



EMOTION DETECTION BASED ON SPEECH USING MACHINE LEARNING

G. Saraswathi^{#1}, M. Lakshmi Sai^{*2}, D.V.S. Praneeth^{#3}, G. Yamini^{*4}, L. Anuja^{#5}

Information Technology

Assistant Professor^{#1}, Students^{*(2,3,4,5)}

Abstract:

Developers use human-system interactions frequently these days. Speech is a common technology that is used for interacting with people. It offers a channel to convey one's thoughts or emotional condition to others. However, a very important task in interactions between people and technology is their failure to recognize emotions in speech. Speech Emotion Recognition (SER) analyses the speaker's emotions through their speech module. Any machine with a small amount of processing capacity, when necessary, can identify common feelings including joy, sorrow, rage, fear, disbelief, and neutrality. The system develops a special challenge to take input from users and presents the emotional state using dataset like RAVDESS, TESS, and Emotional-DB for training as well as testing the efficiency of the device.

Keywords: Speech Emotion Recognition, Data Set, Machine Learning, Neural Network.

I. Introduction

The primary interactional medium in the modern period is speech. When more than one people communicate with each other, they can quickly tell how each other is feeling by looking at each other's faces or using words. Human-machine interaction is widely used in research nowadays. Lack of ability to interpret human emotions through speech is the biggest barrier to human-machine communication.

The important objective of emotion recognition includes determine how individuals perceive or respond to given conditions. Understanding the

speaker's mood is also helpful. It provides applications in huge amount, particularly in call centre applications, robotics engineering, medical science, and applications.

Hence, it is very much important to create an algorithm that is capable of precisely recognizing the emotions expressed in speech. In this field, various parameters for detecting a speaker's emotional state through speech have been discovered. Happy, sad, angry, disgusted, fearful, and neutral emotions can all be identified with one another.

Over the years, similar software has developed, and each uses a different set of features. Artificial networks are frequently employed as voice signals. According to our research, using Support Vector Machines (SVM) and Convolution Neural Networks (CNN) provides an edge during the recognition phase. The ability of machines in identifying people emotional nature is becoming increasingly relevant and enhances interactions between people and machines.

II. Related work

Kumar et al. [1] explores the application of audio communications to determine emotions in an intelligent assistant system. The system was designed to regulate electrical devices for alert actions and uses a multilayer neural network for voice emotion identification. It is about helping people in a variety of contexts, including homes, hospitals, and remote areas. Seven emotions are detectable by the suggested system: anxiety, surprise, neutral, sorrow, happiness, rage, and love. The system's development, testing and training on comparison datasets, and evaluation according to metrics like time, accuracy, and error rate are all covered in this work. When compared to current technology, the suggested approach exhibits positive results.

Plaza et al. [2] offers an innovative approach for call centre and contact centre emotion recognition. By accurately identifying the emotional states of clients as well as agents during talks, the technique aims to improve the efficacy of virtual assistants. The suggested strategy may identify emotions in speech and text channels, providing opportunities for developing behavioural profiles that enhance client satisfaction and agent productivity. This research explores the application

of programmed transcription of recordings to assess voice channel emotions. The suggested approach is appropriate for real-world use in contact centre and call centre systems as the experimental findings show that emotional states may be successfully classified.

Yan et al. [3] suggests applying the AA-CBGRU network model for identifying feelings in audio. The model utilizes a bidirectional gated recurrent unit (BGRU) network with an attention layer to gather deep time series information, spectrogram features, and spatial data using a convolutional neural network with residual blocks. Using the IEMOCAP sentiment corpus, the model's accuracy rises.

Zhang et al. [4] explains the increasing curiosity in multi-modal emotion detection and the important role of recognizing feelings in human interaction. The authors recommend an approach to increase the accuracy of emotion identification utilizing text, video, and audio modalities. After preprocessing, they extract deep emotional features from the data and integrate the data at the feature level. The model's findings on the IEMOCAP dataset are discussed in the release, exhibiting increased accuracy over speech emotion identification alone.

Han et al. [5] recommends using a deep residual shrinkage network with a bidirectional gated recurrent unit (DRSN-Bi-GRU) to identify speech emotions. The approach makes use of the Mel-spectrogram, an attribute for speech that has information in both the historical and frequency worlds. A convolution network, residual shrinkage network, bi-directional recurrent unit, and fully-connected network are all included in the DRSN-Bi-GRU model. To improve feature learning and

screen out distracting information, the self-attention mechanism is used. The approach beats existing models with accuracy rates of 86.03%, 86.07%, and 70.57% on three emotional datasets (CASIA, IEMOCAP, and MELD).

Wani et al. [6] gives an in-depth examination of systems for Speech Emotion Recognition (SER). The design components and methodologies of SER systems, including databases, preprocessing, feature extraction, and classification techniques, are addressed. Along with highlighting the research invalid in the subject, the study additionally tackles the difficulties encountered in SER.

Yang et al. [7] explains the evolution of the discrete emotion model-based spoken emotion recognition research. It provides an overview of speech emotion feature parameters and frequently utilized emotion databases. The study offers a description of the emotion recognition and methods for extraction of features employed in current Chinese research. It also discusses the challenges in recognizing emotions in speech and the directions that future research and growth might proceed.

Barhoumi et al. [8] demonstrates a real-time voice emotion identification system developed via data augmentation and deep learning methods. The goal is to use the tone of the voice alone to identify emotions. The system utilizes the use of three separate datasets and a variety of feature selection techniques, including chroma, Root Mean Square Value (RMS), Mel spectrograms, Zero Crossing Rate (ZCR), and Mel Frequency Cepstral Coefficients (MFCC). Emotion recognition can be accomplished by three distinct deep learning models: Convolutional

Neural Network (CNN), Multi-Layer Perceptron (MLP), and a hybrid model incorporating CNN with Bidirectional Long-Short Term Memory (Bi-LSTM). By evaluating the suggested system's efficacy in real-time scenarios, the CNN + Bi-LSTM model appears to be stronger.

Uthayashangar et al. [9] draws attention to speech emotion recognition (SER) and its potential applications in a number of fields. The study uses Mel Frequency Cepstral Coefficients (MFCCs) to extract attributes from voice data and convolutional neural networks (CNNs) to characterize emotions. Preprocessing speech data, feature selection, and background noise reduction are all part of the recommended approach. Using data augmentation techniques increases the model's dependability. The CNN algorithm is utilized for classification because of its adaptability and history of success with classification problems. When compared to earlier methods, the findings demonstrate that the suggested method achieves great precision in speech emotion identification.

Olatinwo et al. [10] proposes using the Internet of Things to develop a WBAN (Wireless Body Area Network) system that is emotion-aware and capable to grasp patients' expressed emotions. The technology uses a combination of machine learning algorithms and IoT sensors to assess and forecast patients' moods based on their speech. The writers look at several methods for extracting features, techniques to normalization, and algorithms for deep learning and machine learning. Additionally, they create a regularized CNN model and a hybrid deep learning model to lower computational complexity and boost prediction accuracy. The accuracy of the suggested models is around 98% when compared to an existing model.

Iliev et al. [11] investigates the application of deep learning techniques in artificial intelligence to determine emotions through speech. It discusses how essential emotions are in human communication and how difficult it may be to separate emotions clearly from used signals. The chapter looks at and compares the performance of several deep learning and machine learning classifiers used in emotion detection. The limitations of these approaches are also covered, as well as how important emotions are for interactions, making decisions, and overall well-being.

Pucci et al. [12] presents a chatbot virtual assistant system that makes use of machine learning neural networks to identify emotions with the aim to simplify contact tracking during the COVID-19 epidemic. Utilizing a transfer learning strategy, the system trained on an Italian language dataset with identifiers and acquired a 92% testing accuracy. In contact tracing conversations, the importance of recognizing emotions is highlighted since it may be used to identify stress, psychiatric disorders, and possible frauds. The study issue of emotions in contact tracing conversations, the new data set offered by Blu Pantheon, and the use of transfer learning are among the different novelties presented in this work.

Saini et al. [13] investigates the implementation of machine learning methods to voice recognition of emotions. The authors examine two separate datasets including samples of text and speech data to evaluate the efficiency of three different machine learning techniques: multinomial Naive Bayes (MNB), logistic regression (LR), and linear support vector machine (LSVM). The results indicate that LSVM performs

more effectively than either of the two methods. The study highlights how important emotion detection is to improving decision-making across a number of industries.

Koppula et al. [14] describes a unique hybrid firefly-based recurrent neural network approach to speech emotion recognition (SER). Preprocessing and feature analysis algorithms have been incorporated into the system to allow the classification of human emotions through speech input. When compared to other methods currently in usage, the model established efficacy, resilience, and high accuracy. The suggested approach could find applications in a number of fields, including security and medical.

Tambat et al. [15] focuses on building a speech-based emotion prediction system using CNN classifiers. The researchers use the Mel-frequency cepstral (MFCC) as an extracted spectral characteristic and evaluate the effectiveness of their method using the Database of Emotional Speech and Song (RAVDESS). The results show that using CNN classifiers yields good performance in recognizing emotions.

Jayanthi et al. [16] suggests an extensive framework that combines speech and still photographs of faces to identify emotions. With regard to individual recognition methodologies, the framework demonstrates outstanding precision because of its use of deep classifier fusion. The purpose of this tool is to identify a person's mental condition and offer auto-suggestions for enhancing their mental health.

Cai et al. [17] develops a multimodal emotion detection model that increases the performance of the emotional recognition system through the integration of audio and text data. The

model acquires textual and audio information using long short-term memory (LSTM) networks and convolutional neural networks (CNN), respectively. The fusion attributes are subsequently generated and categorized by a deep neural network. The suggested approach performs more effectively than single-modal models with greater precision in text and speech emotion identification, according to evaluations undertaken on the IEMOCAP database.

Abbaschian et al. [18] examines and contrasts deep learning methods for speech emotion recognition (SER) with traditional machine learning methods. The goal of the study is to offer a general description of the issue of discrete speech emotion identification by looking at neural network techniques, datasets, and current methodologies. The importance of SER in human-computer interaction is addressed along with its numerous applications in online courses, online therapy sessions, smart speakers, virtual assistants, and automobile safety systems. The different kinds of training datasets used for SER and the traditional methods employed at the time of the development of deep learning are also addressed in the research. A summary of possible future SER research directions came to a close.

Liu et al. [19] suggests an algorithm for understanding speech emotions in a small number of cases circumstance. The irrelevant traits and unstable data in emotion recognition are tackled by the model. To minimize the impacts of sample imbalance, it provides the Selective Interpolation Synthetic Minority Over-Sampling Technique (SISMOTE), a data imbalance processing strategy. Additionally, redundant features are eliminated using a feature selection technique called gradient

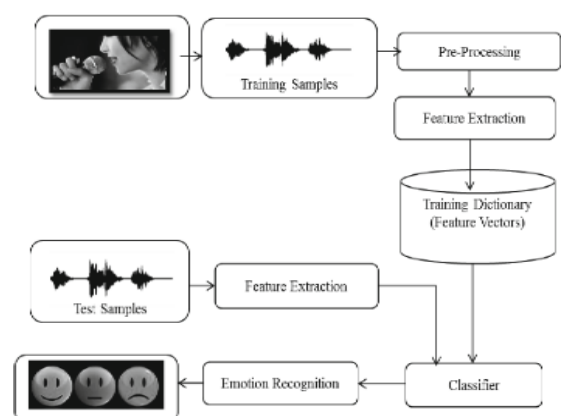
boosting decision tree (GBDT), which is based on variance analysis. The suggested technique improves state-of-the-art approaches in terms of recognition accuracy, based on experimental findings on three databases.

Issa et al. [20] offers an innovative framework for the identification of emotions in speech. The spectrum contrast features, ranging chromagram, Mel-scale spectrogram, Tonnetz representation, and Mel-frequency cepstral coefficients are just a few of the features that the authors extract from sound files. To identify emotions, a one-dimensional Convolutional Neural Network (CNN) uses these properties as inputs. The suggested approaches exceed current frameworks and achieve excellent classification accuracy, providing a new standard for emotion identification.

III. Methodology

Methodology:

The suggested system relies on emotion detection and uses some specified dataset for system training. Following by training, several preprocessing methods are used, and feature extraction techniques is then carried out. This dataset is utilized by the proposed method to



classify the emotions into different categories. CNN and Support Vector Machine (SVM) are two classification methods used by the

system. Training data is utilized for classification.

1.Data Collection: To train the emotion recognition system, the researchers gather audio recordings from 24 people in the RAVDASS speech dataset.

2.Feature Extraction: Mel-frequency cepstral coefficients (MFCC), Chromogram, Mel scaled spectrogram, Spectral contrast, and Tonal Centroid were some of the Acoustic characteristics which

are obtained from speech data. These features capture multiple speech signal characteristics that are crucial for emotion recognition.

3.Deep Neural Network Model: It is a classification model for emotion recognition. It also uses other models like SVM and Control Neural Network (CNN).

4.Training and Evaluation: The researchers trained the speech DNN using the resultant characteristics and collected audio recordings.

IV. Approaches

S. No	Methods	Parameters	Challenges
1	MFCC, LPCC, DELTA, FFT, PLP.	Accuracy, Error rate, Time.	-----
2	Data balancing techniques, Vectorization methods, Word embedding techniques, Dedicated transcription method, Speech signal descriptors	-----	Lack of a dedicated emotion recognition method, Unavailability of methods considering the audio signal parameters.
3	Model - AA-CBGRU Network Model. Methods - Data input, Spatial feature collection, Time series feature collection, and Classification.	-----	Gradient disappearance, Poor learning ability of time series information.
4	Deep Learning Techniques for Multi-Modal Emotion Recognition.	IEMOCAP Database which contains audiovisual data from 10 actors.	Complexity and Diversity of Human Emotion Expressions, Achieving Robust and accurate Emotion Detection.
5	Mel-spectrogram Feature Extraction, Deep Residual Shrinkage Network (DRSN), Bidirectional Gated Recurrent Unit (Bi-GRU).	-----	Finding Robust and Universal Emotional Features of Speech, Constructing Models with High Recognition Accuracy.
6	Interdisciplinary Knowledge from Fields such as Speech Emotion	Significance of Acoustic Features, Potential Integration	Availability of labelled data for Training Supervised Learning

	Recognition, Applied Psychology, Human-Computer Interface.	of linguistic, Facial, and Speech Information in Emotion Recognition.	Systems, Need for improving the accuracy of Feature Extraction Techniques.
7	Selecting the Feature Subset from Existing Features, Using Neural Networks to Extract new Features.	Acoustic Emotional Features, Semantic Emotional Features.	Lack of Acknowledged Speech Emotion Features, Difficulty of Converting Qualitative Emotional States into Quantitative Spatial Coordinates.
8	Data Augmentation Techniques, Feature Extraction Algorithms such as MFCC, ZCR, Mel spectrograms, RMS, and Chroma.	-----	Highlights the Complexity of Emotion Recognition, Need for Effective Feature Extraction and Classification Methods.
9	Voice data Preprocessing Techniques, Data augmentation methods, CNN algorithm as the classification approach.	Kaggle open-source RAVDESS dataset for training and testing.	Lack of data and Poor Model Accuracy in SER Research, Limited availability of labelled whisper voice data.
10	Machine Learning and Deep Learning Algorithms, Different Optimization Strategies and Regularization Techniques, Normalization Techniques are Explored.	Accuracy, Precision, Recall, F1 Score, Confusion Matrix.	Low Prediction Accuracy, High Computational Complexity, Delay in Real-time Prediction.
11	Process of Image Preprocessing, Face detection, Facial Landmark Detection, Feature Vector Creation.	Accuracy, Scores.	Human Error in Identifying Emotions Solely through Speech Signals.
12	Neural networks.	Labelled Italian-language Dataset (EMOVO Corpus).	Unavailability of labelled Emotions in the Dataset provided by Blu Pantheon.
13	Multinomial Naive Bayes (MNB), Logistic Regression (LR), Linear Support Vector Machine (LSVM), Artificial Neural Networks (ANN), Gaussian Mixture Models, K-Nearest	Accuracy, Precision, Recall, F1-score.	Analysis of Expressions in Long-Distance Communication, Identifying the most effective method for Speech Emotion Recognition.

	Neighbour, Hidden Markov Models (HMM).		
14	Hybrid Firefly-based Recurrent Neural Network (FbRNSR).	Features Extracted from the Speech Signal, Firefly Fitness for Optimization.	Filtering Noise Content, Extracting Emotional Features, Complexity and Cost Associated with Incorporating Digital Filters.
15	MFCC as a spectral characteristic for emotion identification, Feature selection.	Librosa module in Python, Samples from the RAVDESS Database.	Difficulty in differentiating between various emotions in spoken words, Limitations of the system in handling multiple speakers simultaneously
16	Convolutional Neural Network (CNN).	Spectral Features, Pitch Features, Energy Features, Intensity, Rate of Spoken Words.	Difficulty in Annotating Audio Recordings with Associated Emotions, Collecting unbiased audio data.
17	Combination of CNN and LSTM, Fusion of Features, CNN-Bi-LSTM-Attention (CBLA) Model,L2 Regularization.	-----	Limited Emotion Information in Single Mode, Traditional Feature Extraction Methods, Modelling Acoustic and Textual Features.
18	Autoencoders, Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), Long Short-term Memory (LSTM) networks.	-----	Complexity of implementing a deterministic system based on Pleasure, Arousal, and dominance measures and the variability of prosody, voice quality, and spectral features across different emotions and speakers.

19	Selective Interpolation Synthetic Minority Over-Sampling Technique (SISMOTE), Variance Analysis, Gradient Boosting Decision Tree (GBDT).	-----	Data Imbalance is common problem in emotional Corpora, Excessive High-level Emotional Feature sets often have Redundant Features.
20	Convolutional Neural Networks.	-----	Uncertainty in choosing the right features, Presence of background noise in audio recordings.

IV. Input and Output

Input: Takes the required number of datasets for training of the system and apply some of the techniques for data preprocessing and feature extraction. And also takes the audio sample input from the user.

Output: Takes the input from the user and classify them into the category in which the emotion is identified.

V. Findings and Trends

An essential component for recognizing emotions is speech. Emotion Recognition, Speech Emotion Recognition, Dataset and Accuracy, Deep Neural Networks, and Feature Comparison are some of the trends in Emotion Recognition.

Emotion Recognition: The primary objective of the study is to recognize emotions in speech. In recent years, there is a lot of interest in the field of emotion detection from speech signals. For a large number of applications, including human-machine interaction, psychological health, decision-making,

medical science, robotics engineering, and contact centre applications, it is seen as crucial.

Speech Emotion Recognition: Speech is a crucial element in understanding emotions. Some of the trends in emotion detection are Emotion detection, Speech Emotion Recognition, Dataset and Accuracy, Deep Neural Networks, Literature Survey, and Feature Comparison.

Dataset and Accuracy: The RAVDASS speech dataset, which consists of 1440 audio recordings from 24 different people, was used by the researchers to train the voice DNN. When compared to other algorithms like KNN, LDA, and SMO, the accuracy rate for emotion identification utilizing the DNN was observed to be 96%.

Deep Neural Networks: It was discovered that deep neural networks, when utilized simply for voice emotion detection, had a significant advantage in accurately identifying and classifying emotions from speech data. Deep neural networks are used in this example to show how they can

comprehend intricate styles and representations in speech notifications.

Feature Comparison: The comparison of MFCC, linear regression forecast cepstral variables (LPCC), and short time log frequency power coefficients (LFPC), three spectral features utilized for emotion identification, is briefly discussed in the text. According to the findings, LFPC was regarded as a better feature for emotion categorization than traditional features.

VI. Conclusion

Deep learning algorithms can produce fruitful outcomes. We successfully described a model for emotion recognition, and it scored 96% in testing. You should be aware that expecting feelings is arbitrary and that different listeners may give any piece of music different emotional values. The algorithm occasionally generates inconsistent results when trained on human-rated emotions for the same reason. The system was trained using datasets such as RAVDESS, which says mainly the speaker accent may result in unexpected results. As a result, it seeks to convey the speaker's emotional state more accurately through speech.

VII. References

[1] Kumar, Sandeep, Mohd Anul Haq, Arpit Jain, C. Andy Jason, Nageswara Rao Moparathi, Nitin Mittal, and Zamil S. Alzamil. "Multilayer Neural Network Based Speech Emotion Recognition for

Smart Assistance." *Computers, Materials & Continua* 75, no. 1 (2023).

[2] Płaza, Mirosław, Robert Kazała, Zbigniew Koruba, Marcin Kozłowski, Małgorzata Lucińska, Kamil Sitek, and Jarosław Spyrcza. "Emotion Recognition Method for Call/contact Centre Systems." *Applied Sciences* 12, no. 21 (2022): 10951.

[3] Yan, Yu, and Xizhong Shen. "Research on speech emotion recognition based on AA-CBGRU network." *Electronics* 11, no. 9 (2022): 1409.

[4] Zhang, Xue, Ming-Jiang Wang, and Xing-Da Guo. "Multi-modal emotion recognition based on deep learning in speech, video and text." In *2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP)*, pp. 328-333. IEEE, 2020.

[5] Han, Tian, Zhu Zhang, Mingyuan Ren, Changchun Dong, Xiaolin Jiang, and Quansheng Zhuang. "Speech Emotion Recognition Based on Deep Residual Shrinkage Network." *Electronics* 12, no. 11 (2023): 2512.

[6] Wani, Taiba Majid, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. "A comprehensive review of speech emotion recognition systems." *IEEE access* 9 (2021): 47795-47814.

[7] Yang, Chunfeng, Jiajia Lu, Qiang Wu, and Huiyu Chen. "Research progress of speech emotion recognition based on discrete emotion model." In *Journal of Physics: Conference Series*, vol. 2010, no. 1, p. 012110. IOP Publishing, 2021.

- [8] Barhoumi, Chawki, and Yassine Ben Ayed. "Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation." (2023).
- [9] Uthayashangar, S. "Speech Emotion Recognition Using Machine Learning." *Journal of Coastal Life Medicine* 11 (2023): 1564-1570.
- [10] Olatinwo, Damilola D., Adnan Abu-Mahfouz, Gerhard Hancke, and Hermanus Myburgh. "IoT-Enabled WBAN and Machine Learning for Speech Emotion Recognition in Patients." *Sensors* 23, no. 6 (2023): 2948.
- [11] Iliev, Alexander I. "Perspective Chapter: Emotion Detection Using Speech Analysis and Deep Learning." (2023).
- [12] Pucci, Francesco, Pasquale Fedele, and Giovanna Maria Dimitri. "Speech emotion recognition with artificial intelligence for contact tracing in the COVID-19 pandemic." *Cognitive Computation and Systems* 5, no. 1 (2023): 71-85.
- [13] Saini, Anu, Amit Ramesh Khaparde, Sunita Kumari, Salim Shamsher, Jeevanandam Joteeswaran, and Seifedine Kadry. "An investigation of machine learning techniques in speech emotion recognition." *Indonesian Journal of Electrical Engineering and Computer Science* 29, no. 2 (2023): 875-882.
- [14] Koppula, Neeraja, Koppula Srinivas Rao, Shaik Abdul Nabi, and Allam Balaram. "A novel optimized recurrent network-based automatic system for speech emotion identification." *Wireless Personal Communications* 128, no. 3 (2023): 2217-2243.
- [15] Tambat, Aditi Manoj, Ramkumar Solanki, and Pawan R. Bhaladhare. "Sentiment Analysis-Emotion Recognition." *Int. J. of Aquatic Science* 14, no. 1 (2023): 381-390.
- [16] Jayanthi, K., and S. Mohan. "An integrated framework for emotion recognition using speech and static images with deep classifier fusion approach." *International Journal of Information Technology* 14, no. 7 (2022): 3401-3411.
- [17] Cai, Linqin, Yaxin Hu, Jiangong Dong, and Sitong Zhou. "Audio-textual emotion recognition based on improved neural networks." *Mathematical Problems in Engineering* 2019 (2019): 1-9.
- [18] Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models." *Sensors* 21, no. 4 (2021): 1249.
- [19] Liu, Zhen-Tao, Bao-Han Wu, Dan-Yun Li, Peng Xiao, and Jun-Wei Mao. "Speech emotion recognition based on selective interpolation synthetic minority over-sampling technique in small sample environment." *Sensors* 20, no. 8 (2020): 2297.
- [20] Issa, Dias, M. Fatih Demirci, and Adnan Yazici. "Speech emotion recognition with deep convolutional neural networks." *Biomedical Signal Processing and Control* 59 (2020): 101894.