Harmonizing with Emotions: A Linguistic Analysis of Speech

Gaurav Verma
AIT-CSE (BDA) Department
Chandigarh University
Gharuan,Punjab
gv385699@gmail.com

Rituparna Seal
AIT-CSE Assistant Professor
Chandigarh University
Gharuan,Punjab
sonaseal2000@gmail.com

Namit Chawla
AIT-CSE Assistant Professor
Chandigarh University
Gharuan,Punjab
namit.e11486@gmail.com

Abstract-
Speech recognition has been an important topic in human computer interaction applications for many years. Emotions co me into play in human psychology. It is a means of expressing o ne's thoughts or feelings to others. Speech Emotion Recognition (SER) is a technique used to identify emotions conveyed by a speaker through their speech. It is particularly useful for computers with constrained processing capabilities. Power can be designed to optionally display or generate some universal emotions, such as neutral, angry, happy, and sad.
This function collects the following features: These include mel frequency cepstral coefficients (MFCC), chromatogram, mel scale spectrogram, spectral contrast, and tone center. In this study, emotions were detected using deep neural networks, with Softmax used to distribute speech across the output layer. Training utilized 1440 recordings from 24 individuals sourced from the RAVDASS speech database. The DNN achieved an impressive recognition accuracy of 96%, surpassing other algorithms like KNN, LDA, and SMO. The automatic recognition of emotions from human speech is increasingly prevalent, aiding in the seamless interaction between computers and humans. In the realm of human-computer interaction, the investigation of speech perception has been an ongoing research endeavor. Emotion plays an important role in humanpsychological life and is a means of expressing one's thoughtsor ideas to others. Speech Emotion Recognition (SER) involvesextracting the speaker's emotions from the speech signal.Universal emotions include neutrality, anger,happiness, sadness, fear, etc. is available. These emotions can be recognized or combined by low-budget smart machines. This work focuses on the extraction of speech features, including Mel Frequency Cepstral Coefficients(MFCC),chromatograms, Mel-scale spectrograms, as well as spectralcontrast and color center functions. In this study, deep neuralnetworks (DNN) were used to classify emotions.

Keywords: Emotion recognition, Deep neural network, Audio files, Chromogram, Spectrogram.

I. INTRODUCTION

Numerous applications today capitalize on human-computer interaction, with audio standing out as one of the most interactive mediums. A significant challenge in human-machine collaboration lies in the comprehension of speech. Moreover, various natural expressions can serve as indicators of emotions [1] [2]. This study's primary aim is to delve into behavioral aspects of speech. When two individuals engage in communication, they can readily discern the emotions conveyed through each other's words. Cognitive simulation aims to replicate how individuals perceive events [3]. The utility of speech recognition is manifold. Emotions can sway decisions, and if the underlying sentiments within a conversation can be accurately identified, the system can respond aptly. This underscores the importance of understanding emotions in speech for effective communication. If the thought in the conversation can be accurately detected, the system can respond effectively. Medicine, robotics engineering, contact centre applications, and other industries can benefit from the power of emotional intelligence. [4] [5].There are many ways to communicate; Music stands out as one of the fastest and most common ways for people to interact. This performance makes speech an effective means of communication between humans and machines. People always use all their senses to understand emotions expressed in words in machines. Therefore, the purpose of emotional intelligence is to use emotional intelligence to improve human-machine communication. In this system, speech recognition is directly affected by the extraction quality. Video removal takes all emotional lines as a unit and removes various acoustic features such as time structure, amplitude structure, frequency structure and structure. By comparing emotional expressions with non-emotional sentences in these contexts, the system identifies emotional distribution patterns and classifies emotional speech accordingly. Deep Neural Networks (DNN) have become incredibly successful in speech and image recognition. However, the use of DNNs in speech recognition is still limited. This article shows a way to extract emotions from audio using the library package in Python. Use 5-layer deep DNN to extract emotional speech and combine features of continuous segments to generate high-level features. The SoftMax classifier layer is employed for speech emotion classification, achieving an accuracy of 73.38% in speech recognition. Other techniques utilized in this context include K-nearest neighbor (KNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), Artificial Neural Network (ANN), Gaussian Mixture Model (GMM), among others. Traditional machine learning methods such as.< br>What is it used for? classification theory. The main problemsof speech recognition include signal processing to extract necessary features and classification for behavioral analysis. The average accuracy of most classifiers on speech-independent systems is lower than the average accuracy on speech-dependent systems. Improving automatic understanding of human speech could help improve interaction between humans and machines.

II. LITERATURE SURVEY

In their work [6], the author explores the application of Deeply Sprint Convolutional Neural Network (DSCNN) for the segmentation stage. They also propose recommendations aimed at filtering out undesirable sounds and ensuring the production of clear, clean audio. The goal is to develop a speech recognition system using an immutable and adaptive DSCNN model. [15] performed acomprehensive evaluation of various CNN and LSTM-RNN models. CNN architecture performs better than other LSTM and RNN architectures. Various assumptions for spectrum- and phoneme-based classification theory are given in [3]. [13] The researchers showcased the implementation of an end-to-end deep neural network for emotion recognition. They utilized extralinguistic information within words to establish communicative links and vice versa . Feature vectors are generated by the study of the finite element coupled to the latent method in [14], where the authors studied various features generated by traversing spectral images from AlexNet for speech recognition. A new way to classify sentences by sound. The algorithm employs categorical Hidden Markov Model (HMM) alongside short-term

logarithmic frequency power coefficients (LFPC) to characterize both the algorithm and the speech signal. The system categorizes thoughts into six groups, then trains and tests the new system on a single file. LFPC's effectiveness is compared with Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC) to assess the proposed strategy's efficiency. The results demonstrated that the average recognition rate and maximum recognition rate achieved 78%. Furthermore, studies indicate that LFPC excels particularly in classifying normal emotions [12]. In discussions involving cultural or semantic information, Rong et al. [1] introduced a system based on Random Forests for Trees (ERF Trees), which is characterized by several features aimed at enhancing curiosity.. This technique is used on small files with many features. We conducted a study to assess the motivation of Chinese speakers to evaluate the application process, and the results showed that curiosity increases speed. Moreover, EF Trees demonstrates superior performance compared to ISO Map, as well as well-known techniques such as PCA and Complex Scaling (MDS). The least effective function only achieves 16% accuracy in natural data across 84 parameters, while the most accurate of the 16 features in the female dataset achieves an impressive accuracy of 82.54%. Addressing the ordering problem in cognitive theory, Ayadi et al. [4] introduced a Gaussian Mixed Vector Autoregressive (GMVAR) technique that combines Gaussian Mixture Models (GMM) with outlier handling. The core concept of GMVAR revolves around its capability to distribute data across various media types and construct systems based on speech characteristics. The GMVAR method was applied to the Berlin Sentiment Dataset for evaluation. The outcomes indicated a classification accuracy of approximately 76%, surpassing feedback neural networks (67%), k-NN (67%), and HMM (71%). This method offers an advantage over HMM by enabling the differentiation of unconscious high and medium-high states [4].

III. Using the Template

Adaptive techniques of deep neural networks for speech recognition. To extract information from recordings, the approach combines Mel Frequency Cepstral Coefficients (MFCCs), chromatograms, Mel scale spectrograms, spectral contrast, and tone center features. The model categorizes speech into eight emotions: justice and peace, joy, sadness, anger, fear, disgust, and surprise. A 5-layer deep neural network (DNN) is trained for speech processing.

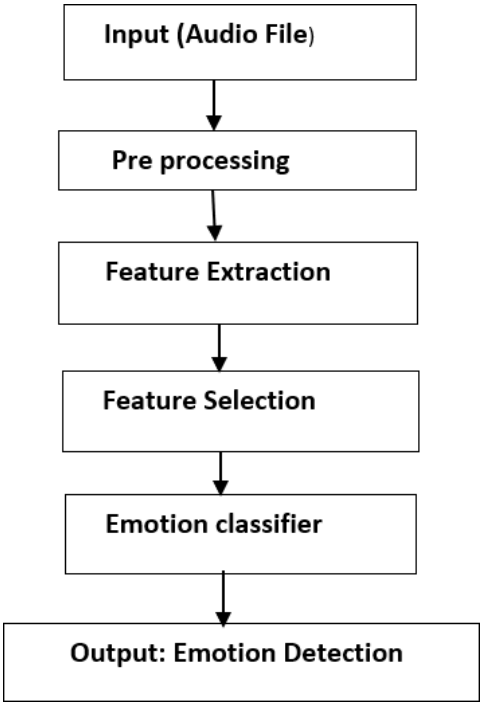


Fig. 1. Example of a figure caption. (*figure caption*)

A. Data Collection: This study uses the Ryerson Audiovisual Emotional Speech and Song Database (RAVDESS) dataset stored in .wav file format. The librosa package in Python is utilized to access audio files. For this study, the Ryerson Audiovisual Emotional Speech and Song Database (RAVDESS) was chosen as the data source. Specifically, a class comprising 24 participants (balanced by gender) was selected, resulting in a total of 1440 recordings.

B. Data preprocessing: This steps involves the process of converting dirty data into correct data. The word "purified information". It includes removing white space, replacing empty strings with important data, removing boilerplate code, removing transitions, and removing unnecessary attributes. If the dataset contains data from a category, replace categorical variables with numeric values. In total, our model consists of 5 layers.

C. Feature Extraction: The features retrieved include Mel Frequency Cepstral Coefficients (MFCC), chromatograms, and spectral contrast and color center features.

1. Mel frequency cepstrum coefficients (MFCC):

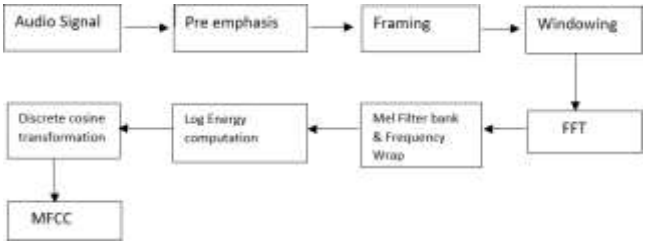


Fig. 2. Block Diagram of MFCC

Fig. 2. The MFCC Feature Extraction Process has the following steps:

Pre-emphasis: Pre-emphasis is necessary to show that it is more dynamic. This technique involves sending a speech sample into an amplifier to increase its liveliness. Adding state power provides additional detail.

K(z)=1- pz -1---------- (1)

K(z)= It is function of z.

Pz = p is a constant and z is a variable.

Framing: This step involves dividing the speech pattern into frames of 20-40 ms. Since the duration of the human voice may vary, this process is needed to adjust the volume of the speech. Although the nature of the sound is not constant (its amplitude will change over time), it behaves like a transmission for a short period of time.

Winding: This step is done after framing. The windowing technique reduces signal discontinuities at the beginning andend of each image. This process includes a 10 ms phase shift, meaning that half of the content from the previous pixel is replicated in each subsequent pixel.

Fast Fourier Transform (FFT): The spectrum of each frame is generated using Fast Fourier Transform (FFT). This transformation converts each sample of a time series frame into frequency lines. FFT is used to identify all frequencies available in a region. Each image was filtered using a mel-level filter bankconsisting of 20–30 triangles. The Mel scale filter determines the electric current value in a given image. Pi

The following calculation can be utilized to convert the frequency F into a decision using a Mel scale filter.

The following calculations can be used to convert the normal frequency f into a Mel scale filter:

$$Mel = 1127.01048 * \log(f/700 + 1) \text{ ---------- (2)}$$

Calculation of logarithmic power: The logarithmic function is applied after receiving the filter branches of each frame. Mammalian hearing serves as another source of inspiration. Human auditory perception does not operate on a linear scale. When noise is high, the human ear cannot detect large electrical changes. Linear calculation of power gives good sound that people can hear clearly.

Discrete Cosine Transform (DCT): In the final stage, logarithmic bank values are utilized to calculate the Discrete Cosine Transform (DCT). We usea scrolling time of 10 ms and a processing time of 25 ms. Moreover, 26 bandpass filters were employed, leading to the calculation of 13 Mel Frequency Cepstral Coefficients (MFCCs) for each frame. We have a total of 13 MFCC features in each frame. Additionally, we estimated the power inherent in each frame. By computing the temporal evolution of energy and MFCCs, we derived 13 accelerations and simultaneously obtained 13 MFCC features.

To create delta features based on M samples before and after, where Cm (t) represents the frame, use the following formula:

$$DCTu,v = Gu,v \sum_{n=0}^{N-1} \sum_{p=0}^{N-1} (Mn,p \cos((2n+1)\, u.\pi/ 2N).\ \cos((2p+1)\, v.\frac{\pi}{2P})) \text{ ---------- (3)}$$

Where Cm(t) represents the static coefficient of the frame, and the delta property is calculated as before and after M square.

2. Spectral Contrast:

Spectral contrast measures the harmonic strength of a sound at any given time. This feature is valuable because many audio recordings include sounds that vary in strength over time.

Measuring vibration may seem difficult. This change in energy can be measured using the spectral difference. Simple broadband pulses usually have a large contrast ratio, while broadband signals usually have a lower contrast ratio. While high frequency difference is most often associated with shortband pulses, wideband audio is often associated with negative consequences. Here the difference between the average power at the peak power pole and the average power the power difference is determined using the bottom or flat power pole..

3. Tonal Centroid:

The Tonal Centroid help identify changes in music or differences in pitch of the sound.

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)\ x(n)}{\sum_{n=0}^{N-1} x(n)}$$

- ------(4)

The following characteristics of the tonal centroid are considered:

4. Frequency and pitch :

Pitch and amplitude resonance are crucial aspects of the signal. Besides volume, duration, pitch, and area, other significant aspects of vibration pertain to sound. The simplest waveform is depicted by the sinusoid x(t), where a is the maximum amplitude, t is the duration in seconds, f is the frequency, and the initial phase. The frequency f of the simple model refers to the number of cycles repeated per second.

5. Melody : The frequency of sound is often considered music. According to this interpretation, music is "a musical ensemble collected in musical time according to certain rules and restrictions."

6. Harmony : The term "harmony" denotes the combination of sounds known as chords and their progression over time. Apart from describing sounds and their relationships, the term frequently refers to the mathematical process by which they are combined. In this latter context, harmony has several practical applications. We only remember the details of the harmonic content, that is, how the pitches come together in harmony and affect the tonality of the product. As we have seen, rhythm and rhythm are often associated with writing, which can be synchronous or sequential. Distinguishing the harmony of music in this way would be difficult because they interact.

7. Tonality : The vagueness of the concept of sound is one of the reasons for the lack of agreement on the various inductive methods of music. Castil-Blaze first coined the word "tonality" in 1821. Today, it generally refers to the process of interaction between different tones that creates harmony and music, including its most important element, the tonic or fundamental pitch level. (or stable) components.

It describes the sequence of sound events in a broader way than expected. The tone center aids in detecting harmonic changes.

C. D. System Design:

1. Application For Emotional Recognition Model : Users upload audio files containing a person's voice and utilize the model to predict the speaker's emotional state. This model is implemented in web applications constructed using the Flask architecture. The system comprises two main aspects: the system itself and the user interface. The user interface provides functionalities for managing user access and authentication. Initially, new users are prompted to input their information, typically including name, email, and password. This information is stored in a MySQL database. Registered users whose information is stored in the MySQL database can access the web-based application using their login credentials. The program allows them to predict emotions only if the inputs are accurate.

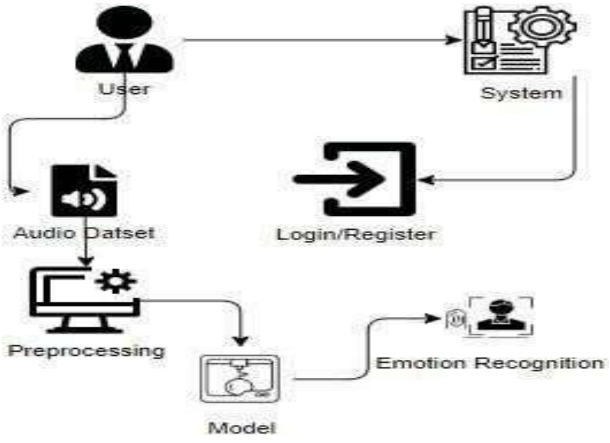
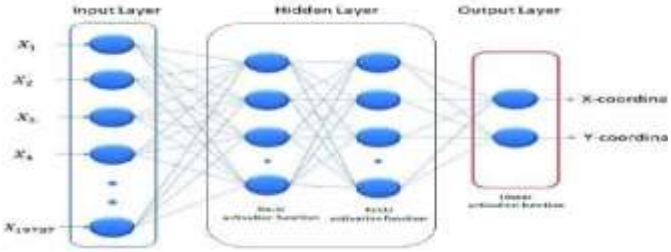


Fig. 3.Overall Workflow of the Emotion Recognition Application

E. Model for Deep Neural Network to Emotion Recognition: Deep learning can be very efficient while using fewer processors. Deep learning often uses deep-connectivity neural networks (DNN) to build models for activities that are impossible or difficult to perform with machine learning algorithms.

1. Architecture Diagram of DNN

Our 1440 sound files are split into training files (1008 sound files) and test files (432 sound files). In this instance, the first dataset comprises 80% of the data.



A five-level Sequence() structure is created. These data were trained 700 times using the training data.

F. Performance Metrics:

1. Hyper parameters

In the test data, the classification rate is computed by summing the number of correct predictions (diagonal elements) and dividing the result by the total sample size.

2. Recall

The recall is calculated as the sum of true positives divided by the sum of true positives and false negatives. In essence, it represents the proportion of actual positives that were correctly identified. The recall score ranges from 0.0 (indicating no true positives) to 1.0 (reflecting complete or perfect recall).

3. Precision Value

Increasing precision helps in reducing the number of false positives, while increasing recall helps in reducing false negatives.

4. F-Measures

Recall and accuracy can be combined into a single measure using the F-measure, which has two properties. The F-measure is equal to the product of the recall and precision factors.

5. ROC

Rate of Change (ROC) is a technical indicator that measures the percentage change in price between the current price and the price after a specified period of time.

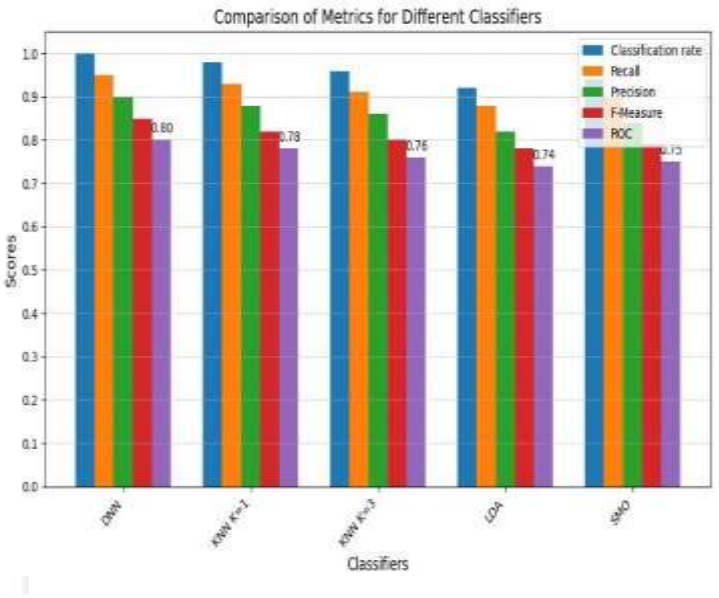


Fig. 5. Performance Metrics

G. Experimental And Results

We began by partitioning all the data into two groups: 80% for training and 20% for testing and 20% for verification purposes. We then calculate the MFCC features for each data in the training and testing data. We take a deep neural network that processes the acquired features as the first proposal. We use DNN using five convolutional layers. It has been created 700 times for the network we operate. The last output node has the activation function we use, called “softmax”. We also calculated losses using categorical linear precision. After 700 repetitions, training performance is up to 96% and testing accuracy is up to 80%. The confusion matrix of he training data is depicted in the figures.

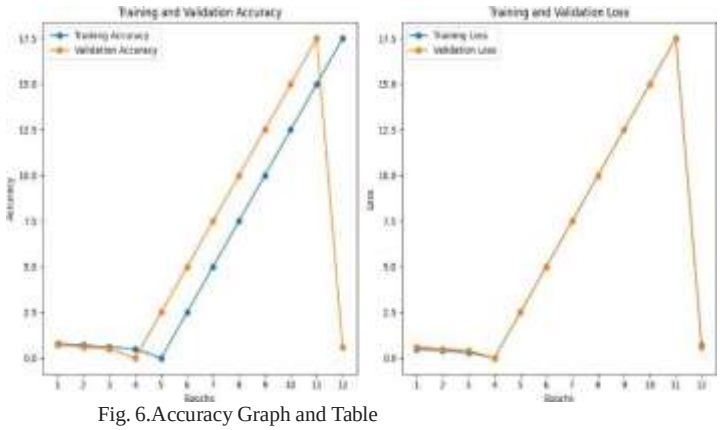
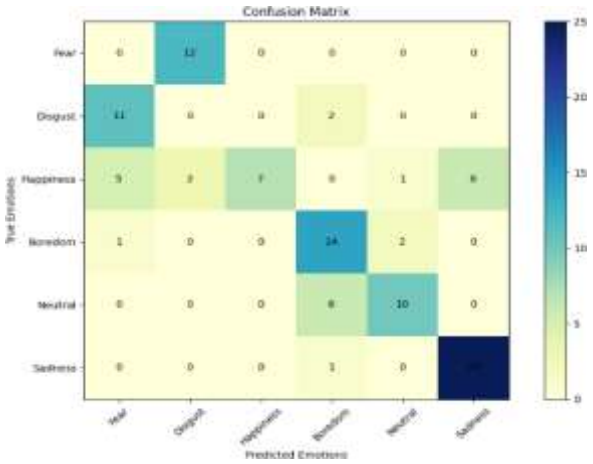


Fig. 6.Accuracy Graph and Table

Current research mainly focuses on the study explores various aspects of speech and their relationship with the context of speech..New features such as TEO-based features have also been developed. Different feature selection methods also try to find the best features for this task. However, the results of different studies are inconsistent. The main reason for this may be that each research study has only one perspective of the speech database.

TABLE I. Confusion Matrix for Validation Data



Correct identification of different behaviors is represented by the diagonal elements of the confusion matrix. Our network demonstrates accurate determination for most views.

IV. CONCLUSIONS

Deep learning algorithms can yield remarkable results. We have successfully developed a deep learning model for emotional intelligence, achieving a test score of 96%. Remember that feeling is a decision and different people will feel differently about the same music. This is why the output of algorithms trained by humans to measure behavior is sometimes inconsistent. Since the model is trained using the RAVDESS dataset, the speaker's accent will yield inconsistent results since the model is trained using the North American accent dataset. Our solution offers an effective way to recognize emotions in human speech through neural networks. A deep learning model was successfully developed using deep neural network architecture to predict speaker emotions in audio. The project was created as a web-based application that leverages the Flask architecture and integrates the user registration system into the user interface. The training model achieved an accuracy of 73.4%. It should not be forgotten that the perception of emotion is the perception of the source and that different people's evaluations of the same sound will be different. This algorithm has learned from human evaluations and can produce inconsistent results. Additionally,Since the model is trained on the RAVDESS dataset, variations in accent can lead to unpredictable results, as the model is specifically trained on the North American accent dataset.

REFERENCES

[1] Szegedy, Christian &Toshev, Alexander & Erhan, Dumitru. (2013). Deep Neural Networks for Object Detection. 1-9.

[2] Benk, Sal & Elmir, Youssef &Dennai, Abdeslem. (2019). A Study on Automatic Speech Recognition. 10. 77-85.10.6025/jitr/2019/10/3/77-85.

[3] Ashish B. Ingale & D. S. Chaudhari (2012). Speech Emotion Recognition. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2 Issue-1, March 2012.

[4] Chenchen Huang, Wei Gong, Wenlong Fu, Dongyu Feng, "A Research of Speech Emotion Recognition Based on Deep BeliefNetwork and SVM", MathematicalProblems in Engineering, vol. 2014, ArticleID 749604, 7 pages, 2014https://doi.org/10.1155/2014/749604

[5] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.

[6] E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", Pattern Recognition 44, PP.572-587, 2011.

[7] I. Chiriacescu, "Automatic Emotion Analysis Based On Speech", M.Sc. THESIS Delft University of Technology, 2009.

[8] T. Vogt, E. Andre and J. Wagner, "Automatic Recognition of Emotions from Speech: A review of the literature and recommendations for practical realization", LNCS 4868, PP.75-91, 2008.

[9] S. Emerich, E. Lupu, A. Apatean, "Emotions Recognitions by Speech and Facial Expressions Analysis", 17th European Signal Processing Conference, 2009.

[10] P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.

[11] M. Borchert, A. Dusterhoft, Emotions in speech—experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments, in: Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE'05 2005, 2005, pp. 147–151.

[12] L. Bosch, Emotions, speech and the asr framework, Speech Commun. 40 (2003) 213–225.

[13] S. Bou-Ghazale, J. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Trans. Speech Audio Process. 8(4) (2000) 429–442.

[14] R. Le Bouquin, Enhancement of noisy speech signals: application to mobile radio communications, Speech Commun. 18 (1) (1996) 3–19

[15] C. Breazeal, L. Aryananda, Recognition of affective communicative intent in robot-directed speech, Autonomous Robots 2 (2002) 83–104.

[16] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.

[17] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining Knowl. Discovery 2 (2) (1998) 121–167.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, in: Proceedings of the Interspeech 2005, Lissabon, Portugal, 2005, pp. 1517–1520.

[19] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE Trans. Audio Speech Language Process. 17 (4) (2009) 582–596.

[20] J. Cahn, The generation of affect in synthesized speech, J. Am. Voice Input/ Output Soc. 8 (1990) 1–19.

[21] D. Caims, J. Hansen, Nonlinear analysis and detection of speech under stressed conditions, J. Acoust. Soc. Am. 96 (1994) 3392–3400.

[22] W. Campbell, Databases of emotional speech, in: Proceedings of the ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, 2000, pp. 34–38.

[23] C. Chen, M. You, M. Song, J. Bu, J. Liu, An enhanced speech emotion recognition system based on discourse information, in: Lecture Notes in Computer Science—I (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3991, 2006, pp. 449–456, cited by (since 1996) 1.

[24] L. Chen, T. Huang, T. Miyasato, R. Nakatsu, Multimodal human emotion/ expression recognition, in: Proceedings of the IEEE Automatic Face and Gesture Recognition, 1998, pp. 366–371.

[25] Z. Chuang, C. Wu, Emotion recognition using acoustic features and textual content, Multimedia and Expo, 2004. IEEE International Conference on ICME '04, vol. 1, 2004, pp. 53–56.

[26] R. Cohen, A computational theory of the function of clue words in argument understanding, in: ACL-22: Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting on Association for Computational Linguistics, 1984, pp. 251–258.

[27] R. Cowie, R.R. Cornelius, Describing the emotional states that are expressed in speech, Speech Commun. 40 (1–2) (2003) 5–32.