Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 08, August-2024

Emotion Recognition From Children Speech – A Review

Mrs. Dipti Kale
Department of Electronics and Telecommunication
D.J. Sanghavi College of Engineering
Mumbai, India

***Abstract*—Emotion recognition from speech is gaining attention as a research area from last two decades. Speech emotion recognition is one of the important element of Human-Machine Interactions (HMI). Recognition of emotion from speech deals with finding speakers emotions from the varied pallet of emotions in the human behavior. SER targets to extracting the various features from human preprocessed speech signal and then extracted features get classified in set of 7 basic emotions like Happiness, Sadness, Neutral, fear, boredom, disgust and anger. Most of the studies focuses on recognition of emotions in adults as the standard databases are available to experiment on. However most of applications of Speech emotion recognition like education, security, healthcare, psychology, cognitive sciences, gaming are targeting towards children as end user. So the need arises to have study of emotion recognition in children as well. In this paper brief overview is presented for the current state of research in this area focusing different techniques being used to detect emotional states in children vocal expressions. In addition to this, approaches for extracting linguistic features from children speech database and machine learning techniques with emphasis on classifiers are analyzed. This paper discusses the major application areas where emotion recognition in children speech is to be implemented.**

***Keywords*— Human-Machine Interaction, recognition of Emotion, SER, linguistic features, machine learning**

I. INTRODUCTION

Speech and Language is the primary way of communication in humans. While communicating using speech the information or message not only contains an informative message but it matters a lot the way how that message is being delivered. The message can be combined by both ways of communication that is verbal (speech) and nonverbal (facial expressions). In order to recognize emotions from speech a study conducted that confirmed that the facial expression of a speaker has information contents for about 55 percent of the effect, 38 percent of the impact is conveyed by voice intonation and 7 percent conveyed by the spoken words [1]. To recognize emotions from speech along with spoken words it becomes important to study voice characteristics of a speaker as well. Wide range of emotions can be detected in the adult speech as the clear utterance as well as other acoustics properties of voice are communicated effectively in adults. [2] So most of the researchers focus on speech emotion recognition (SER) systems in adults.

There are very limited studies on the effective recognition of emotions from children speech [3]. There are few problems for conducting research on the same as listed below.

The children are proficient in judging basic emotions from propositional and paralinguistic cues in speech [3][4]. It becomes more difficult to children to express emotions in acting speech if a child has not experienced that emotions before. The other reason for lesser research in children emotion recognition from speech identified as, Child's acoustic voice characteristics are different from that of an adult. Adults linguistic contains are more consistent than child. Child's voice clarity and consistency is low [4] so it becomes a major challenge in recognition of emotion in the children.

Most of the applications of emotion recognition directs towards children as an end user like education, security, healthcare, gaming with speech based interfaces. Therefore, accurate method for recognition of emotion in children will be beneficial to them for their implementation.

One of the effective parameter in recognition of emotion is an accurate database. Over the past few decades many emotional databases have been created on acted as well as non-acted corpus. Few of the creations are featured with children's emotional speech as well in English, German, French, Mandarin, Sesotho, Filipino and Russian languages. The variation in database is made not only in different languages but also with selection of different age group, group of children with language disorder as well. [5]

The work by [6] focused on creation of database that contains emotional speech in the Russian language of younger school age (8–12-year-old) children. By using classical machine learning algorithm like Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) the result described for recognizing four emotion classes, Neutral-Joy-Sadness-Anger. Result confirmed that for Russian language database proved to valuable resource in child and computer interactions.

The research dedicated to detection of emotional state of a child in conversational computer game for detecting "frustration", "politeness" and "neutral" attitudes [7]. Experiment performed on a voice activated computer game over 103 children of 7-14 years of age. Results confirmed that lexical information has more discriminative power than acoustic and contextual cues for detecting politeness, whereas context and acoustic features perform best for frustration detection. Their fusion has best classification

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 08, August-2024

results. Moreover 10-11 years old females have higher classification accuracy.

The research [8] investigated recognition of emotions from vocal expressions improves throughout childhood and part of adolescence. For that they tested 225 children (age 5-17) and 30 adults using vocal bursts expressing four basic emotions (anger, fear, happiness and sadness). Results verified by Mixed model logistic regression showed that during childhood affective prosody understanding improves, In the pivotal period of social maturation that is from early childhood to mid- adolescence vocal affect recognition improves.

The research paper [9] presented emotion recognition model to extract emotion from speech in children using FAU-AIBO children's speech emotion database. In the speech of children lot of lengthy fragments are present so instead of semantics the emotions show presence in the acoustic aspects like tones and timbers of voice. The human auditory speech based features are extracted using Mel Frequency Cepstral Coefficients (MFCC). 13 triangular filters are used to calculate corresponding mel frequency which reflects static characteristics of speech. Dynamic characteristics are found by describing static characteristics with change in time. Bi-directional Long Short-Term Memory (BiLSTM) extracts temporal features of speech. Convolutional neural network classifies CNN the MFCC based forty dimensional features in addition with BiLSTM extracting simultaneously features form frequency as well as time domain. Attention mechanism used on few frames for detecting emotions from children speech. This model improves accuracy of emotion recognition than established models such as LSTM - CNN and 2D-CNN-LSTM.

II. SPEECH EMOTION RECOGNITION SYSTEM

All the concept behind Human- Computer Interface is to develop a system that can communicate with humans through interpreting biological signals. The speech emotional recognition system intends to train a machine with relevant speech samples for identifying set of emotions and then while interacting with humans the machine can effectively identify emotions which show human-like intelligence. Speech emotion recognition system for processing children's speech, the emotional features suitable for children evaluation need to be extracted from audio signal, which is very important for the recognition of Children's emotions. The process followed while detecting emotions from the speech is depicted in Fig 1. The ways to find emotions can be from natural speech or from the acted one. For both of the ways the most important input is nothing but the database we refer for training and testing purpose.

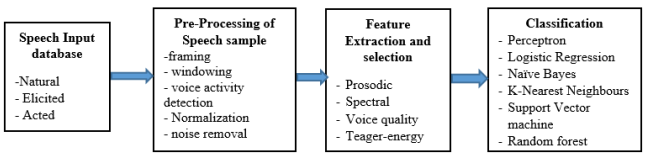

Fig.1 Speech Emotion Recognition system

To recognize emotions from speech the way featured in fig.1 is generalized approach whereas numerous approaches are present to display emotions. The selection of process depends on the objective to be achieved by the system.

A. Speech input database

For emotion recognition the other way is face recognition, as emotions recognized from facial expressions are universal. Whereas Vocal signatures shows variations in different category of emotions [10]. Based on method of data sample collection the database can be of three categories: Natural database – in this data samples are collected from spontaneous speech of real data. It includes recorded data from call center conversations, cockpit recordings, conversation between patient and doctor, emotional conversation at public place etc. Simulated (Acted) database compiled by collecting speech samples from experienced, trained actors which are referred as full blown emotions. Elicited or induced database fabricated where emotions were induced for example artificial emotional system without any knowledge about speaker.

Only few children related work is done in creation of database. Over past decade database is created in English language and also so of German, French, Spanish, Mandarin, Sesotho, Filipino, and Russian languages as well. some databases are designed in last few years are MESD (Mexican Emotional Speech Database) is uttered by 6 non-professional actors with average age of 9.83 years [11]. IESC-Child (Interactive Emotional Children's Speech Corpus) is induced database obtained as conversation between robots and 174 children spoke in Mexican Spanish language [12]. EmoReact (Multimodal Dataset for Recognizing Emotional Responses in Children) is spontaneous emotional dataset of 63 children generated from videos and gaming devices [13]. FAU-AIBO (Friedrich-Alexander-Universität corpus of spontaneous, emotionally colored speech of children interacting with Sony's pet robot Aibo) has corpus collected from recordings of 51 children of age 10 to 13, interacting with Sony's pet robot Aibo in German/English language [14]. ASD-DB (Autism Spectrum Disorder Tamil speech emotion database) is of spontaneous speech samples of age 5 to 12 years with autism spectrum disorder.[15].

B. Pre-processing of speech signal

Speech sample recorded from speaker contains signals other than the speech utterances of speaker which are clutter and need to be removed. For this number of steps followed to get speech corpus in enhanced format these steps are referred as pre-processing of speech sample. The process of dividing continuous speech samples into chunks of 20-30ms is framing is performed to get speech segments of fixed length. After framing each frame is passed through windowing function wherein discontinuities at the edge are reduced. Usually Hamming window is applied for this application. Every speech segment consists of unvoiced, voiced and silence part. To process only voiced part Voice activity detection is implement over segment of speech. Furthermore, signal is normalized to a maximum value of signal. Signal other than speech corpus affects or mask the accuracy of system so noise reduction is done over the recorded speech signal.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 08, August-2024

C. Feature extraction and selection

The accurate pattern recognition takes place only when extracted features and its representation is accurate. A key factor of accurate Speech emotion recognition system lies in preparation of appropriate database, selection of suitable feature set, design of proper classification method [16].

While processing speech from children voice data samples the most powerful feature needs to be selected which can give performance with any age group, any gender, any speaking style and with any language. Most well-known features are Acoustic features of prosodic extracting pitch, energy, duration. Spectral feature like MFCC, GFCC, LPCC, PLP, formants. Voice quality features like jitter, shimmer, harmonics to noise ratio, normalized amplitude quotient.

- Prosody:

Prosody of a speech segment indicates the flow of speech signal and consists of information like duration, intensity intonation and sound units.

Acoustic correlates of prosodic features included in features like pitch, energy, duration and their derivatives. Prosody features give linguistic level information as utterances are related with language only. Articularly movements are physical movements of muscles in throat related with prosody. Intensity in speech measured by fundamental frequency. [17]. So standard set of prosody features include pitch,

- Spectral:

Vocal tract features are called as spectral features. For analysis of different spectral features, a speech segment of length 20-30 ms is generated and Fourier transform of speech frame is taken. Further features derived from this like, MFCCs (Mel Frequency Spectral Coefficients), PLPs (Perpetual Linear Prediction Coefficients) and LPCCs (Linear Predictions Cepstral Coefficient. [18].

For emotion recognition form speech of children, the spectral feature extracted for static as well as dynamic characteristics of speech using MFCC (Mel Frequency Spectral Coefficient). MFCC performs good for static and dynamic features. [9] . MFCC deals with the extraction of feature from the speech from perspective of human auditory system. The speech samples obtained from human speech converted into short term frames, then these frames are converted into a new Mel frequencies spectrum. The 13 Mel filters are utilized to extract standard 13-dimensional MFCC. 13 dimensional MFCC reflect the static characteristics of speech. For dynamic characteristics of speech change of static parameters in time estimated. Static and dynamic features are combined to improve the accuracy of speech emotion recognition. Difference in MFCC reflects the dynamic characteristics of speech, and combines dynamic and static features. [9].

- Voice quality:

The glottal source qualities are defined by voice quality features. When the vocal tract is expiated to large extent by inverse filtering some changes observed in speech signal by that it may distinguish between various emotions further investigate by features like like harmonics to noise ratio (HNR), shimmer, and jitter. Voice quality features give compelling correlation between emotional content and voice quality of the speech.[18]

- Teager – energy

Teager energy operator (TEO) was introduced by Teager and Kaiser. In the research conducted it has been investigated that energy in different class of emotional speech is detected because of hearing process. When the class of emotional speech is Anger or stressful due to change of harmonics bands distribution change in fundamental frequency or critical frequency bands is observed. Stressful condition affects the speaker's muscle tension so sound produced from vocal tract modifies as airflow in vocal tract gets modified. From the sampled speech signal this correlation in the modified characteristics are quantified as features like frequency modulation variation, normalized TEO autocorrelation envelope area, critical band based TEO autocorrelation envelope area are introduced to help recognizing emotions from these varied characteristics of emotional speech in children.

While detecting effective features from speech of children, the features such as acoustic and linguistic give high variability, whereas clarity and consistency of child's voice is low. So features such as MFCC gives good measure of power in terms of Intensity and perceptive loudeness [9], acoustic features [19] proved to be good.

D. Classifiers

There are many classification methods like perceptron, logical regression, naïve bayes, k-nearest neighbours, support vector machine, random forest classical used to recognize emotions in speech. machine learning (ML) algorithms, such as Support Vector Machines (SVM), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Neural Networks (NN), and Multi-Layer Perceptron (MLP) are the most popular in recognizing emotions from speech. [20].

Based on number of emotions to be classified and classes of data the classifier algorithm needs to be selected. In case of emotion recognition from children speech a classifier to be selected must be effective in multidimensional spaces, it should give netter result even for relatively small size of database and with lower sensitivity.

Children's emotions expression from database FAU-AIBO children's speech emotion database are used to propose speech emotion recognition model (9). For recognizing emotions from emotional speech of children the frequency domain feature extracted is MFCC and classifier associated for classification of emotions is convolutional neural networks (CNN). For the mentioned system CNN gives enhanced results with MFCC features used. The long-term dependent learning features are classified well with Bi-directional Long Short-Term Memory BiLSTM which solve the problem of poor performance. Aattention mechanism is used for only a few frames contain emotional features in the children speech signal. In comparison speech emotion recognition models used in context like LSTM, CNN and 2D-CNN-LSTM for recognition of emotion from speech in children recognizes and improves the accuracy up to 71.6% on the FAU-AIBO children's speech emotion database.(9)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 08, August-2024

For the spectral features of speech most researchers use Convolution Neural Network (CNN) as classifier and MFCC to extract the features of the sound spectrum in the frequency domain, and classify the emotions. For the features like MFCC, the cyclic neural network is used for time domain feature recognition to complete the subsequent classification and recognition operations.[21]

Generative models such as Gaussian Mixture Model (GMM) when trained learns to cover feature space in a particular class of mentioned emotion. Discriminative classifier models such Artificial Neural network (ANN) or Support Vector Machine (SVM) learns the boundaries between subsequent feature classes in a discriminative way. According to experiments[19] SVM gives superior performance when used with features like MFCC as SVMs view data as two sets of vectors in a multi-dimensional space, and construct a separating hyperplane in that space.

In the children's speech, there are a lot of lengthy fragments are present and only a few feature frames contain emotional features. Therefore, in the process of implementing emotion recognition of children's speech, it is more effective to extract acoustic emotion features rather than semantic features.

III. CONCLUSION

The emotion recognition from children's speech system is a need of many of the applications intended to work with children as end user such as educational interface, healthcare, security, entertainment and many more. In this paper a precise analysis of this system is carried out. For emotion recognition from children speech the foremost important parameter is precise database. Elicited as well as acted databases are present for children speech in some of the languages for research ranging in different ages of children. The effective database can be used for providing data while training the module. Feature extraction is done after the speech signal has undergone pre-processing. The precise results are obtained from the best suited features. The features in children speech mostly investigated are as prosodic and spectral acoustic features such as formant frequencies, spectral energy of speech, speech rate and fundamental frequencies, and some feature extraction techniques like MFCC, LPCC, and TEO features performs best for classes of emotions while considering children speech. Two classification algorithms SVM and MLP for emotion recognition over classical Random Forest, K-NN, etc. classifiers [3] and also over different deep neural network classifiers are suitably used to recognize emotions. Deep learning classifier can be effective only with large and precise dataset. The development in emotional database with children will lead towards extracting more specific features and classifier like SVM can be grate to classify emotional classes with children.

REFERENCES

[1] C. C. Chibelushi and F. Bourel, "Facial expression recognition: Abrief tutorial overview," CVonline: On-Line Compendium ofComputer Vision, vol. 9, 2003
[2] Sandeep Kumar Panda,Ajay Kumar Jena,Mohit Ranjan Panda,Susmita Panda, "Speech emotion recognition using multimodal feature fusion with machine learning approach", Multimedia Tools and Applications (2023) 82:42763–42781
[3] Yuri Matveev 1,*, Anton Matveev 1 , Olga Frolova 1 , Elena Lyakso 1 and Nersisson Ruban ,"Automatic Speech Emotion Recognition of Younger School Age Children" Mathematics 2022, 10, 2373. https://doi.org/10.3390/math10142373
[4] Prashanth Gurunath Shivakumar*, Panayiotis Georgiou ,"Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendation," Computer Speech & Language 63 (2020) 101077
[5] List of Children's Speech Corpora. Available online: https://en.wikipedia.org/wiki/List_of_children%27s_speech_corpora
[6] Serdar Yildirim, Shrikanth Narayanan, Alexandros Potamianos, " Detecting emotional state of a child in a controversational computer game", Computer Speech & Language, 2011
[7] Serdar Yildirim,Shrikanth Narayanan,Alexandros Potamianos," Detecting emotional state of a child in a conversational computer game" Computer Speech & Language, Elsevier, January 2011
[8] Marie-HélèneGrosbras1, Paddy D. Ross2 & Pascal Belin, "Categorical emotion recognition from voice improves during childhood and adolescence" SCIENTIFIC REPORTS | (2018) 8:14791 | DOI:10.1038/s41598-018-32868-3
[9] Guitao Cao1 , Member, IEEE, Yunming Tang1 , Jiyu Sheng1 , and Wenming Cao*2, Member, IEEE," Emotion Recognition from Children Speech Signals Using Attention Based Time Series Deep Learning" 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
[10] R. Munot and A. Nenkova, ''Emotion impacts speech recognition performance,'' in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Student Res. Workshop, 2019, pp. 16–21, doi: 10.18653/v1/n19-3003
[11] Duville, M.M.; Alonso-Valerdi, L.M.; Ibarra-Zarate, D.I. Mexican Emotional Speech Database (MESD). Mendeley Data 2021, V2, 1644–1647
[12] Pérez-Espinosa, H.; Martínez-Miranda, J.; Espinosa-Curiel, I.; Rodríguez-Jacobo, J.; Villaseñor-Pineda, L.; Avila-George, H. IESC-Child: An Interactive Emotional Children's Speech Corpus. Comput. Speech Lang. 2020, 59, 55–74.
[13] Nojavanasghari, B.; Baltrušaitis, T.; Hughes, C.; Morency, L. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI), Tokyo, Japan, 12–16 November 2016; pp. 137–144.
[14] Batliner, A.; Hacker, C.; Steidl, S.; Nöth, E.; D'Arcy, S.; Russell, M.; Wong, M. "You Stupid Tin Box"—Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, 26–28 May 2004; pp. 171–174.
[15] Ram, C.S.; Ponnusamy, R. Recognising and classify Emotion from the speech of Autism Spectrum Disorder children for Tamil language using Support Vector Machine. Int. J. Appl. Eng. Res. 2014, 9, 25587–25602.
[16] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, ''A review on emotion recognition using speech,'' in Proc. Int. Conf. Inventive Commun. Comput. Technol., 2017, pp. 109–114
[17] 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.
[18] M. Swain, A. Routray, and P. Kabisatpathy, ''Databases, features and classifiers for speech emotion recognition: A review,'' Int. J. Speech Technol., vol. 21, no. 1, pp. 93–120, Mar. 2018
[19] Tim Polzehl 1 , Shiva Sundaram 1 , Hamed Ketabdar 1 , Michael Wagner 1,2, and Florian Metze 3," Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features" Interspeech, 6-10 September, Brighton UK
[20] S. Asiri, ''Machine learning classifiers,'' Towards Data Sci., Jun. 2018. [Online]. Available: https://towardsdatascience.com/machine-learningclassifiers-a5cc4e1b0623
[21] C. Hema a , Fausto Pedro Garcia Marquez," Emotional speech Recognition using CNN and Deep learning techniques", Applied Acoustics 211 (2023) 109492, 2023 Elsevier Ltd.