**Mid semester evaluation**

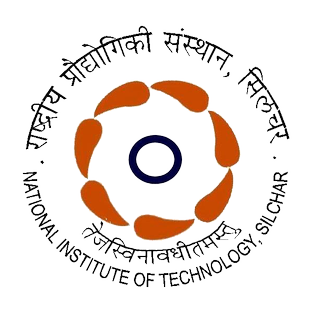
**Design of speech-emotion recognition system using deep networks**

**Submitted by**

AJAY TALUKDAR (18-14-118)

YUVARAJ SAI (18-14-125)

7th Semester, ECE, SEC-B



**Department of Electronics and Communication Engineering**

**NATIONAL INSTITUTE OF TECHNOLOGY, SILCHAR**

**CONTENTS**

* **Introduction**
* **Literature review**
* **Research gap in Recognition**

# **Objective**

# **Progress/probable solution**

# **Conclusion**

# **References**

**Introduction**

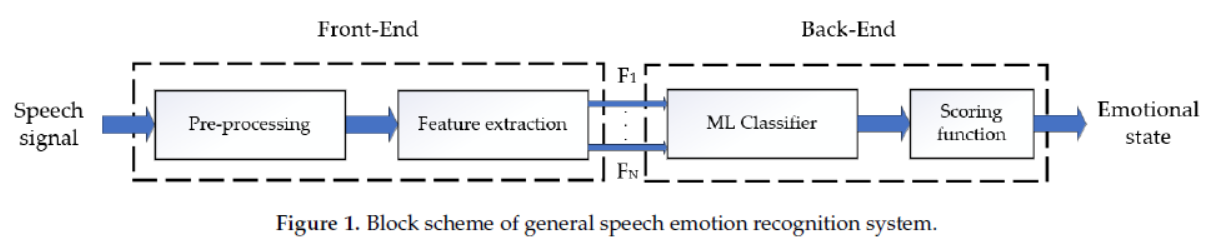
Human communication through the spoken language is the base for Information exchange and it is the main aspect of the society since the first human settlements. Speech is the fastest and best normal way of communicating amongst humans. Speech production is a multi-step process by which thoughts are generated into spoken utterances. Speech signals convey not only words and meanings but also emotions.

In the same way, Emotions play an extremely important role in human mental life. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Since emotions form an integral part of human interactions, they have naturally become an important aspect of the development of HCI-based applications. Emotions can be technologically captured and assessed in a variety of ways, such as facial expressions, physiological signals, or speech. With the intention of creating more natural and intuitive communication between humans and computers, emotions conveyed through signals should be correctly detected and appropriately processed. This has opened up a new research field called automatic emotion recognition, having basic goals to understand and retrieve desired emotions. Compared to many other biological signals (e.g., electrocardiogram), speech signals usually can be acquired more readily and economically. This is why the majority of researchers are interested in speech emotion recognition (SER).

Speech Emotion Recognition (SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. While humans can efficiently perform this task as a natural part of speech communication, the ability to conduct it automatically using programmable devices is still an ongoing subject of research.

Studies of automatic emotion recognition systems aim to create efficient, real-time methods of detecting the emotions of mobile phone users, call center operators and customers, car drivers, pilots, and many other human-machine communication users. Adding emotions to machines has been recognized as a critical factor in making machines appear and act in a human-like manner. Robots capable of understanding emotions could provide appropriate emotional responses and exhibit emotional personalities.

The task of speech emotion recognition (SER) is traditionally divided into two main parts: feature extraction and classification, as depicted in Figure 1. During the feature extraction stage, a speech signal is converted to numerical values using various front-end signal processing techniques. Extracted feature vectors have a compact form and ideally should capture essential information from the signal. In the back-end, an appropriate classifier is selected according to the task to be performed.



Examples of widely used acoustic features are mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCC), short-time energy, fundamental frequency (F0), formants, etc. Traditional classification techniques include probabilistic models such as the Gaussian mixture model (GMM), hidden Markov model (HMM), and support vector machine (SVM). Over the years of research, also various artificial neural network architectures have been utilised, from the simplest multilayer perceptron (MLP) through extreme learning machine (ELM), convolutional neural networks (CNNs), to deep architectures of residual neural networks (ResNets) and recurrent neural networks (RNNs). In particular, long short-term memory (LSTM) and gated recurrent units (GRU)-based neural networks (NNs), as state-of-the-art solutions in time-sequence modelling, have been widely utilised in speech signal modelling. In addition, various end-to-end architectures have been proposed to learn jointly both extraction of features and classification.

**Literature review**

Over the last years, an excessive investigation has been completed to recognize emotions by using speech statistics. Farah Chenchah, Zied Lachiri (2014)[1] set up a speech emotion recognition system based on the wavelet packet energy and entropy features. To test the effectiveness of the system, they performed it on two sets of data SAVEE and IEMOCAP, one is full acted and other is scripted with spontaneous recording. They have examined wavelet packet energy and entropy features applied to Mel, Bark and ERB scale applied with Hidden Markov Model (HMM) as classification system. The results show that wavelet packet filter bank with ERB scale give promising classification accuracy for both of databases.

Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, Bjorn W. Schuller (2021)[2] proposed a novel framework that can capture both the semantic and the paralinguistic information in the signal. Both semantic and paraliguistic features are then combined to a unified representation using a novel attention mechanism. The unified feature vector is passed through a LSTM to capture the temporal dynamics in the signal, before the final prediction.The proposed model is evaluated on the SEWA dataset and produces state-of-the-art results on the valence and liking dimensions.

El Ayadi et al. (2007) [3] proposed a Gaussian mixture vector autoregressive (GMVAR) approach, which is mixture of GMM with vector autoregressive for classification problem of speech emotion recognition. The key idea of GMVAR is its capability to multi-modality in their dissemination and design the dependency between speech feature set. Berlin emotional dataset was used for evaluation of GMVAR. The experimental result shows classification accuracy achieves 76% when for HMM reached 71%, for k-NN 67% and 55% for feed-forward neural networks. The advantage of this method better differentiation amongst high and low arousal with neutral emotions compare to HMM .

Wu et al. (2003)[4] proposed a new modulation spectral features (MSFs) human speech emotion recognition. Appropriate feature extracted from an auditory-inspired long-term spectro-temporal by utilizing a modulation filterbank and an auditory filterbank for speech decomposition. This method obtained acoustic frequency and temporal modulation frequency components for convey important data which is missing from traditional short-term spectral features. For classification process, SVM with radial basis function (RBF) are adopted. Berlin and Vera am Mittag (VAM) are employed to evaluate MSFs. In experimental result, the MSFs display capable performance in comparison with MFCC and perceptual linear prediction coefficients (PLPC). When MSFs utilized augment prosodic features, there is a considerable improvement in performance of recognition. Furthermore overall recognition rate of 91.6% is achieved for classification.

Lee et al. (2011)[5] represent a hierarchical computational structure to identify emotions. This method via following layers of binary classifications, maps input speech signal in one of the corresponding emotion classes. The main concept of different level in tree is to solve the classification task in easiest way to diminish error propagation. AIBO and USC IEMOCAP databases are employed to evaluate the classification method. Over the baseline SVM, the absolute result improve the accuracy archives an absolute improvement of 72.44%- 89.58%. The consequence proves the reported hierarchical method is efficient for classifying emotional speech in various databases .

Chen et al. (2012) [6] aimed to improve speech emotion recognition in speaker-independent with three level speech emotion recognition method. This method classify different emotions from coarse to fine then select appropriate feature by using Fisher rate. The output of Fisher rate is an input parameters for multi- level SVM based classifier. Furthermore principal component analysis (PCA) and artificial neural network (ANN) are employed to reduce the dimensionality and classification of four comparative experiments, respectively. Four comparative experiments include Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN. Consequence indicates in dimension reduction Fisher is better than PCA and for classification, SVM is more expansible than ANN for emotion recognition in speaker independent is. The recognition rates for three level are 86.5%, 68.5% and 50.2% separately in Beihang university database of emotional speech (BHUDES).

Albornoz et al.(2010)[7] investigate a new spectral feature in order to determine emotions and to characterize groups. In this study, based on acoustic features and a novel hierarchical classifier, emotions are grouped. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design a novel hierarchical techniques for classification of emotions. The innovation of the proposed method is two things, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. Experimental result in Berlin dataset demonstrates the hierarchical approach achieves the better performance compare to best standard classifier, with decuple cross-validation. For example, performance of standard HMM method reached 68.57% and the hierarchical model reached 71.75% [16].

Grimm et al. (2017) [8] proposed a multi-dimensional model by utilizing emotion primitives for speech emotion recognition. Three dimension were made by composing of three different value of emotion primitives, which is called valence, activation, and dominance. The value of these factors assumed to be in the range of [-1, +1]. A textfree, image-based method was introduced to assess the emotion primitives and achieves best inter-evaluator agreement. For extracting acoustic feature such as energy, pith and spectral specifications, both fuzzy logic and rulebased estimator are employed. The approached are validate by testing two EMA and VAM datasets, which are acted emotion and spontaneous speech emotion. Both dataset are recorded form talk-show in German TV. Finally, for mapping the emotion primitives to certain emotion category, k-NN was employed as a classifier. K-NN achieves total recognition rate up to 83.5% .

**Research gap in Recognition**

1. In the future works, it is advised to use others classification methods such Support Vector Machines (SVM) or Neural Networks. Moreover, it is proposed to add different noises to data to test the robustness of the proposed system. Combining emotional signals with psychological perception can be a promising idea to ameliorate the effectiveness of the system.
2. In future work, we intend to use a single network to simultaneously capture the semantic and the paralinguistic information in the speech signal. This will result in simplifying, and at the same time, reducing the number of parameters of the model. Additionally, we intend to investigate the performance of the proposed method on categorical emotion recognition datasets.
3. To further improve the classification performance, we shall study the implementation of a two-stage classifier. In the first stage, emotions are classified into high arousal, low arousal, and neutral emotions using our proposed method. In the second stage, another classifier is used to distinguish between emotions in the same category.
4. With possible refinement in future work, the performance of modulation domain features could be further improved. Hence, further research on the use of MSFs for SER can be beneficial
5. The major limitation of the approach described here is the empirical nature of the proposed hierarchical structure. While the proposed method has the advantage of being intuitive and efficient to design, it does not ensure an optimal solution. Our future work plans to investigate an automatic procedure to generate the hierarchical structure. This can minimize the need for several iterations of empirical testing. A specific related question for future work surrounds the derivation of a hierarchical structure that will not only optimally balance performance accuracy and combinatorial complexity but also yield results that are intuitively interpretable in light of psychological theories of emotions.
6. Further research should focus on the following aspects: first, the combination of special emotional information should be paid attention to, such as the fundamental frequency rise in the end of surprising sentence, the shaking sound of fear, etc. Second, use fuzzy theory to find the probability of some kind of emotions.
7. In future works the hierarchical classifier will be tested with noisy signals. Furthermore, these results will be compared with another model which take into account gender variability in an explicit manner. Similar analyses on other languages are also planned.
8. In our future work we plan to fuse other modalities with the acoustic analysis. The integration can be accomplished in a straightforward manner by implementing additional rules in the rule base. In this framework, speaker-dependent emotion expression variations are explicitly described by speaker models: this can in turn lead to model-based or parameter-driven emotion recognition that is adapted to individual speakers. Finally, the integration of the automatic emotion estimation into specific man–machine interaction applications, such as a humanoid robot, will indicate further needs in this field of research.

# **Objective**

# **Progress/probable solution**

# **Conclusion**

# **References**

1. Farah Chenchaha\*, Zied Lachiri “Speech emotion recognition in acted and spontaneous context” 6th International conference on Intelligent Human Computer Interaction, IHCI 2014
2. Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, Bj¨orn W. Schuller “SPEECH EMOTION RECOGNITION USING SEMANTIC INFORMATION” Mar 2021
3. M. M. H. El Ayadi, M. S. Kamel, and F. Karray, “Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models,” in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07, 2007, vol. 4, pp. IV–957–IV–960.
4. S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.
5. C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011.
6. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.
7. E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Spoken emotion recognition using hierarchical classifiers,” Comput. Speech Lang., vol. 25, no. 3, pp. 556–570, Jul. 2011.
8. M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” Speech Commun., vol. 49, no. 10–11, pp. 787–800, Oct. 2007.

**Introduction** should cover speech, emotion and why it is necessary to do the estimation of emotion from speech.

**Literature review**

* **Ref(author and year)**
* **Database(no of indeviduals,no of speech samples, included/natural)**
* **Methods/findings(what they have done, handcrafted?)**
* **Constraints of limitations**

**Research gap in Recognition**

# **Objective**

# **Progress/probable solution**

# **Conclusion**

# Speech, language refer to the means of communication used by people. Speech is the expression of ideas and thoughts by means of articulate vocal sounds, or the faculty of thus expressing ideas and thoughts.

**Speech** is human vocal [communication](https://en.wikipedia.org/wiki/Communication) using [language](https://en.wikipedia.org/wiki/Language).

Speech production is a multi-step process by which thoughts are generated into spoken utterances. the communication or expression of thoughts in spoken words

**emotion**, a complex [experience](https://www.britannica.com/topic/experience-philosophy-and-psychology) of [consciousness](https://www.britannica.com/topic/consciousness), bodily [sensation](https://www.britannica.com/topic/sensation), and behaviour that reflects the personal significance of a thing, an event, or a state of affairs.

“are all those feelings that so change men as to affect their judgements, and that are also attended by [pain](https://www.britannica.com/science/pain) or pleasure. Such are [anger](https://www.britannica.com/science/anger), pity, fear and the like, with their opposites.”

an affective state of consciousness in which joy, sorrow, fear, hate, or the like, is experienced, as distinguished from cognitive and volitional states of consciousness.

a conscious mental reaction (such as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body

Speech Emotion Recognition (SER) is the task of recognizing the emotional aspects of speech irrespective of the semantic contents. While humans can efficiently perform this task as a natural part of speech communication, the ability to conduct it automatically using programmable devices is still an ongoing subject of research.

Studies of automatic emotion recognition systems aim to create efficient, real-time methods of detecting the emotions of mobile phone users, call center operators and customers, car drivers, pilots, and many other human-machine communication users. Adding emotions to machines has been recognized as a critical factor in making machines appear and act in a human-like manner. Robots capable of understanding emotions could provide appropriate emotional responses and exhibit emotional personalities.

Emotion classification[1] is one of the most challenging tasks in a speech signal processing domain. The problem of speaker or speech recognition becomes relatively an easier one when compared with recognizing emotion from speech. Sound signal is one of the main medium of communication and it can be processed to recognize the speaker, speech or even emotion. The basic principle behind emotion recognition lies with analysing the acoustic difference that occurs when uttering the same thing under different emotional situations. In addition to the features corresponding to the speaker and/or the speech, the sound signals do have some features that represents the emotional state of the speaker. The paper addresses the problem of emotion classification for human speech. The study is aimed at exploring dependencies the nature of utterance have with the human emotional state. Since the emotions have a direct influence on the nervous system, the heart rate also is affected by them. So the heart rate of a person can also be measured to get information about the emotional status of person[2, 3]. It is interesting to note that the speech signals are also a representative of the heart rate of the speaker since the heart rate also affects the speech. The work in [4] says that if there is a negative stimuli that causes negative emotion the heart rate decelerate more actively than when there is positive stimuli.

**Literature review**

|  |  |  |  |
| --- | --- | --- | --- |
| **Ref** | **Database** | **Methods/findings** | **Constraints of limitations** |
| Albornoz et al. (2010) |  | based on acoustic features and a novel hierarchical classifier, emotions are grouped. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design a novel hierarchical techniques for classification of emotions. The innovation of the proposed method is two things, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. | In future works the hierarchical classifier will be tested with noisy signals. Furthermore, these results will be compared with another model which take into account gender variability in an explicit manner. Similar analyses on other languages are also planned. |
| Yang & Lugger(2009) |  | propose a new set of harmony features for automatic emotion recognition from speech signals. They are based on the psychoacoustic harmony perception known from music theory. Starting from the estimated pitch contour of an utterance, we calculate the circular autocorrelation of the pitch histogram on the logarithmic semitone scale. It measures the occurrence of different two-pitch intervals which cause a consonant or dissonant impression.  In Classification step, Bayesian classifier plays an important rule with a Gaussian class-conditional likelihood. | Future works include experiments with the triad features in (8), an optimization of the dissonance function dðsÞ in (6), a study of these harmony features in other classifiers or combinations of classifiers, and an evaluation of these harmony features with more realistic non-acted emotional speech. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |