Step-by-step Approach to Automatic Speech Emotion Recognition

Purnima Chandrasekar
Asst. Prof., Dept. of Electronics & Telecommunication Engg.,
Thakur College of Engineering & Technology

Shailendra Pratap Shastri
Asso. Prof., Dept. of Electronics & Telecommunication Engg.,
Thakur College of Engineering & Technology

ABSTRACT
Humans use emotions to express themselves naturally either through facial expressions or through speech. Emotions play an important role in influencing the decision-making capability of human beings as human mind is influenced by personal experiences as well as physiological, communicative and behavioral reaction to external stimulus. While considering emotions displayed through speech, one needs to understand that a speech signal not only conveys the emotional state of the speaker which is visible from the intent of the message as well as the gender of the person and the language spoken. While an effective communication between humans through speech ensures exchange of right amount of ideas, messages and perceptions, interaction between human and machine with the same intent becomes challenging as a machine is expected to mimic the mechanism of human perception. Automatic Speech Emotion recognition (ASER) systems has found usefulness in several applications viz. healthcare, counseling, call center communication etc. Primary to this system are three basic components viz. creation of emotional speech corpus, extraction of features relevant to emotion detection and classification of emotion in the test speech using appropriate classifiers. This paper surveys extensively the prominent features extracted, several dimension reduction techniques and classifiers commonly used in recent times. It also throws light on the concept of auto encoders being used in recent times in the process of ASER.

General Terms
Pattern Recognition

Keywords
ASER, feature extraction, dimensionality reduction, auto encoders

1. INTRODUCTION
ASER system can be immensely used in the field of medical science, robotics engineering, call center application as well as in applications like identifying mental state of the driver to initiate his/her safety, diagnostic tool for therapy, etc. [1,2]. While implementing an ASER system, the primary task is to choose an emotion recognition corpus followed by identifying relevant features related to speech and an appropriate choice of a classification model. A corpus that is contextually rich is preferred and if the intent is to build an efficient affect detection system, it is important that the corpora or database consists of real and natural emotional speech spoken by a large number of male and female persons [1].

There exist many known databases utilized in the literature survey viz. EmoDB, Danish Emotional Speech Database, SUSAS, INTERFACE etc. [2] whose speech samples have been used to evaluate, rank and compare different features and discrimination techniques. Creating such emotionally labeled speech corpus becomes a daunting task as it involves collecting, labeling and segmenting into predefined or unknown emotion categories, spontaneous, induced or acted emotional speech samples. [3]

Two models have become common referred to in speech emotion recognition: discrete emotional model, and dimensional emotional model. The former is based on Dr. Paul Ekman's theory of six categories of basic emotions viz. sadness, happiness, fear, anger, disgust, and surprise which people experience in their data to day life while the latter is an alternative model that uses a small number of latent dimensions to characterize emotions such as valence, arousal, control, power. In the dimensional emotional model approach, emotions are not independent of each other but analogous to each other in a systematic way. [4]

Features are essential to the emotion classification process as their extraction converts the original data (in this case speech sample) to its most important characteristics. The choice of features to be extracted is crucial as it directly influences the accuracy of the classification results. In the literature survey, several features relevant to emotion recognition has been used and broadly they can be categorized into 4 groups viz. prosodic, spectral, voice quality, and features based on Teager energy operator. Once these features are extracted, they are then passed to the classification system which has a wide range of classifiers available to them. [4,5]

The process of selecting relevant and useful subset of the given set of extracted features called as feature selection is important to solve the most common issue of curse of dimensionality. Feature selection algorithms are useful as one can extract many features and there is no certain set of features to model the emotions [4]. In recent times, autoencoders have also been explored. The autoencoder algorithm belongs to a special family of dimensionality reduction methods that is implemented using artificial neural networks. [6]

Identifying the appropriate emotion thereby improving the system performance is an important criterion while selecting the correct classifier. Many classifiers have been chosen for speech emotion recognition system, but it is very difficult to conclude which performs better-there is no clear winner. Recent works in this direction has brought focus mostly on deep neural network and architecture, hybrid classifiers and fusion methods for emotion recognition system [7].

The rest of the paper is organized as follows. In section 2, the different emotional speech databases explored as far as emotion recognition is concerned is discussed. In section 3, different features prominently extracted for emotion recognition is discussed. In section 4, the need for dimensionality reduction is discussed and the concept of auto encoders is introduced. In

section 5, the different classifiers used have been discussed. In section 6, the results obtained by few of the authors by combining appropriate extracted features, dimensionality reduction techniques and suitable classifiers have been discussed.

2. EMOTIONAL SPEECH DATABASES

Three kinds of emotional speech databases are available for developing an ASER system viz. natural, simulated (Acted) and elicited (Induced) emotional speech database. Natural databases are usually developed on spontaneous speech of real data. Such databases are developed from recordings of call center conversations, cockpit recordings during abnormal conditions, conversation between a patient and a doctor, conversation with emotions in public places etc. Simulated databases are developed with the help of professional experienced, trained and professional actors. These are databases that can vouch for recordings with full blown emotions. Elicited databases are those which are developed with artificial emotional situations created without the knowledge of the speaker in order to induce appropriate emotions within the speaker. [7]

Several emotional speech databases across multiple languages have been used in the literature survey of ASER. Speech under simulated and actual stress (SUSAS) is an English database in which isolated-word utterances under simulated or actual stress has been recorded with emotions like anxiety, fear, anger, depression, stress widely covered. The Berlin Database of Emotional Speech is a German database covering emotions like anger, boredom, disgust, fear, joy, sadness and neutral recorded with the help of non-professional actors. The Danish Emotional Speech Database is a Danish database that has recordings of actors familiar with radio theater uttering single words, sentences and passages of fluent speech in different emotional states viz. Anger, happiness, sadness, surprise and neutral. The AIBO database has been developed in English and German consisting of recordings of children interacting with a WOZ robot. Different emotions like angry, bored, emphatic, helpless, joyful, motherese, reprimanding, surprised, touchy (irritated), neutral and rest were recorded. RUSLANA is a Russian database that has been developed with recordings of actors expressing emotional sentences with anger, fear, happiness, sadness, surprise and neutral. [8]

As far as Indian languages are concerned, few databases have been developed. Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP-SESC) which has been recorded using the professional artists from All India Radio (AIR), Vijayawada, India in Telugu covering emotions like anger, compassion, disgust, fear, happy, neutral, sarcastic and surprise [9]. Yet another database developed by Indian Institute of Technology Kharagpur called as Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) comprises of recordings of professional artists from Gyanavani FM radio station, Varanasi, India using neutral (emotion free) text prompts covering emotions like anger, disgust, fear, happy, neutral, sad, sarcastic and surprise [10]. Details of few more emotional speeches databases based on Indian languages can be obtained from [11].

While developing an emotional speech database, some obstacles that may arise include annotating an audio recording i.e. deciding whether to label the whole conversation, a single part, which one(s), human intervention, time and cost as it requires trained personnel to analyze speech samples and give an annotation, collecting data from mediums like news and films as emotions are not natural and can be with bias and so on [12]. Another major issue while creating a database is the reduction of noise which can easily get incorporated with the speech data sample. [1]

3. FEATURE EXTRACTION

The next step after creating an emotional speech database is that of feature extraction. This is a crucial step as it involves detecting and choosing salient acoustic features to feed the appropriate classifiers [3]. Characteristics of the human vocal tract and hearing system is represented by different features of speech signal [1]. Fig. 1 depicts the categories of features that are commonly used in emotion speech recognition, the description of which is given in Table II.

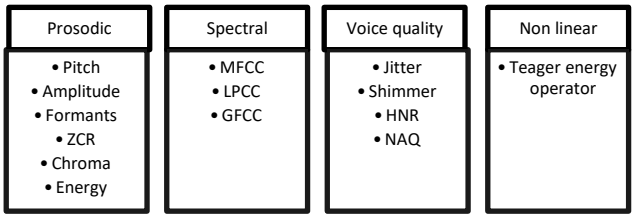

Fig 1: Categorization of commonly used speech features

3.1 Prosodic features

These features are also known as para-linguistic features and are associated with those elements of speech that are properties of large units as in syllables, words, phrases, and sentences. Prosodic features are known to convey the most distinctive properties of emotional content from speech for the purpose of emotion recognition. Fundamental frequency, energy and duration are the most distinctively used prosodic features. These features are pertinent to emotion recognition as they have been found useful in detecting certain emotions viz. high arousal emotions such as anger, happiness or surprise yields increased energy while disgust and sadness result with decreased energy; F0 contour decreases over time during the expression of anger which in contrast, increases over time during the expression of joy; Low-level arousal such as sadness yields lower mean F0, F0 variability, and vocal intensity compared to natural speech, while also F0 decreases over time; Gross statistics such as the mean, maximum and minimum values, and the range of the F0 are found to be the most salient aspects of F0 contour [4]. Variation in pitch, differences in the time duration of the speech, characteristics of stress, shortening/lengthening of speech are also other prosodic characteristics that are known to provide cues about a specific emotion [13].

3.2 Spectral features

The sound that comes out is determined by the shape of the vocal tract and the characteristics of the vocal tract are well represented in the frequency domain [4]. Also known as spectral or segmental features, for vocal tract features to be extracted, a speech segment of length 20–30 ms is generally required [7]. Typically, Spectral features are obtained by transforming the time domain signal into the frequency domain signal using the Fourier transform [4].

3.3 Voice quality features

Voice quality features are determined by the physical properties of the vocal tract to which if any involuntary changes occur, leads to production of a speech signal with a particular emotion [4]. In the literature of ASER it has been stated that voice quality features, as the characteristic auditory colouring of an individual voice, have been shown to be discriminative in

expressing positive or negative emotions and by combining these features with prosodic features, performance of ASER has been shown to improve [14].

3.4 Nonlinear speech feature

Certain applications like predicting the mood of car driver to avoid any untoward mishaps, understanding a student's mental state so that proper counselling can be given, etc. requires detecting stressed emotions from the speech. The speech under stressful conditions affects the nonlinear flow of air in the vocal tract system when the speech signal is produced. Hence, these non-linear speech features are very important for detection of stress. Teager energy operator is one such feature that is used for detecting the stressed emotions like Lombard, angry, loud versus neutral emotions etc. [15]. A description of few features extracted for the purpose of speech emotion recognition in the literature survey is as shown in Table 2.

4. DIMENSIONALITY REDUCTION

In order to have an effective speech emotion characterization, feature extraction should be combined with dimensionality reduction. Dimensionality reduction is necessary approach to downsize data by reducing the number of features that describe the data. In order to reduce the dimension of the large sets of the descriptors, two kinds of practice are widespread [16]: projecting features in a reduced dimension space (i.e. linear discriminant analysis or principal component analysis) and selection of a subset of discriminant features (i.e. Fisher selection algorithm or genetic algorithm). In recent times, auto encoders have been used for dimensionality reduction purpose.

Auto encoders comprise of three layers viz. an input layer and an output layer of the same size, and hidden layers that contain fewer neurons than the input and output just like other neural networks. These are designed to reconstruct the original input data as output for which it comprises of an encoder that compresses the input data and transforms into a denser representation and a decoder that reconstructs the data [4].

The significance of using auto encoders is that unlike traditional dimensionality reduction methods like PCA, this does not rely on selecting meaningful features from the entire list of components thereby reducing subjectivity and significant human interaction from the analysis. It is also observed to retain all the information of the original data set by encoding all the information into the reduced layer, leading to the decoder, in turn, being better equipped to reconstruct the original data set. The most basic architecture of an auto encoder has the same number of dimensions in the input layer as well in the output layer, but the hidden layer has a smaller number of dimensions which is where the dimension reduction occurs. The disadvantage as observed includes requirement for greater computation and the tuning. [6]

There are several types of auto encoders such as variational auto encoder (VAE), denoising auto encoder (DAE), sparse auto encoder (SAE), adversarial auto encoder (AAE) etc. [4] that have been used for the purpose of speech emotion recognition. Table 1. describes the different auto encoders that have been explored till date.

Table 1. Few auto encoders used in recent times

Sr. No.	Name of auto encoder	In-between
1.	Variational auto encoder	Works on variational Bayes approach with input values described in probabilistic manner [17]
2.	Denoising auto encoder	More useful while reconstructing the original input that is corrupted due to noise [18]
3.	Sparse auto encoder	Helps finds a common structure in a small target data which is then used to reconstruct source data so that useful knowledge transfer from source data to a target task can be completed. [19]
4.	Adversarial auto encoder	This works through an encoder-decoder neural network by imposing the encoder output distribution into some known prior distribution, such as Gaussian [20]

Table 2. Description of different types of features that can be extracted from speech sample

Feature	Basic description
Pitch	Pitch is a feature related to vocal folds vibration. It's defined as the number of periods of vocal folds vibration per second [21]
Amplitude	When the speaker is angry or happy, the volume of speech is generally high. When speaker is sad or depressed, the volume of speech is generally low. [22]
Formants	They are an acoustic resonance of the human vocal tract which is measured as an amplitude peak in the frequency spectrum of the sound. [23]
Mel frequency cepstral coefficients (MFCC)	They are coefficients which represent audio, based on perception of human auditory systems with frequency bands positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT [23].
Linear prediction cepstral coefficients (LPCC)	LPCC features are extracted from spectra, using the energy values of linearly arranged filter banks, which equally emphasize the contribution of all frequency components of a speech signal [24]
Zero Crossing Rate (ZCR)	It indicates the frequency of signal amplitude sign changes
Chroma features	It is a quality of pitch class that refers to a colour of a musical pitch whose features aim at representing the harmonic content of a short-time window of audio [25]
Energy	High arousal emotions such as anger happiness or surprise yields increased energy while disgust and sadness result with decreased energy [4]

Gamma tone frequency Cepstral coefficients (GFCC)	GFCC is calculated using the ERB filters [26]
Mel Energy Spectrum Dynamic Coefficients (MEDC)	MEDCs are used to represent energy features of the emotional states (and is same as MFCC) with the difference that logarithmic mean of energies is taken after filtering process
Jitter	Jitter and shimmer are measures of the cycle-to-cycle variations of the fundamental frequency and amplitude [27]
Shimmer	
HNR	It is a pitch-based feature based on the auto correlation of the input signal and calculated as log HNR to better model human perception [28]
Normalized amplitude quotient (NAQ)	NAQ correlates better with arousal than with valence for both genders [29]

Table 3. Review of few papers describing the end-to-end ASER system proposed and designed

Database used	Emotions in it	Features extracted	Dimensionality reduction technique	Classifier	Accuracy
Ravdess Dataset, *Toronto Emotion Speech Set (TESS)* [6]	Neutral, calm, happy, sad, angry, fearful, disgust, and surprised; *anger, happiness, disgust, fear, sadness, neutral and pleasant surprise*	MFCC	Simple auto encoder	SVM, **Decision tree**, **1D-CNN**	91%, **90%**, **91%** (TESS), 40%, **75%**, **80%** (Ravdess),
RAVDESS dataset [30]	Neutral, calm, happy, sad, angry, fearful, disgust, and surprised	Pitch, MFCC	-	SVM, **Decision tree**	91%, **62%**
EMO Db [21]	fear, neutral, anger, boredom, disgust, joy and sadness	Pitch based features	PCA	Decision tree, **KNN**, *SVM* and Subspace discriminant algorithms	72%, **78.7%**, *77.3%*, 72% (Simple classification); 51.74%, **57.1%**, *56.44%*, 59.32% (Hierarchical classification)
Spontaneous spoken interactions between Spanish elderly people and a simulated virtual coach [16]	Calm, sad, happy, puzzled and tense	ZCR, Energy, Spectral features, MFCC, Chroma features	-	SVM	80.75%
Interactive Emotional Dyadic Motion Capture (USCIEMOCAP) Database [18]	Angry, happy, sad and neutral	Acoustic features viz. MFCC, F0, Jitter, Shimmer etc.	Denoising autoencoder (DAE)	SVM	61.5% (using 800 hidden nodes of DAE)
Ravdess Dataset [31]	Anger, sad, fear, excitement, happiness and neutral	MFCC, DWT, Pitch, Energy, ZCR	Min, Max, Mean, Median and standard deviation	SVM, *Decision tree*, and LDA	70%, *85%*, 65%
Interactive Emotional Dyadic Motion Capture (IEMOCAP) [32]	Neutral, angry, sad, and happy	openSMILE features (Combination of spectral, prosody and energy based features)	PCA, LDA, auto-encoder and adversarial auto-encoder	SVM	57.88% (Max value of Unweighted Average Recall which has been used as performance metric)

5. FEATURE CLASSIFICATION

The ability to map a speech signal to a proper emotional category based on the suitable features extracted from the speech is the primary role of a classification system. Theory states that choosing a proper classifier is mostly based on past references. With classification majorly belonging to the category of supervised learning, features that are extracted, are fed to the classification model which adjusts its weight vector through training method to ensure that the model has been fitted properly. An activation function is then used to generate the output from the model which mapped each input to a predefined emotion class. According to the nature of activation function classifiers can be grouped into two different categories, which are linear classifier and nonlinear classifier. Linear classifiers will be able to classify accurately if the feature vectors are linearly separable, the exact opposite of which are non-linear classifiers which deals with features that are not linearly separable [1]. The authors in [1] have further stated that because in real life scenarios most of the feature vectors are not linearly separable, so a nonlinear classifier is comparatively a better choice. SVM (Support Vector Machine), GMM (Gaussian Mixture Model), MLP (Multi-Layer Perceptron), RNN (Recurrent Neural Network), KNN (K-Nearest Neighbors), HMM (Hidden Markov Model) are few non-linear classifiers that have been used widely as far as automatic speech emotion recognition systems are concerned.

With feature extraction, dimensionality reduction and feature classification being integral parts of ASER, the following block diagram depicts the flow of working of ASER as seen from Fig 2.

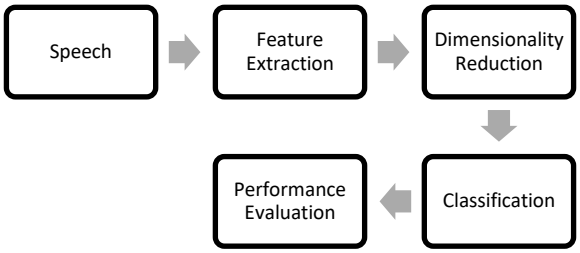

Fig 2: Flow of typical ASER system

Prior to feature extraction, data preprocessing may be required if the raw data set has inconsistencies, missing values, noisy data etc. Auto encoders may be used where dimensionality reduction is mentioned. Performance evaluation will help to understand whether the given combination of extracted feature set, dimensionality reduction (if used) and classifier chosen helps in the objective of creating an ASER with accuracy almost 100%.

Table 3. describes the work done in the literature survey of ASER by prominently highlighting the databases used, the features extracted, dimensionality reduction (if performed and the classifier chosen for the emotion recognition process. As stated by authors in [33], it may be difficult to choose the best classifier among the existing classifiers for emotion recognition purpose. As can be seen from Table 3., Support Vector Machine (SVM) has been commonly used by multiple authors as the chosen classifier for emotion recognition yet the accuracy achieved varies as high as 91% to as low as 57.88%. This iterates the fact that the success of the classifier as measured from the performance metrics, depends on not only type of features extracted but also on its quantity, density distribution of each emotional class and the language. The accuracy obtained using an acted database may differ from the accuracy obtained from a natural database. Choosing appropriate values for the parameters that define the classifier is seen to have significant impact on the accuracy of the classifier.

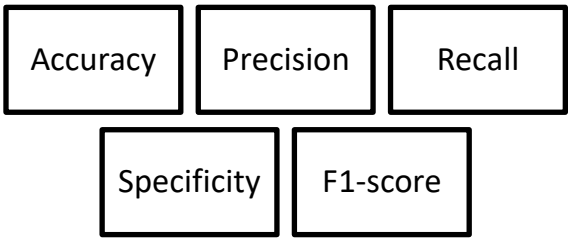

Fig 3: Commonly used performance metrics used in ASER

Several performance metrics have been used in the literature of ASER as can be seen from Fig. 3. These performance metrics are useful to validate the overall performance of the developed ASER model. Accuracy is the most common evaluation criteria used in measuring the ASER performance. It is defined as the ratio of the sum of true positive and true negative over the entire observation. The authors in [34] have further highlighted that there are two types of accuracy viz. weighted accuracy (WA) and unweighted accuracy (UA). WA considers the overall accuracy regarding all the utterances, whereas UA refers to the average score of the different accuracies for different emotions' classification. UA is used as the most important metric to evaluate the SER performance on imbalanced dataset. The authors further mentioned that Weighted Average Recall (WAR), known as the standard accuracy, and Unweighted Averaged Recall (UAR) have also been used to evaluate the performance of SER. Other performance metrics also been taken into consideration for evaluating the performance of ASER includes precision, specificity, recall and F1-score. Precision is defined as the ratio of true positive to all positives (i.e. sum of true and false positives) while recall is the ratio of the number of true positives to the total number of true positives and false negatives [35]. If precision helps understand what proportions of positive identifications are actually correct, then recall helps understand what proportion of actual positives are identified correctly [36]. Likewise, sensitivity is the ratio of true negatives to all the negative outcomes i.e. sum of true negatives and false positives while F1 score gives a combined idea about precision and recall metrics. These performance metrics can be easily calculated from the confusion matrix which is a table that summarizes how successful the classification model is at predicting examples belonging to various classes [32].

6. CONCLUSION

With human-machine interaction playing a significant role in industry and everyday life, in recent times, enabling trained machines to automatically carry out human-like tasks is making it adapt more and more to human needs and habits. Adding a touch of rationale and decision-making capability to machines is one such field of emotion recognition systems. With speech emotion recognition finding interest among several researchers, this paper has attempted to survey the different stages of an automatic speech emotion recognition system viz. feature extraction, dimensionality reduction and feature classification. There is no specific combination of these stages that will ensure a sure shot accuracy of 100% as a lot of factors matter viz. what kind of database is being used, nature of

emotion specific features and its quantity and the classifier chosen with appropriate activation parameters. Lastly this paper has also highlighted the use of auto encoders that have been introduced in recent times.

7. REFERENCES

[1] Basu, S., Chakraborty, J., Bag, A. and Aftabuddin, M. A review on emotion recognition using speech. 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2017, pp. 109-114, doi: 10.1109/ICICCT.2017.7975169.

[2] Ayadi, M., Kamel, M. and Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, vol. 44, Issue 3, pp. 572-587, Mar 2011

[3] Kotsakis, N., Liatsou, A., Dimoulas, C., Kalliris, G. Speech Emotion Recognition for Performance Interaction. Journal of Audio Engineering Society, vol. 66, Issue 6 pp. 457-467, June 2018

[4] Akçay, B., Oguz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116. 10.1016/j.specom.2019.12.001.

[5] Wang, C. et al. Speech emotion recognition based on multi-feature and multi-lingual fusion. Multimed Tools Appl, 81, 4897–4907 (2022). https://doi.org/10.1007/s11042-021-10553-4

[6] Patel, N., Patel, S., Mankad, S.H. Impact of autoencoder based compact representation on emotion detection from audio. Journal of Ambient Intelligence and Humanized Computing, 13, 867–885 (2022). https://doi.org/10.1007/s12652-021-02979-3

[7] Swain, M., Routray, A. and Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, 21, 93–120 (2018). https://doi.org/10.1007/s10772-018-9491-z

[8] Emotional Speech Databases. [Online]. Available: https://link.springer.com/content/pdf/bbm:978-90-481-3129-7/1.pdf

[9] Koolagudi, S. et al. (2009), IITKGP-SESC: Speech Database for Emotion Analysis. In: Ranka, S., et al. Contemporary Computing. IC3 2009. Communications in Computer and Information Science, vol 40. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03547-0_46

[10] Koolagudi, S., Reddy, R., Yadav, J. and Rao, K.S. IITKGP-SEHSC: Hindi Speech Corpus for Emotion Analysis. 2011 International Conference on Devices and Communications (ICDeCom), Mesra, India, 2011, pp. 1-5, doi: 10.1109/ICDECOM.2011.5738540.

[11] Shrishrimal, P., Deshmukh, R. and Waghmare, V. Indian Language Speech Database: A Review. Intl. Journal of Computer Applications, vol.47, no. 5, pp. 17-21, June 2012

[12] How to build your own Speech Emotion Recognition? [Online]. Available: https://vivoka.com/how-to-speech-emotion-recognition/

[13] Alex, S., Mary, L and Babu, B. Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features. Circuits Syst Signal Process 39, 5681–5709 (2020). https://doi.org/10.1007/s00034-020-01429-3

[14] Zhang, S., Zhang, S., Huang, T. and Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. in IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018, doi: 10.1109/TMM.2017.2766843.

[15] Bandela, S., and Kumar, T. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017, pp. 1-5, doi: 10.1109/ICCCNT.2017.8204149.

[16] Letaifa, L., Torres, M. and Justo, R. Adding dimensional features for emotion recognition on speech. 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2020, pp. 1-6, doi: 10.1109/ATSIP49331.2020.9231766.

[17] Alex, S. and Mary, L. Variational autoencoder for prosody-based speaker recognition. ETRI Journal, 45 (2023), pp. 678–689. https://doi.org/10.4218/etrij.2021-0377

[18] Xia, R. and Liu, Y. Using denoising autoencoder for emotion recognition. In Interspeech, pp. 2886-2889. 2013.

[19] Deng, J. Zhang, Z., Marchi, E. and Schuller, B. Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2013, pp. 511-516, doi: 10.1109/ACII.2013.90.

[20] Bhaswara, I.D. (2020) Exploration of autoencoder as feature extractor for face recognition system. [Online]. Available: https://essay.utwente.nl/83138/

[21] Chebbi, S. and Jebara, S. On the use of Pitch-based features for fear emotion Detection from Speech. 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Mar 2018

[22] Huang, C., Gong, W., Fu, W. and Feng, D. A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM. Mathematical Problems in Engineering, vol. 2014

[23] Khulage, A. and Pathak, B. Analysis of speech under stress using Linear Techniques and Non-Linear techniques for Emotion Recognition System. Jul 2012, https://doi.org/10.48550/arXiv.1207.5104

[24] LPCC Features [Online]. Available: https://link.springer.com/content/pdf/bbm%3A978-3-319-17163-0%2F1.pdf

[25] Shah, A., Kattel, M., Nepal, A. and Shrestha, D. Chroma Feature Extraction. Jan 2019

[26] Revathi, A., Sasikaladevi, N., Nagakrishnan, R. et al. Robust emotion recognition from speech: Gamma tone features and models. Int J Speech Technol 21, 723–739 (2018). https://doi.org/10.1007/s10772-018-9546-1

[27] Dmitrieva, E. and Nikitin, K. Design of Automatic Speech Emotion Recognition System. Proceedings of the International Workshop on Applications in Information Technology, pp. 47-50, 2015

[28] Schuller, B., Reiter, S. and Rigoll, G. Evolutionary feature generation in speech emotion recognition. IEEE International Conference on Multimedia and Expo. IEEE, pp. 5-8, 2006

[29] Kadiri, S., Gangamohan, P., Gangashetty, S. et al., Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference. Circuits Syst Signal Process 39, 4459–4481 (2020). https://doi.org/10.1007/s00034-020-01377-y

[30] Amartya, J.G.M., Kumar, S.M. Speech Emotion Recognition in Machine Learning to Improve Accuracy using Novel Support Vector Machine and Compared with Decision Tree Algorithm. Journal of Pharmaceutical Negative Results, vol. 13, no. 4, pp. 185-192, 2022

[31] Koduru, A., Valiveti, H.B. and Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology, vol. 23, pp. 45-55, Jan 2020

[32] Sahu, S. et al. Adversarial Auto-encoders for Speech Based Emotion Recognition. arXiv preprint arXiv:1806.02146 (2018).

[33] Partila, P., Voznak, M. and Tovarek, J. Pattern Recognition Methods and Features Selection for Speech Emotion Recognition System. The Scientific World Journal, vol. 2015, Article ID 573068, pp. 1-7, 2015. https://doi.org/10.1155/2015/573068

[34] Madanian, S. et al. Speech emotion recognition using machine learning — A systematic review. Intelligent Systems with Applications, vol. 20, Nov 2023

[35] Confusion Matrix, Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures [Online]. Available: https://www.linkedin.com/pulse/confusion-matrix-accuracy-precision-recall-f1-score-measures-silwal#:~:text=F1%20score%20is%20a%20weighted,have%20an%20uneven%20class%20distribution.

[36] Classification: Precision and Recall [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall