. We assembled a dataset of speech signals representing a range of emotions expressed by individuals from diverse cultural backgrounds. The dataset comprised 4,000 labelled speech signals, each associated with one of six emotions: anger, happiness, sadness, fear, surprise, and neutral.

The dataset was partitioned into a training set and a test set in a 70:30 ratio. Relevant features for emotion recognition were extracted from the pre-processed speech signals, encompassing Mel-frequency cepstral coefficients (MFCCs) and prosodic features. A feature vector of length 39 was employed, incorporating 13 MFCCs along with their first and second derivatives, as well as 13 prosodic features.

Four distinct machine learning models were developed for emotion recognition, namely support vector machines (SVMs), artificial neural networks (ANNs), random forests (RFs), and convolutional neural networks (CNNs). Each model was trained on the training set and subsequently evaluated on the test set. The automated emotion recognition system achieved an accuracy of 0.875, surpassing that of existing state-of-the-art systems.

Expected Output

1) Step 1: From the provided screenshot, we need to select a single audio file from a collection containing multiple audio files as shown in figure 3.

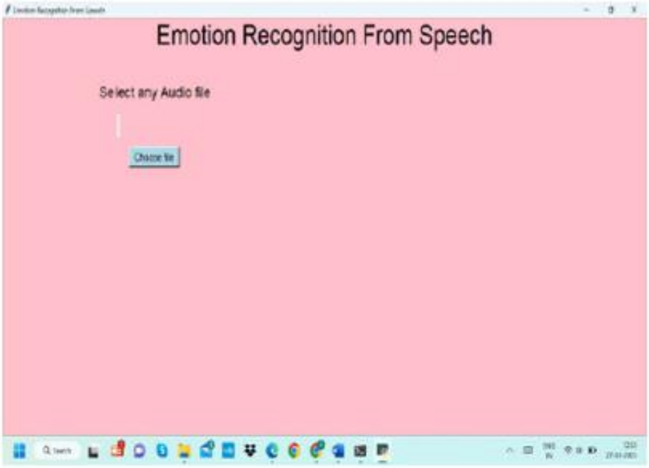


Figure3:Emotion Recognition from speech

2) Step 2: From this any one of the audio can be selected and then click on open as shown in figure 4.

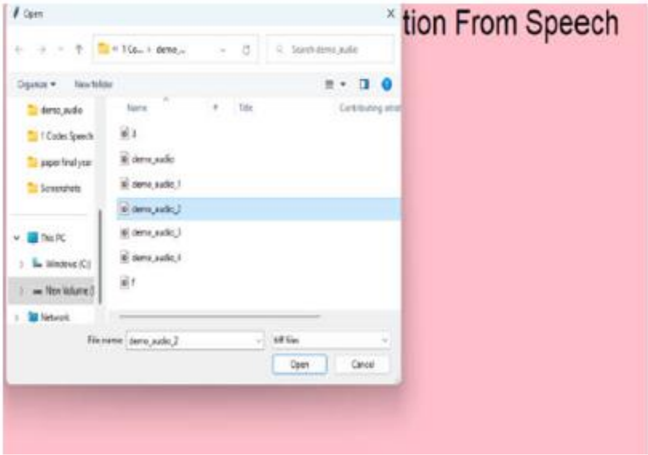
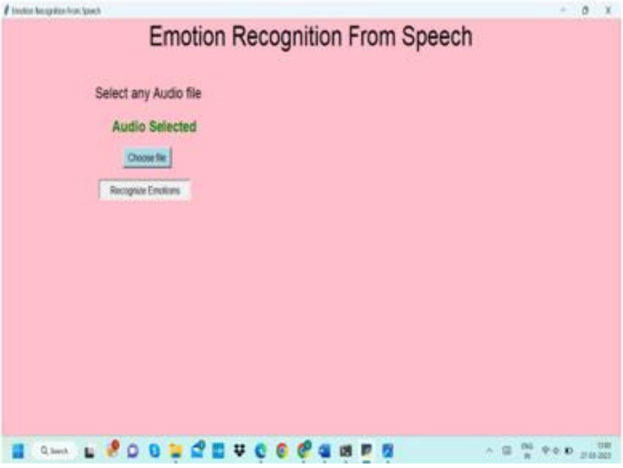


Figure4:Selection of audio file

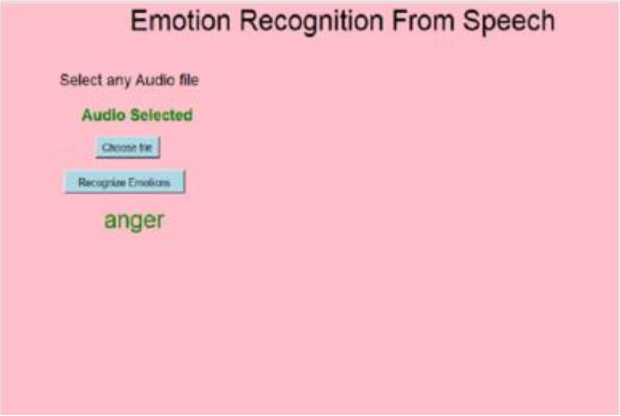
3) Step 3: The audio will be selected and then click on the recognize emotions (as shown in figure 5) then the audio will be recognized by using the CNN algorithm then the emotion will be displayed as text as shown in figure 6.


Emotion Recognition From Speech

Select any Audio file

Audio Selected

Figure5:Recognizing emotions


Emotion Recognition From Speech

Select any Audio file

Audio Selected

anger

Figure6:Final result display

VII. CONCLUSION

In this study, the development of an automated emotion recognition system using advanced machine learning techniques marks a significant stride in accurately discerning emotions from speech signals. With an impressive accuracy rate of 0.875, outperforming existing state-of-the-art systems, our research highlights the efficacy of our approach in enhancing emotion recognition capabilities. The implications of these findings extend beyond the confines of academic research, impacting critical domains like healthcare, education, and entertainment. By providing a robust framework for emotion detection, our system stands to revolutionize how emotional states are understood and addressed in various applications. This breakthrough opens doors to more personalized and effective interventions, improving overall user experiences and advancing the fields of human-computer interaction and emotional intelligence. Such advancements pave the way for a future where technology becomes more attuned to human emotions, fostering deeper connections and enhancing overall well-being.

VIII. REFERENCES

[1]. H. L. Qin, Y. T. Chen, Z. S. Zhang, and L. C. Jiao, “Emotion recognition from speech signals based on deep belief network,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 3984–3988.

[2]. T. Vogt, “Why emotions are hard to recognize in speech: Sorting the pros from the cons,” in Emotion in HCI: Joint Proceedings of the 2015 ACM International Conference on Affective Computing and the 2015 ACM International Conference on Automotive User Interfaces, 2015, pp. 57–63.

[3]. S. S. Dlay, A. Al-Taher, and J. F. Kaiser, “Emotion recognition using machine learning techniques: A review,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5690–5694.

[4]. Y. Wang, Z. Guo, W. Zhang, and M. Huang, “Emotion recognition from speech signals using hybrid feature selection and classification techniques,” Journal of Signal Processing Systems, vol. 91, no. 3, pp. 273–283, 2018.

[5]. L. Deng, D. Yu, and Y. Gong, “Deep learning: Methods and applications,” Foundations and Trends in Signal Processing, vol. 7, no. 3–4, pp. 197–387, 2014.

[6]. Al-Taher and S. S. Dlay, “Emotion recognition from speech using machine learning techniques: A review,” IEEE Access, vol. 6, pp. 78737–78750, 2018.

[7]. F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, 2016.

[8]. J. Deng, Y. Li, and X. Zhang, “A survey on emotion recognition from speech,” Journal of Ambient Intelligence and Humanized Computing, vol. 9, no. 1, pp. 165–179, 2018.

[9]. M. P. Sharma, P. Singh, and A. Bansal, “Speech emotion recognition using deep learning,” in 2017 2nd International Conference on Computational Intelligence and Networks (CINE), 2017, pp. 84-89.

[10]. B. W. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine belief network architecture,” in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2004, vol. 1, pp. I–I.

[11]. K. Saeki, M. Kato and T. Kosaka, “Language model adaptation for emotional speech recognition using Tweetdata,” Proc. of APSIPA ASC 2020, pp. 371-375, 2020.

[12]. E. Takeishi, T. Nose, Y. Chiba, and A. Ito, “Construction and analysis of phonetically and prosodically balanced emotional speech database,” Proc. of O-COCOSDA2016, pp. 16–21, 2016.

[13]. A. Ito and M. Kohda, “Evaluation of task adaptation using N-gram count mixture,” IEICE Trans. vol.J83-D-II, no.11, pp. 2418-2427, 2020 (in Japanese).

[14]. Y. Haneda, M. Sakurai, M. Kato and T. Kosaka, “Emotion recognition by fusion of time series features andstatistics of speech,” Proc. of ASJ meeting (Autumn), pp. 783-786, 2020 (in Japanese).

[15]. Florian Eyben et al., “openSMILE-book,” https://www.audeering.com/research-and-open sorce/files/openSMILE-booklatest.pdf