WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Emotion Recognition System

Illa Jaya Durgesh Kumar[1], Korupolu Uday Kumar[2], Bathina Satish Kumar[3], A. Phani Sridhar[4]

[1,2,3] U.G Student, Department of Computer Science and Engineering, Aditya Engineering College, Surampalem , A.P., India
[4]Associate Professor, Department of Computer Science and Engineering, Aditya Engineering College, Surampalem

ABSTRACT—

A fundamental but challenging part of human-computer interaction is deciphering emotions from voice inputs. (HCI). The literature on voice emotion recognition use a variety of techniques, such as numerous well-known speech processing and classification models, to extract emotions from signals. Recently, deep learning approaches are proposed as an alternative to conventional SER practices. This research paper presents a brief introduction of deep learning aa a subset of machine learning approaches and highlights some recent work that applies these approaches to recognize speech-based emotions. The evaluation discusses the databases utilized, the emotions extracted, the advancements in speech emotion identification, and any relevant restrictions. INDEX TERMS Convolutional neural networks, deep belief networks, deep networks, speech emotion recognition, and deep networks

Keywords— Librosa, Deep Learning, Voice Classification

Introduction

From being specialized topic, emotion identification from speech has become a fundamental component of HCI (HCI). Not other traditional conventional devices as input to the verbal information and make it simple for human listeners to respond, these systems strive to promote natural communication with machines through direct speech interface. Dialogue systems for spoken languages are used in a variety of settings, including customer service discussions, on-board auto driving systems, and the therapeutic use of spoken emotion patterns.

HCI systems still have a lot of problems that need to be resolved, particularly as they transition from being tested in labs to being used in real applications. Thus, initial steps are used to effectively solve these problems and improve the emotion recognition by machines. Assessment of a person's emotional state is a special job that is used for all emotion recognition models. One of the most fundamental methods among the many models used to classify different emotions is the discrete emotional approach. It employs a variety of feelings, such as neutrality, boredom, contempt, surprise, fright, delight, and melancholy.

A three-dimensional continuous space containing elements like arousal, valence, and potency is another significant model that is employed.

The majority of the emotion recognition of speeches technique is composed of the feature extraction and feature classification phases. Several features, In the realm of voice processing, features like vocal traction factors, source-based excitation features, prosodic features have been developed. Feature classification in the second step is done with both linear and nonlinear classifiers.

The Maximum Likelihood Principle , Bayesian Networks, and SVM's are the most used linear classifiers for identifying emotions. The voice signal is randomly taken to be non stationary. As a result, it is believed that non-linear classifiers are beneficial for SER. The GMM and HMM are two examples of non linear classifiers for SER.

They are frequently used to categorize data produced from defining features. Energy-based traits such as melancholy For efficient emotion recognition from speech, power-spectrum dynamic parameters, melancholy frequency power coefficients, linear classifier parameters, and perceptive linear prediction power coefficients are widely used. Principal component analysis, decision trees, and K-Nearest Neighbor are some other classifiers that are used for classifying emotions. In recent years, deep learning has gained greater attention as a machine learning research area since it is thought to be developing. Speech Emotion Recognition for deep learning have a number of benefits over conventional techniques, including the ability to recognize intricate structure and features without relying on human feature extraction.

The foundation of deep neural networks is a feed forward structure with more than two hidden layers placed between the inputs and outputs. convolutional neural networks and Deep neural networks , two examples of feed forward architectures, perform well when analyzing images and videos. Recurrent architectures, on the other hand, are very useful for speech-based classification techniques like Natural Language Processing and SER, as well as Recurrent Neural Networks and Long Short-Term Memory. These models do have a lot of disadvantages besides being used for categorization.

For instance, CNNs have the benefit of learning features from high-dimensional input data, but they also need a lot of storage capacity because they also learn features from small oscillations and distortions. Similar to LSTM-based RNNs, LSTM-based RNNs are capable of managing changeable source data and modelling long ranged linear data.

The three main elements of emotion recognition systems based on digitized speech are signal filtering, extraction of features, and classification. Using strategies like segmentation and acoustic preprocessing, such as denoising, the signal's pertinent components are separated.

The relevant features that are present in the signal are found using feature extraction. The retrieved extracted features are then mapped to the relevant emotions via classifiers. Speech signal processing, feature extraction, and identification are covered in great depth in this section. Given that they relate to the subject, The differences between performed speech and spontaneous speech are also looked at. Noisy components are removed.

The second phase consists of the extraction of features and feature selection. Relevant features from the preprocessed speech signal are extracted, and the characteristics are then applied to choose. The temporal and frequencies evaluation of speech inputs often serves as a foundation for such feature extraction and selection. In the third stage, these features are classified using a range of classifiers, include GMM and HMM. Eventually, feature categorization is used to classify a range of emotions.

Typically, noise from collection taints the source data used for emotion recognition. These issues lead to less accurate feature extraction and categorization. This means that in order for emotion detection and recognition algorithms to function properly, the input data must be improved. The emotional discrimination is preserved throughout this preprocessing stage while the speaker and recording variance are removed.

Related Work

In the subject of Human Computer Interaction, automatic emotion recognition for speech is a popular study area with a wide range of possible applications . The five emotional states that speech emotion recognition algorithms speedy data verbal manifestations of are disgust, boredom, sorrow, neutrality, and happiness. Using speech samples from the frequency cepstrum coefficients , linear prediction coefficients ,Berlin emotional database, linear prediction coefficients, energy, pitch were computed . The Support Vector Machine classifier is used to identify various emotional states. When using only energy and pitch data, the system's classification accuracy is 66.02%; when using only LPCMCC characteristics, it is 70.7%; and when using both, it is 82.5%.

Convolutional neural networks (CNNs) of one and two layers were employed, but these designs failed to extract the best features from challenging voice inputs. By employing data augmentation techniques to extract seven significant feature sets from each syllable, our work created a novel SER architecture to get over this restriction. The recovered feature vector is fed into the 1D CNN with the help of the RAVDESS, EMO DB, and other databases in order to identify emotions. In this study, all of the audio files of typical emotions from the abovementioned datasets were employed in a cross-corpus SER model. In this experiment results given that our proposed Speech Emotion Recognition framework performed more effectively than previous SER frameworks.

Literature Review

"Emotion detection from audio-visual emotional large data using deep learning approach "

In this paper, Big Data for emotion of deep learning is proposed for an emotion recognition system. Speaking and watching video are part of the big data. A Mel spectrogram, that can be used as an image, is created in the proposed system by first processing a voice signal in the frequency domain. A convolutional neural network is then fed this Mel spectrogram . For video signals, the CNN is supplied with a sample of representative frames from a video segment. Two consecutive extreme learning machines are utilised to combine the outputs of the two CNNs.

"System for Emotion detection using Deep Learning "

People are turning their attention away from the real world and towards the realm in today's richer and more materialistic society. Human machine interaction systems have been developed in order to identify and treat people's emotions. The existing human-machine interaction systems frequently facilitate human-robot interaction in a line of sight propagation environment, although the majority of human to human and human to machine communications are non-LOS. We suggest an emotion communication system based on NLOS mode to overcome the limitations of the conventional human-machine interface system. To be more precise, we analyze these emotions at initial stage as a type of multimedia, comparable to sound and video.

"Speech emotion detection in emotional feedback for human-robot interaction"

Robots must be able to identify human emotions in order to communicate with humans and plan their activities autonomously. Nonverbal indicators including pitch, loudness, spectrum, and speech pace are effective emotional messengers for the majority of people. Within this approach, a machine could recognize emotions using such sound properties. It's conceivable that a speaker's voice features reveal significant information about their emotional state. This article evaluated six different types of classifiers to predict six essential universal feelings from nonverbal features of human speech.

Emotion identification from speaking activities using HMM combined with DBN.

The goal of research on emotion recognition is to gain understanding of the temporal characteristics of emotion difficult. Sometimes it becomes difficult to find the emotion it is due to lack of proper recording of speech along with this the data may contain similar type of recordings so it becomes difficult

to identify the speech. Furthermore, high-dimensional noisy data are frequently used by emotion identification systems, making it challenging to identify representative features and develop efficient classifiers. Using DBN, which can simulate intricate and non linear high level interactions between low level features, we attempt to solve this issue. A group of hybrid classifiers based on HMM are proposed and evaluated.

"Speech recognition using deep neural networks"

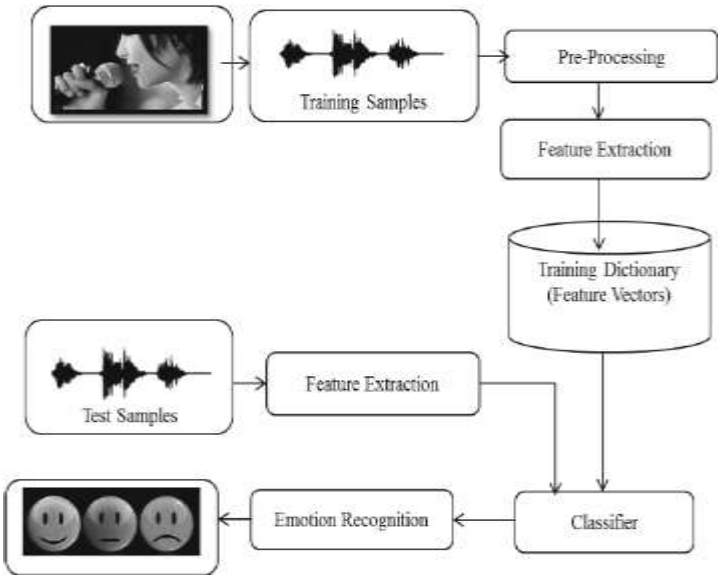
By using deep neural networks, In this existing system emotion detection can be done in an accurate way and also nowadays, research of particular thing will effect the whole project. Because the research can be useful for the study using any technology. Deep neural networks are used as best and suitable systems for the applications.

Methodology Used

An MLP classifier is used in the proposed system to build a voice emotion recognition model. We have proposed a machine learning-based revamp of the current online speech emotion recognition system. The suggested system has the following advantages over the current one.

Humans' mental health depends on their ability to feel emotions. It acts as a means of expressing one's views or mental state to others. Speech emotion recognition is the technique of interpreting the speaker's emotional state from the speech signal (SER). The few universal emotions—Neutral, Anger, Happy, and Sadness—can be taught to be recognized or synthesized by any intelligent system with minimum processing power. Both spectral and prosodic features are used in this study to determine speech emotions because they both contain the emotional information.

- To anticipate the emotion of a human speech, we apply the ANN algorithm in the suggested system.
- To deal with the audio files and turn them into arrays, we used a built-in Python package called librosa (Conv1D)
- The group of arrays is reduced to a single binary number using the ANN method
- Formerly, we had to add the accuracy and loss after passing the input from one layer to another with a weight named "rms prop" (learning rate=0.01).
- Initialize the Augmentation, which is used to grow the dataset size and add more files, if we desire greater precision.
- By using the fit command, we can add some epochs and begin training the dataset.
- We can get 100% accuracy for audio classification using the ANN algorithm.
- It is more accurate than the previous methods at predicting emotions.
- By using this algorithm we can accurately find the emotion of a particular person.
- Based on previous algorithms and systems we can get a reference to the system to use.



Features:

- To use the MLP classifier to enhance the current speech emotion recognition system.
- It makes use of a dataset that actors have meticulously recorded.

- Algorithm, which is used in this project is accurately predict the output.
- It provides the product in terms of emotional realism.

The major goal of this project is to research, categorize, and analyze works on speech emotion identification for a specific human speech.

It can also be used to monitor a person's psycho physiological status in lie detectors. It is also employed in the applications in medical and forensics. The main goal of SER is to improve man-machine interface.

Instead than using antiquated devices to understand verbal content and made it simple for listeners to respond, emotion recognition of speech systems strive to promote the traditional engagement with machines through direct voice interface.

Stages in an MLP Classifier, a sort of sequential model for a dataset:

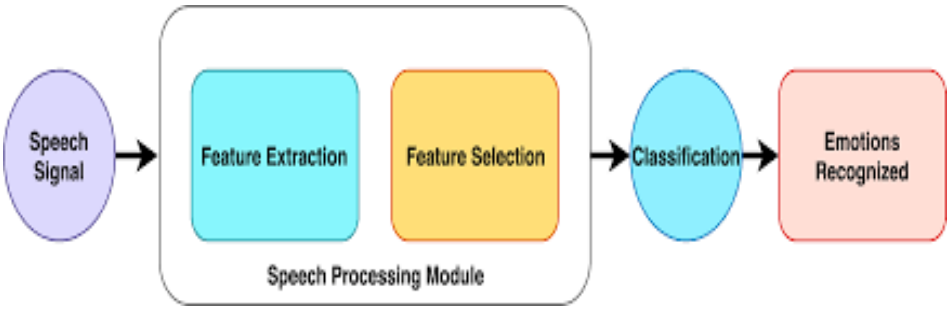
The approach used to accomplish this process of transmitting dataset information to MLP Classifier, a kind of sequential model.

1. Feature extraction:

This module uses pertinent information derived from the voice signal to classify distinct emotions. Examples of frequently used features include spectral features, prosodic features (such as pitch, energy, and duration), and mel frequency cepstral coefficients (MFCC).

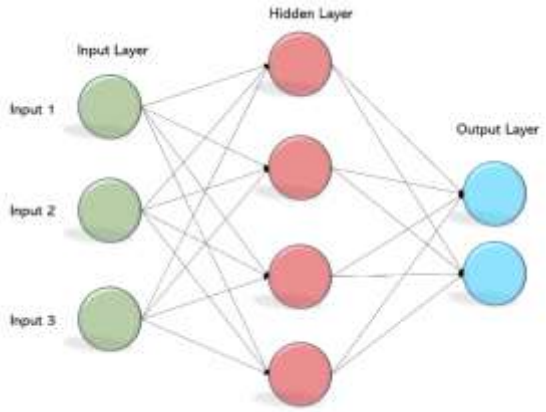
2. Pre-processing: Pre-processing may be used to reduce the impact of noise or normalise the values of the recovered characteristics.

3. Preparing the training data: For training, the ANN MLP classifier requires a large amount of tagged voice data. After being labelled with the target emotions and carefully chosen datasets in this module, the data is split into training and validation sets.

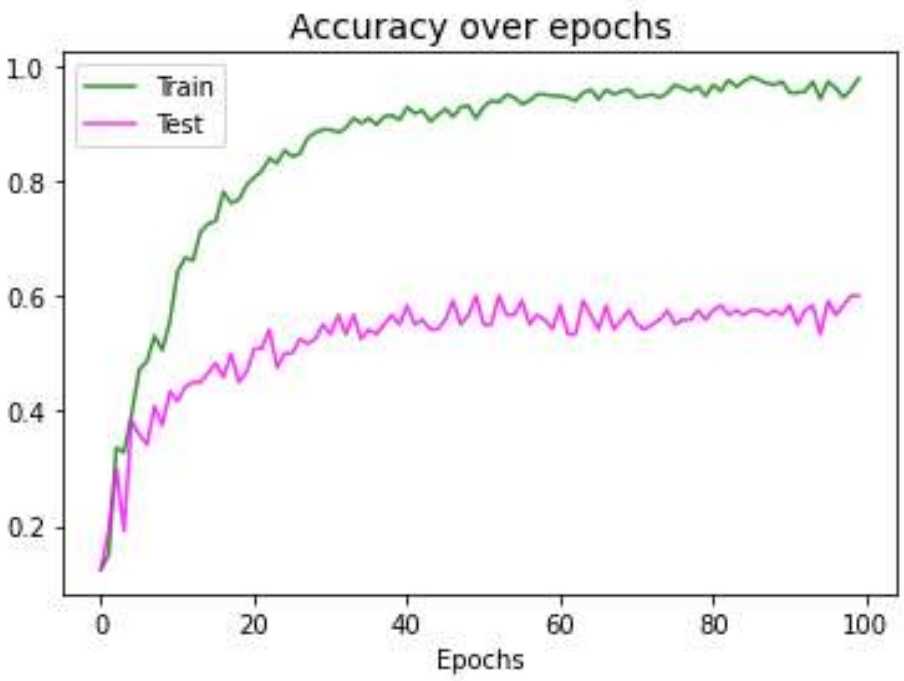


4. Design of neural network architecture: The ANN MLP classifier architecture is established by selecting the number of input and output nodes, hidden layers, and neurons in each layer. The network's learning rate and each neuron's activation capabilities are both selected.

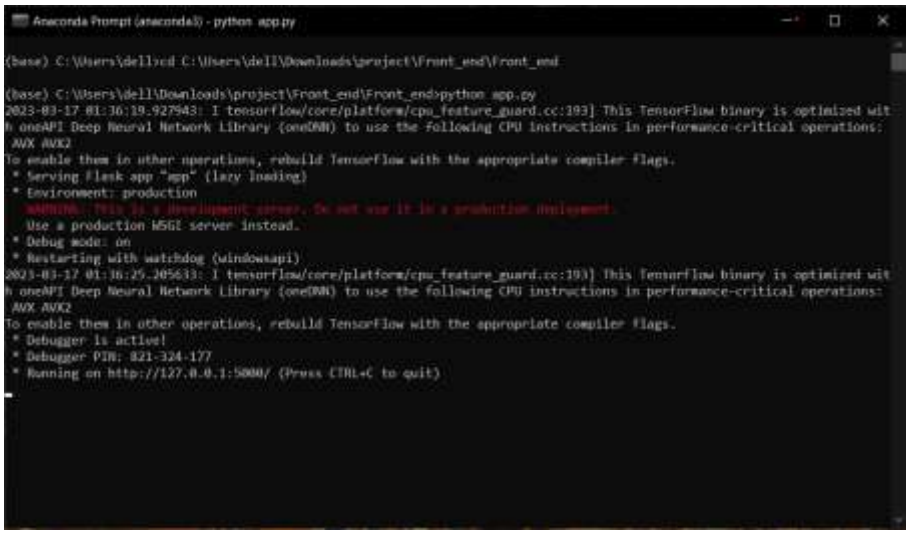
5. Training the ANN MLP classifier: involves feeding the neural network with the pre-processed features and modifying the network weights using a backpropagation technique. Training is conducted on the training set, and performance is monitored on the validation set.



6. Validation and testing: The ANN MLP classifier's performance can be evaluated by putting it to the test on brand-new, undiscovered data after it has been trained. Accuracy, precision, recall, and F1-score performance measures are calculated as part of this module.



7. Implementation: When the system's performance is enough, it can be used in a real-world environment. The system may be integrated with additional modules, such as a speech recognition module or a text-to-speech module, to create a complete speech-based application.



- Installed the Anaconda navigator and run the anaconda prompt. Navigate to the frontend path where we can run the project.
- By using “app.py” command we can start the API calls to the server.
- After running the server the front end will appear on local host 5000.
- The interface will look like above mentioned screenshots where we having a option called choose file.
- We can upload the file and click on upload&predict button to get the output.
- By clicking the upload&predict button we are getting the output at right side which emotion it was.
- Different emotions will predict like happy, sad, disgust etc., based on training set.

Conclusion and Future Work

In conclusion, a useful technique for detecting emotions in speech is the integration of an artificial neural network (ANN) MLP classifier. The method is made up of a number of parts, including feature extraction, preprocessing, preparing training data, designing the neural network architecture, training,

testing, and deployment. By picking out the right datasets, designing a neural network structure that functions effectively, and training the network with sufficient amounts of labelled data, the system is able to accurately identifying a wide range of speech emotions.

Yet, creating such systems is not without its challenges. One of the main challenges is the need for a sizeable amount of annotated speech data, which can be difficult and time-consuming to collect. Another challenge is the possibility of speech signal change caused by variances in accent, pronunciation, and speaking manner. These factors might affect how accurate the system is, necessitating the employment of additional techniques like domain adaptation or transfer learning.

Notwithstanding these challenges, the use of an ANN MLP classifier for speech emotion analysis shows a lot of promise and has the ability to be applied in a variety of real-world settings, such as human-computer interaction, speech therapy, and emotional speech analysis. Further research and development will be done to increase the precision and resilience of these systems, broadening the value and application of these systems.

Language-specific speech emotion recognition algorithms predominate in the market today. Cross-lingual emotion recognition systems that can recognise spoken emotions in several languages may be developed as a result of future research. The vast majority of speech emotion recognition systems currently in use are "black boxes," making it difficult to understand how the system selects what actions to take. Future systems might be developed that can defend their decisions, improving their dependability and openness. Because they are illogical, emotions might vary from person to person. In the future, systems for recognising emotions may be developed that are individual to users and their emotional expressions. Speech is one tool that can convey emotional information. Such emotion detection systems may one day be combined with additional modalities, including physiological signals or facial expressions, to improve the accuracy of emotion recognition.

References

[1] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” Communications of the ACM, vol. 61, no. 5, pp. 90–99, 2018.

[2] M. S. Hossain and G. Muhammad, “Emotion recognition using deep learning approach from audio–visual emotional big data,” Information Fusion, vol. 49, pp. 69–78, 2019.

[3] M. Chen, P. Zhou, and G. Fortino, “Emotion communication system,” IEEE Access, vol. 5, pp. 326–337, 2017.

[4] N. D. Lane and P. Georgiev, “Can deep learning revolutionize mobile sensing?” in Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications. ACM, 2015, pp. 117–122.

[5] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, “Speech emotion recognition in emotional feedback for human-robot interaction,” International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 4, no. 2, pp. 20–27, 2015.

[6] D. Le and E. M. Provost, “Emotion recognition from spontaneous speech using hidden markov models with deep belief networks,” in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013, pp. 216–221.

[7] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” IEEE Access, vol. 7, pp. 19 143–19 165, 2019.

[8] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, “Speech emotion recognition in emotional feedbackfor human-robot interaction,” International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 4, no. 2, pp. 20–27, 2015.

[9] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, “Speech emotion recognition,” in Advances in Electronics, Computers and Communications (ICAECC), 2014 International Conference on. IEEE, 2014, pp. 1–4.

[10] K. R. Scherer, “What are emotions? and how can they be measured?” Social science information, vol. 44, no. 4, pp. 695–729, 2005.

[11] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, “Emotion analysis in man-machine interaction systems,” in International Workshop on Machine Learning for Multimodal Interaction. Springer, 2004, pp. 318–328.

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction,” IEEE Signal processing magazine, vol. 18, no. 1, pp. 32–80, 2001.

[13] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, “Emotion recognition by speech signals,” in Eighth European Conference on Speech Communication and Technology, 2003.

[14] R. W. Picard, Affective computing. Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology, 1995.

[15] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” International journal of speech technology, vol. 15, no. 2, pp. 99–117, 2012. Nuts”, International Journal of Neural System 8(1):55-61

[16] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011. [17] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, “Emotion recognition by speech signals,” in Eighth European Conference on Speech Communication and Technology, 2003.

[18] A. D. Dileep and C. C. Sekhar, "Gmm-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," IEEE transactions on neural networks and learning systems, vol. 25, no. 8, pp. 1421–1432, 2014.

[19] L. Deng, D. Yu et al., "Deep learning: methods and applications," Foundations and Trends® in Signal Processing, vol. 7, no. 3-4, pp. 197– 387, 2014.

[20] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural networks, vol. 61, pp. 85–117, 2015.

[21] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE, 2005, pp. 474–477.

[22] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artificial Intelligence Review, vol. 43, no. 2, pp. 155–177, 2015.

[23] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in Emotion-Oriented Systems. Springer, 2011, pp. 71–99.

[24] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1057– 1070, 2011.

[25] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5005–5009.