Speech Emotion Recognition Using Multi-Layer Perceptron Classifier and Extra Tree Classifier

Project Report

AN CONG JIE
Master in Data Science and Business Analytics (M1 Student)
ESSEC Business School and CentraleSupélec
congjie.an@student-cs.fr

LIU DAYU
Master in Data Science and Business Analytics (M1 Student)
ESSEC Business School and CentraleSupélec
dayu.liu@student-cs.fr

Nicolo' Giacopelli
Master in Data Science and Business Analytics (M1 Student)
ESSEC Business School and CentraleSupélec
nicolo.giacopelli@student-cs.fr

Vanshika Sharma
Master in Data Science and Business Analytics (M1 Student)
ESSEC Business School and CentraleSupélec
vanshika.sharma@student-cs.fr

1. MOTIVATION AND PROBLEM DEFINITION

Humans have the ability to comprehend emotions of another human being via text, speech, visuals, etc. and accordingly respond to it. However, computers and machines cannot. In this rapidly advancing AI world, human computer interactions (HCI) are of extreme importance. The act of attempting to recognize human emotion and affective states from speech is known as Speech Emotion Recognition (SER). Speech is one of the main mediums to communicate emotions and attitudes in a particular language, and thus characteristics like intonation, pitch of voice and relative loudness of specific frequencies can be used as features for a machine-learning problem of ever-increasing importance. In fact, understanding human emotions paves the way to understanding people's needs better and, ultimately, providing better service: SER techniques could have wide applications in marketing, healthcare, customer satisfaction, stress monitoring, social media analysis, gaming experience improvement and much more. SER is in fact the same mechanisms that animals like dogs and horses employ to be able to understand human emotions. Designing a machine-learning algorithm brings about the same challenges as well as additional ones. Firstly, the necessity to apply Fourier transformations and Dimensionality Reduction techniques to the vocal files to make them intelligible; secondly, the problem of extracting the informative features by allowing the model to distinguish between the portion of the vocal content coming from the Vocal Tract, which is the conscious elaboration of spoken words, and the one coming from the Glottal Pulse, which is just the origin of the sound waves and that can be interpreted as noise.

KEYWORDS

Sound waves; Fourier transformations; Spectrograms; Emotion classification.

2. RELATED PRIOR WORK

In the designing a speech emotion recognition system, the most important task is to identify and extract different emotion related speech features. Since proper selection of the features affects the classification performance, it is critical to combine appropriate audio features in speech emotion recognition. Due to the importance of Speech Emotion Recognition in human-computer interaction and the development of artificial intelligence systems, there are multiple other recent publications and surveys on Speech Emotion Recognition. There were many approaches to recognizing emotion from speech, and each study used different speech features.

In this section, we review the most recent studies related to the current work.

In 2018, Swain et al. [7] reviewed studies between 2000 and 2017 on SER systems based on three perspectives: database, feature extraction, and classifiers. The research has an extensive section on databases and feature extraction; however, only traditional machine learning methods have been considered as classifying tool, and the authors are regretting neural networks and deep learning approaches.

In 2019, Khalil et al. [8] reviewed discrete approaches in SER using deep learning. Several deep learning approaches, including deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoder, have been mentioned along with some of their limitations and strengths in the study. However, the research is not addressing the accessible approaches to overcome weaknesses.

Very recently, Anjali et al. [9] published a review as a summary of speech emotion detection methods. A brief discussion of various features used in speech emotion recognition and review of methods used for this purpose from 2009 to 2018 has been provided in the research. The drawback of the paper is the depth of analysis. Yet, it can be considered a start point.

In 2020, Basu et al. [10] published a brief review on the importance of speech emotion datasets and features, noise reduction; ultimately, they analyze the significance of different classification approaches, including SVM and HMM. The strength of the research is the identification of several features related to speech emotion recognition; however, its weakness is the leak of more modern methods’ investigation and briefly mentions convolutional and recurrent neural networks as a deep learning method.

A few studies also extracted features from speech and recognized emotions utilizing Gaussian Mixture Models (GMMs) [11] and Hidden Markov Models (HMMs)[12]. Recently, with the rapid development of deep-learning algorithms, RNN have been applied to various fields of speech analysis [13-15]. The main focus of the approach used in ‘High-Level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition’ is to classify the emotions in an utterance, rather than classifying the emotions using the frames of the utterance. To extract global features, the authors fed 32-dimensional frame features directly into an RNN. Then, the global features were fed into an extreme learning machine to classify the emotions. This approach recognizes emotions from an utterance, rather than using the frames of the utterance, and therefore requires a lot of computational power to train the network.

In another approach, Wieman (Analyzing Vocal Patterns to Determine Emotion) [16] found that binary decision trees can determine the features most relevant to emotions. However, their experiment was performed using a small dataset.

The study ‘Recognizing Emotion from Speech Based on Age and Gender using Hierarchical Models’ [17] assumed that the characteristics of speech vary in each person, and that emotions are affected by the age, gender, and acoustic features of the speaker. The authors focused on speech emotion recognition by grouping speech data by age and gender. They proposed a hierarchical gender- and age-based model and utilized different feature vectors of OpenSmile [18] and eGeMAPS [19]. The results indicated that building a separate classifier for each gender and age group produces better performance than having one model for all genders and ages.

Many studies have demonstrated the correlation between emotional voices and acoustic features [20-25]. However, because there does not exist overt and deterministic mapping between features and emotional state, speech emotion recognition still has lower recognition rate than other emotion-recognition methods, such as FER. For this reason, finding the appropriate feature combination is a critical task in speech-based emotion recognition.

3. METHODOLOGY

In this project, we worked on a machine learning model based on Speech Emotion Recognition. The model predicts the emotion of the speaker by analyzing the recorded audio clip of the speaker’s voice.

Clearly, a speech contains three different sources of information, namely the lexical features (the words used), the acoustic features (the pitch and tone) and the visual ones (carried by facial expressions in the presence of a video). Our project focuses on using the acoustic features of the speech. For future development, one can consider the possibility of leveraging the lexical aspect of the data as well.

The following steps were taken during the project:

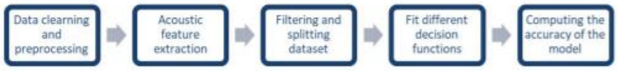

Data clearning and preprocessing
Acoustic feature extraction
Filtering and splitting dataset
Fit different decision functions
Computing the accuracy of the model

Figure 1: Project Pipeline

For the project, we created a dataset by merging four popular speech databases, that are the Crowd-Sourced Emotional Multimodal Actors Dataset (Crema-D), The Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess), Surrey Audio-Visual Expressed Emotion (Savee) Database and Toronto emotional speech set (Tess). These datasets classify the audio recordings according to seven recognizable emotions, namely anger, fear, disgust, sadness, happiness, calm and surprise.

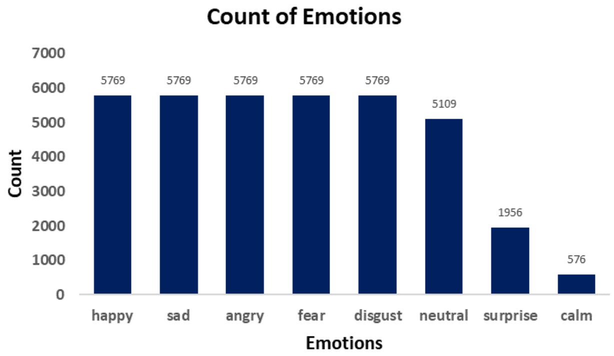
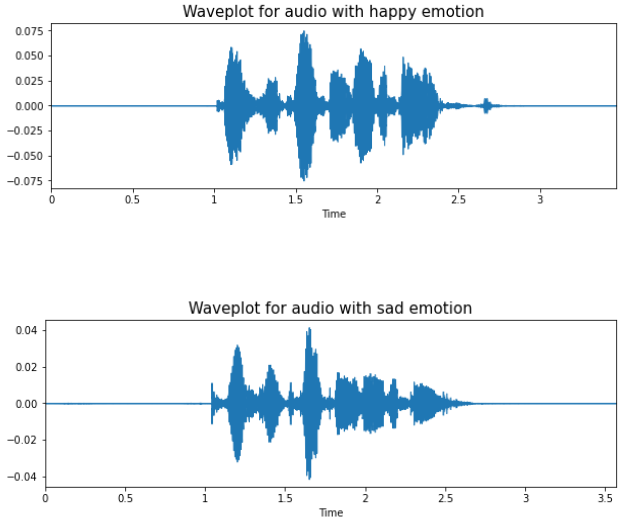


Figure 2: Count of Audios by Different Emotions in the combined dataset

Using the Python package Librosa, we created digital representation of the audio clips for better understanding and used it to extract features from a raw audio waveform. Here are the wave plots for the top 5 emotions from our combined dataset:



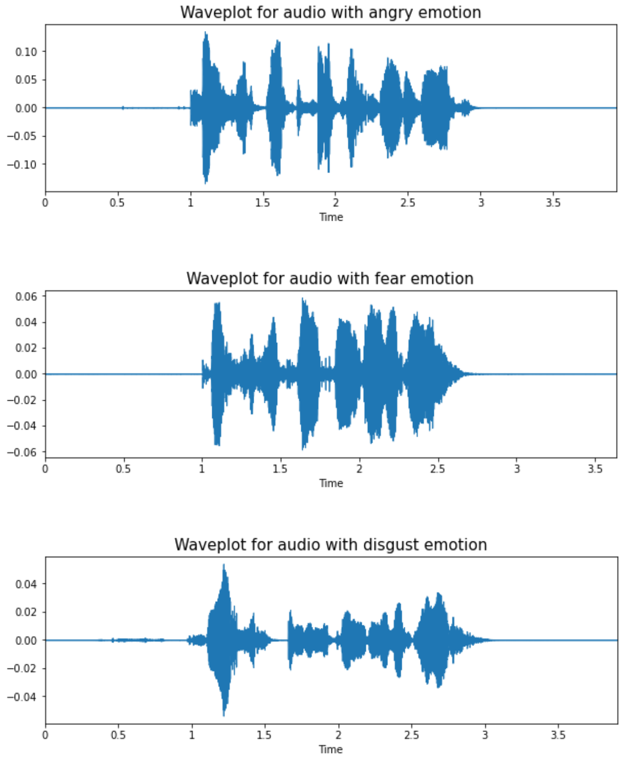


Figure 3: Wave plots of emotions – happy, sad, angry, fear, and disgust

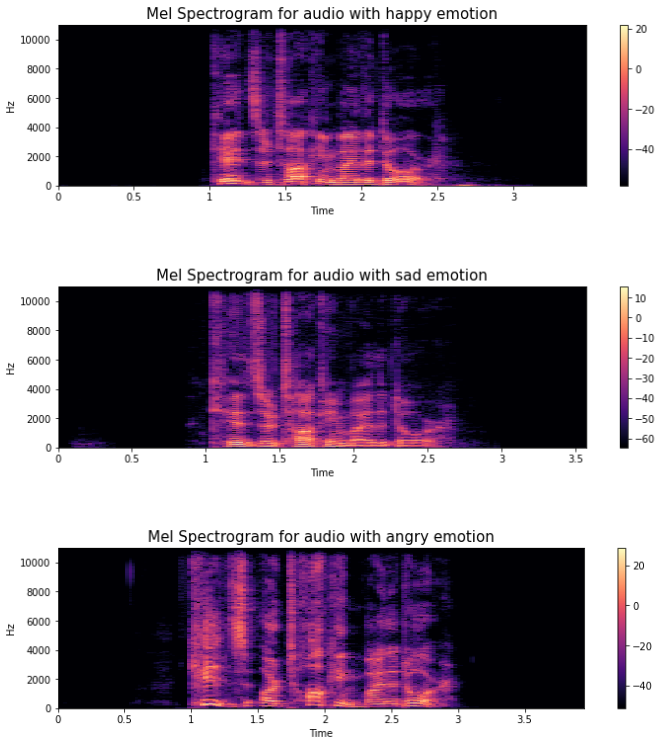
About Wave Plots: The above plots describes the change in amplitude (loudness) of a signal over time domain. The next challenge is extracting the significant features from this wave form that can easily help to distinguish emotions embedded.

There are numerous ways to extract features from a raw audio waveform using signal processing such as zero crossing rate, spectral centroid, zooming in, etc. After our experiments, we decided to move forward with a combination of three main acoustic features which will be discussed below in detail. In the case of audio waves, the variations in the amplitudes and frequencies contained in it can provide insights. A single audio wave consists of multiple single frequency signals. Extracting these individual, single frequency signals from the audio is called spectrum analysis.

About Fourier Transformation: Human brains are a kind of spectrum analyzer which automatically split up the frequency signals in the audio and help us focus more on the main component, ignoring the single frequency signals of background noise. This same feature of splitting up the single frequencies in an audio file can be obtained by a mathematical technique called Fourier Transform. Using this, one could basically transform a time-domain signal into a frequency-domain signal. With this, we have information on how amplitude and frequency vary with time (separately, not together).

About Spectogram: To get the frequency and amplitude variations together with time, Spectogram comes in handy. Spectogram carries a snapshot of the information regarding the relation between amplitude (i.e., decibels), frequency (i.e. the pitch), and time. Using three different versions of Spectograms, we carried out feature selection.

1. Mel Spectogram: It shows how the frequency and amplitude evolve over time. The x-axis represents time, the y-axis represents frequency, and color represents amplitude. The brighter the color, the higher the amplitude. This spectrogram uses a special scale for frequency called the Mel scale. Below are the images of the Mel spectrogram of an audio wave with our top 5 emotions from combined dataset.



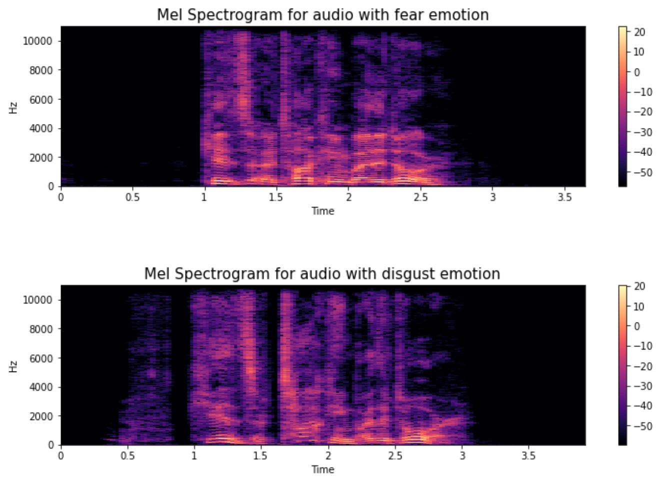


Figure 4: Mel Spectograms of emotions – happy, sad, angry, fear, and disgust

2. Mel-Frequency Cepstral Coefficients (MFCC): The mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10-20) which concisely describe the overall shape of a spectral envelope. This is one of the most common ways to perform feature extraction from audio waveforms. MFCC features are very efficient in preserving the overall shape of the audio waveform. Below is a sample image for MFCC:

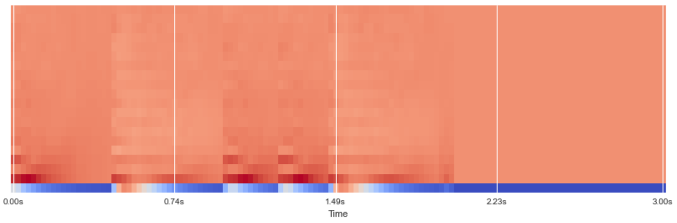


Figure 5: Sample image of Mel-Frequency Cepstral Coefficients (MFCC)

3. Chroma: Chroma is a specific kind of spectrogram based on the chromatic scale. It is a descriptor, which represents the tonal content of a musical audio signal in a condensed form. Therefore, chroma features can be considered as important prerequisite for high-level semantic analysis, like chord recognition or harmonic similarity estimation.

The chromatic scale or 12 tone-scale is a set of 12 pitches used in tonal music.

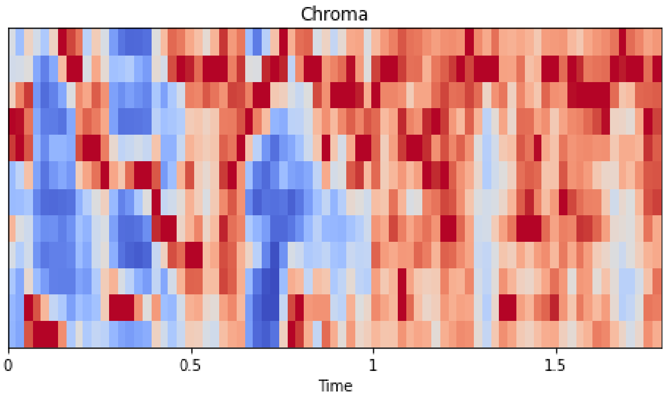


Figure 6: Chroma

These three different versions, namely Mel Spectrograms (a log-scale spectrogram), Mel-Frequency Cepstral Coefficients (MFCC) and Chroma, are all computed through different applications of the mathematical technique of the Fourier Transform and they allow to extract the most valuable information by differentiating easily between the Spectral envelope (the Vocal Tract) and the noise (the Glottal Pulse).

After performing the above steps, we split the dataset between training and test dataset. The final step was to build the learning model which takes spectrogram features of an audio file as input and predicts the emotion embedded in it. For the same, we trained the models using different techniques and carried out evaluation to identify which model is more suitable.

MODEL EVALUATION AND SELECTION

By performing the above steps, we basically converted our problem into a multi-class classification problem. Based on our experience with e-mail multi-class classification problem (i.e., kaggle competition), we used various models and below are the accuracy scores we achieved.

Modelling Algorithm	Accuracy Score
MLP	0.6352
Decision tree	0.5136
Extratree	0.6376
SVM	0.5875
RandomForest	0.6163

For the purpose of evaluation i.e., calculating the accuracy of the predictions, we performed tests using the dataset which has been extracted from the main database (and hasn't been used during the training). We also tested the model based on audio recordings from various tv-shows and edited the identified emotion in the video for the presentation.

Since we are dealing with a multiclass classification problem, we based our model evaluation on the different estimates computable from the Confusion matrix.

Multi-Layer Perceptron Classifier (MLP) - We chose MLP as one of the selected models since it gave us almost the best accuracy. However, using MLP generally requires more time in modeling and predictions, and our MLP model didn't perform very well for features extracted from sounds, even after the standardization and gridsearchCV process (which is used to find the best parameters for the model and caused lots of time).

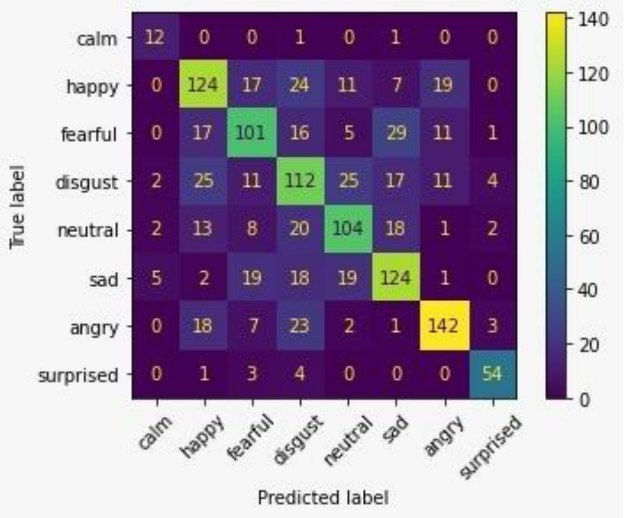
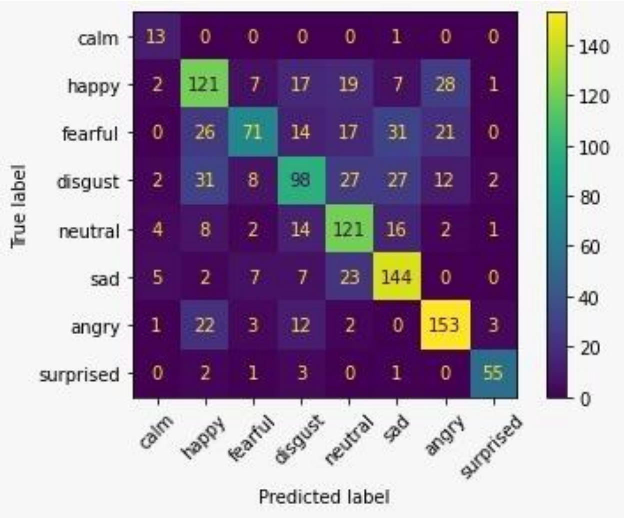


Figure 7: Confusion Matrix for MLP

Thus, we decided to choose Extra tree as the second selected model. It gives us a slightly higher accuracy than MLP. An extra-trees classifier implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.



	calm	happy	fearful	disgust	neutral	sad	angry	surprised
calm	13	0	0	0	0	1	0	0
happy	2	121	7	17	19	7	28	1
fearful	0	26	71	14	17	31	21	0
disgust	2	31	8	98	27	27	12	2
neutral	4	8	2	14	121	16	2	1
sad	5	2	7	7	23	144	0	0
angry	1	22	3	12	2	0	153	3
surprised	0	2	1	3	0	1	0	55

Figure 8: Confusion Matrix for ExtraTree

CONCLUSION

Decision Tree and SVM didn't give us satisfactory results, whereas RandomForest, MLP and ExtraTree Classifier gave us a better accuracy.

As a part of future development, we could try techniques for data augmentation. Data augmentation is the process by which we create new synthetic data samples by adding small perturbations on our initial training set. To generate syntactic data for audio, we can apply noise injection, shifting time, changing pitch and speed. These techniques may further increase accuracy of our prediction and avoid over-fitting issues.

Another possibility could be to leverage the lexical aspect of data as well.

DATA SET LINKS/ARTICLES

- The Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS) - https://zenodo.org/record/1188976#.YX0Ldp5BxZW
- Crowd-Sourced Emotional Multimodal Actors Dataset (Crema-D) - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313618/
- Surrey Audio-Visual Expressed Emotion (SAVEE) Database- http://kahlan.eps.surrey.ac.uk/savee/
- Toronto emotional speech set (TESS)- https://tspace.library.utoronto.ca/handle/1807/24487

REFERENCES

[1] El Ayadi, M., M. S. Kamel, and F. Karray (2011): "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern recognition, 44, 572–587.

[2] Hajarolasvadi, N. and H. Demirel (2019): "3D CNN-based speech emotion recognition using k-means clustering and spectrograms," Entropy, 21, 479.

[3] Pan, Y., P. Shen, and L. Shen (2012): "Speech emotion recognition using support vector machine," International Journal of Smart Home, 6, 101–108.

[4] Satt, A., S. Rozenberg, and R. Hoory (2017): "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms." in Interspeech, 1089–1093.

[5] Wang, K., N. An, B. N. Li, Y. Zhang, and L. Li (2015): "Speech emotion recognition using Fourier parameters," IEEE Transactions on affective computing, 6, 69–75.

[6] Zhao, J., X. Mao, and L. Chen (2019): "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical Signal Processing and Control, 47, 312–323.

[7] Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. Int. J. Speech Technol. 2018, 21, 93‑120, doi.org/10.1007/s10772-018-9491-z. [CrossRef]

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

[9] Tripathi, A.; Singh, U.; Bansal, G.; Gupta, R.; Singh, A.K. A Review on Emotion Detection and Classification using Speech. In Proceedings of the International Conference on Innovative Computing and Communications (ICICC), Online, 15 May 2020.

[10] Basu, S.; Chakraborty, J.; Bag, A.; Aftabuddin, M. A Review on Emotion Recognition using Speech. In Proceedings of the International Conference on Inventive Communication and Computational Technologies (ICICCT 2017), Coimbatore, India, 10‑11 March 2017.

[11] Lee, C.M.; Yildirim, S.; Bulut, M.; Kazemzadeh, A.; Busso, C.; Deng, Z.; Lee, S.; Narayanan, S. Emotion Recognition Based on Phoneme Classes. In Proceedings of the Eighth International Conference on Spoken Language Processing, Jeju Island, Korea, 4‑8 October 2004.

[12] Schuller, B.; Rigoll, G.; Lang, M. Hidden Markov Model-Based Speech Emotion Recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 6‑10 April 2003.

[13] Lee, J.; Tashev, I. High-Level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6‑10 September 2015.

[14] Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic Speech Emotion Recognition using Recurrent Neural Networks with Local Attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5‑9 March 2017; pp. 2227‑2231.

[15] Chen, R.; Zhou, Y.; Qian, Y. Emotion Recognition using Support Vector Machine and Deep Neural Network. In

Proceedings of the National Conference on Man-Machine Speech Communication, Lianyungang, China, 11–13 October 2017; pp. 122–131.

[16] Wieman, M.; Sun, A. Analyzing Vocal Patterns to Determine Emotion. Available online: http://www.datascienceassn.org/content/analyzing-vocal-patterns-determine-emotion (accessed on 3 September 2016).

[17] Shaqra, F.A.; Duwairi, R.; Al-Ayyoub, M. Recognizing Emotion from Speech Based on Age and Gender using Hierarchical Models. Procedia Comput. Sci. 2019, 151, 37–44.

[18] Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.

[19] Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans. Affect. Comput. 2015, 7, 190–202.

[20] Iliou, T.; Anagnostopoulos, C. Statistical Evaluation of Speech Features for Emotion Recognition. In Proceedings of the 2009 Fourth International Conference on Digital Telecommunications, Colmar, France, 20–25 July 2009; pp. 121–126.

[21] Kao, Y.; Lee, L. Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.

[22] Luengo, I.; Navas, E.; Hernáez, I.; Sánchez, J. Automatic Emotion Recognition using Prosodic Parameters. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal, 4–8 September 2005.

[23] Rao, K.S.; Koolagudi, S.G.; Vempada, R.R. Emotion Recognition from Speech using Global and Local Prosodic Features. Int. J. Speech Technol. 2013, 16, 143–160. [CrossRef]

[24] Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. Speech Commun. 2011, 53, 1062–1087.

[25] Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E. INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013; Interspeech: Lyon, France, 2013.