Nat. Volatiles & Essent. Oils, 2021; 8(4): 555-560

NVEO
Natural Volatiles & Essential Oils

Emotion Recognition from Speech Signal through DWT - LPC & Convolution Neural Network

Niranjan Samudre [1], Dr. Prashant Sharma[2], Dr. Chandan Singh Rawat[3]

[1]*Research Scholar, Department of Electronics and Communication Engineering, PAHER University, Udaipur, Rajasthan, India. Email- nasamudre@gmail.com*

[2] *Department of Computer Engineering, PAHER University, Udaipur, Rajasthan, India. Email- prashant.sharma@pacific-it.ac.in*

[3] *Department of Electronics and Telecommunication Engineering, VESIT, Mumbai, Maharashtra, India. Email- csrawat3@gmail.com*

Abstract

Tremendous development in microcircuit technology and the Web of Things has fuelled growth in virtual personal assistant devices and systems like Alexa, Siri, and Google Assistant. These virtual assistant devices receive commands through speech signals and are trained to deliver necessary actions quickly and accurately. But these virtual assistant devices are fairly trained to receive speech commands however have to enhance their emotion recognition ability to semantically method request from the user. In this work, implementation of speech emotion recognition through Discrete Wavelet Transform (DWT) – LPC and convolution neural network (CNN) is tried. Speech signals obtained from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset are processed, transformed using DWT, reduced-order linear predictive coding coefficient, and convolution neural network (CNN). The convolution neural network was enforced for training, classification, and recognition of emotion. Relatively higher recognition accuracy was obtained through DWT - LPC & Convolution Neural Network as compared with different ways revealed within the literature.

Keywords – Emotion Recognition, Discrete Wavelet Transform, Convolution Neural Network

I. INTRODUCTION

Report generated by the UN has foretold a tremendous increase in the figure of individuals interacting through virtual assistant systems in the next few years [1]. Virtual assistant devices like Alexa, Siri, and Google Assistant receive commands through speech signals and are trained to deliver necessary actions quickly and accurately. Speech signal has been a major source of communication everywhere the globe through numerous wired and wireless communication systems for decades. Linguistic info through verbal communication using numerous languages like English, French, Chinese, Hindi, and paralinguistic info through emotions, pitch, tone, gestures, expressions, etc are major ways used for social communication of knowledge. It's largely ascertained that the paralinguistic methodology of human action info is simpler in understanding the status of the sender through numerous emotions and gestures [2]. "Ryerson Audio-Visual information of Emotional Speech and Song (RAVDESS) dataset may be valid multi-modal information of emotional speech and song".

In this section, the implementation of speech emotion detection through DWT – LPC and convolution neural networks are mentioned. Speech signals obtained from "Ryerson Audio-Visual information of Emotional Speech and Song (RAVDESS) dataset are processed, transformed using DWT, reduced-order linear predictive coding coefficient, and convolution neural network (CNN). Improved illustration of speech signal was achieved through LPC acquired from discrete wavelet transform decomposed sub-bands in dyadic fashion. The convolution neural network was enforced for coaching, classification, and recognition of emotion.

II. RELATED WORKS

For speech signal process and recognition, the audio signals are obtained through microphones that are converted into electrical signals. Pre-processing of the speech signals to get rid of background and silence at the beginning and finish of the signal is crucial. Filtering, windowing, and framing are applied within the pre-processing stages of the speech signal recognition systems. Spectral subtraction and Wiener filtering are used for filtering noise from the corrupted speech signals and obtained clean signals for any process. "Feature extraction" is one of the many steps within the speech recognition systems and "convolution neural network". Speech signal feature extraction was performed using wavelet transform and wavelet packet tree [3]. Mel filtered sub-band energies were replaced with energies of speech signal was obtained from frequency decomposed sub-bands. it's ascertained that speech signal contains each high and low-frequency parts, high frequencies are present within the starting whereas low-frequency parts are present until the tip or for a very long time. Thus variable time-frequency processing of wavelet transform is considered as a viable option for processing such non-stationary speech signals. In the beginning, the voice signal is split into four frequency bands uniformly or in a dyadic manner to resolve the parameters of the speech wave. During this investigation voice wave is separated into sub-bands by "DWT" then "LPC coefficients" are calculated that are concatenated to offer "wavelet decomposed LPC features" [4]. "LPC" has bound benefits such as its capability of higher differentiation among the words that have separate vowel sounds [5]. However within "LPC", it's perceived that "DWT" is acceptable to represent the particulars of the unvoiced sound portion of the speech. Collective edges of "LPC" and "DWT" are also explored by application of the "LPC" technique on every sub-band concatenated in traditional and two fashions through wavelet decomposition [6]. "Convolution neural networks" (CNN) [9,10,11] are thought better for automatically learning relevant features input training dataset. "CNN" employs shared weighted kernels to see the associated composition of information. Max-pooling is introduced into "CNN" to scale back quality and choose applicable elevated "dimensional features" that are educated for a spread of tasks like "classification, segmentation and then on [7,12, 13]. Since wavelet decomposition of the speech signal may be befittingly applied using convolution operation and completely different filter coefficients (weighted kernels) for numerous frequencies range. "CNN's" superior performance might facilitate to model the underlying feature with efficiency from "short time-frames leading to a state of the art performance in speech-based" sentiment acknowledgment theme.

III. PROJECTED METHODOLOGY

Initially, pre-processing of voice waves was performed that enclosed "pre-emphasis, framing and windowing". The input voice wave was "pre-emphasis using first-order high pass digital filter to stress high-frequency energy using" (1)

$$y[n]=x[n]-0.9\,x[n-1] \quad (1)$$

"The output of the pre-emphasis filter is blocked to N samples followed with windowing using a hamming window to minimize the signal discontinues at starting and finish of the frame".

$$h[n]=0.54-0.46\cos(2\pi n/(N-1)) \quad (2)$$

where n = zero, 1, 2, …, N-1.

The wavelet decomposition of speech signal contains the data of various frequency scales that help in determining the corresponding band. so the LPC feature vector obtained from frame i by applying two methodologies may be pictured by (3)

$$f_i=[a_{A3}, a_{D3}, a_{D2}, a_{D1}] \quad (3)$$

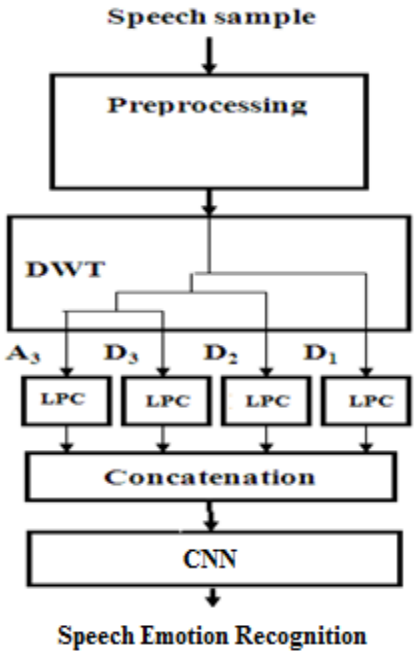


Figure 1. Speech emotion detection system using DWT, LPC, and CNN

Convolution neural network (CNN) has incontestable higher performance for speech emotion recognition systems [8]. The projected CNN consists of 2 convolution layers. Initial one is the absolutely connected layer, and second maybe a Softmax layer. Convolution layer one consists of 128 kernels whereas the convolution layer consists of 256 kernels. Figure 1 depicts the methodology adopted for implementing speech emotion detection system exploitation DWT, LPC and CNN. Figure 2 depicts the projected CNN design adopted in the speech emotion recognition system.

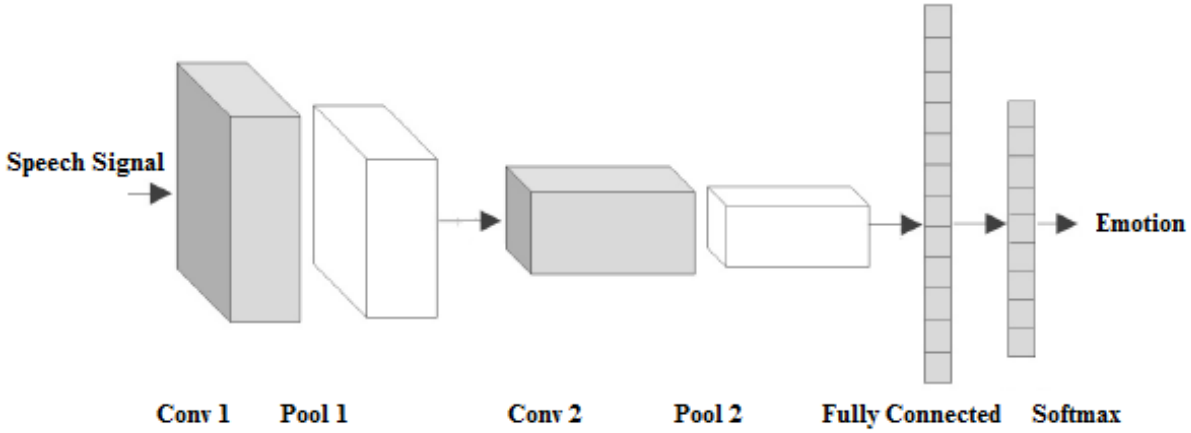
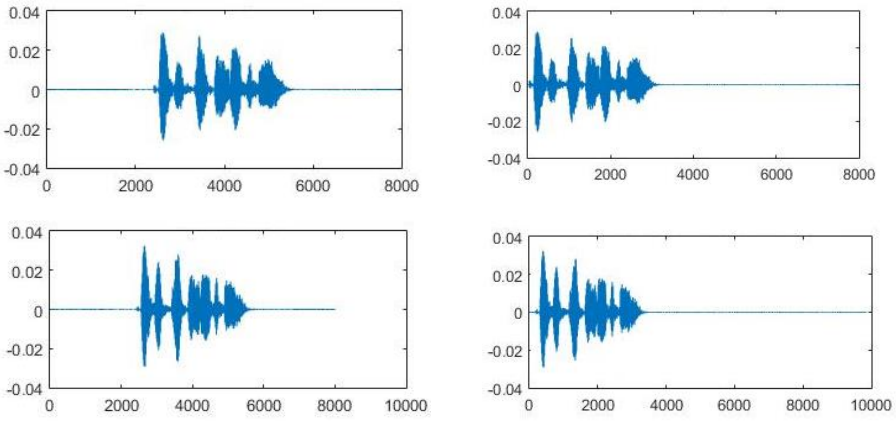


Figure 2. CNN design adopted for speech emotion recognition system

IV. IMPLEMENTATION AND RESULTS

"Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset consists of twenty-four skilled actors, every performing arts 104 distinctive vocalizations with emotions that include: happy, sad, angry, fearful, surprise, disgust, calm, and neutral". Each actor act out a pair of statements for every emotion: "Kids are talking by the door" and "Dogs are sitting by the door." "These statements were additionally recorded in 2 different emotional intensities, traditional and powerful, for every emotion, apart from neutral (normal only), actors recurrent every vocalization double. There are a complete of 1440 speech utterances and 1012 song utterances. Every audio file contains a 7-part numerical symbol each denoting the modality, vocal channel, emotion, emotional intensity, statement, repetition and also the actor respectively".



(a) Input speech signal

(b) Output after pre-processing

Figure.3. Speech signal after pre-processing

"Pre-emphasis of speech signal was obtained using filter coefficients and so the signal was divided into frames of 250 ms having 2200 samples". Fig 3 shows the speech signal after pre-processing. "Speech frame was decomposed into sub-bands using 3-level Daubechies's wavelet transform. Prediction

coefficients with sixth-order LPC were estimated from the sub-bands. The prediction coefficients estimated from the sub-bands were then concatenated to create vector fi (for ith frame)". Presentation of these structures was tested using CNN on the database. Table 1" gives the performance of DWT LPC and CNN in terms of percentage recognition rate". Eight emotions have been used. Database contains total 180 signals of each emotion. The proposed algorithm gives highest recognition rate of 82 for angry signal and lowest rate for neutral signal. Relatively higher recognition accuracy was obtained through DWT - LPC & Convolution Neural Network as compared with different ways like SVM, Random Forest, and HMM.

Table -1 Performance parameters

Emotion Recognition Rate (DWT + LPC + CNN)								
Emotions	**Neutral**	**Calm**	**Happy**	**Sad**	**Angry**	**Fearful**	**Disgust**	**Surprise**
Total signals in database	**180**	**180**	**180**	**180**	**180**	**180**	**180**	**180**
Recognized signals from database	**114**	**134**	**146**	**143**	**148**	**139**	**126**	**144**
Percentage recognition rate	**63**	**74**	**81**	**79**	**82**	**77**	**70**	**80**

V. CONCLUSION

Here, implementation of "emotion recognition from speech signals" through DWT – LPC, and CNN is attempted. Speech signals obtained from the "Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" dataset were pre-processed, transformed utilizing "discrete wavelet transform", "reduced-order linear predictive coding coefficient" and "convolution neural network (CNN)". The Convolution neural network was enforced for training, classification, and recognition of emotion. Eight emotions have been used. Database contains total 180 signals of each emotion. The proposed algorithm gives highest recognition rate of 82 for angry signal and lowest rate for neutral signal. Relatively higher recognition accuracy was obtained through DWT - LPC & Convolution Neural Network as compared with different ways like SVM, Random Forest, and HMM.

REFERENCES

United Nations Educational, Scientific, and Cultural Organization. (2019), "I'd blush if I could: closing gender divides in digital skills through education, "website: http://unesdoc.unesco.org/images/0021/002170/217073e.pdf.

Yamashita, Y. (2013), "A review of paralinguistic information science for natural speech," Acoustical science Science and Technology, 34, (2), pp: 73–79. DOI: 10.1250/ast.34.73

Jagannath H Nirmal, Mukesh A Zaveri, Suprava Patnaik and Pramod H Kachare (2013), "A novel voice conversion approach using permissible wavelet decomposition," Springer EURASIP Journal on Audio, Speech, and Music process, pp: 1 – 10.

M Krishnan, CP Neophytou, G Prescott, "Wavelet remodel speech recognition using vector quantization, dynamic time warping, and artificial neural networks," in International Conference On Spoken Language Processing, Yokohama, Japan, pp: 1191–1193.

Firoz Shah, RajiSukumar A., and BabuAnto. P (2010), "Discrete wavelet-Transforms and Artificial Neural-Networks for SER," International Journal of Computer Theory and Engineering, vol. 2, no. 3, pp: 319-322.

Paul A.K., Das D., Kamal M.M., "Bangla SR System using LPC and ANN," 7th IEEE International Conference on Advances in Pattern Recognition", pp: 171 – 174.

Ji, S., Xu, W., Yang, M., & Yu, K. (2010), "3D Convolutional Neural Networks for act Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence," 35, pp: 221-231.

Kim, Y., Lee, H., Provost, E.M., "Deep-learning for strong feature generation in audio-visual emotion recognition," Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, BC, Canada, pp: 3687–3691.

Sharma, P., Saxena, K., & Sharma, R. (2016). Heart disease prediction system evaluation using C4.5 rules and partial tree doi:10.1007/978-81-322-2731-1_26.

Aashkaar, M., & Sharma, P. (2016). Enhanced energy efficient AODV routing protocol for MANET. Paper presented at the International Conference on Research Advances in Integrated Navigation Systems, RAINS 2016, doi:10.1109/RAINS.2016.7764376.

Saxena, R., Johri, A., Deep, V., & Sharma, P. (2019). Heart diseases prediction system using CHC-TSS evolutionary, KNN, and decision tree classification algorithm doi:10.1007/978-981-13-1498-8_71.

Sharma, P., Alshehri, M., Sharma, R., & Alfarraj, O. (2021). Self-management of low back pain using neural network. Computers, Materials and Continua, 66(1), 885-901. doi:10.32604/cmc.2020.012251.

Alshehri, M., Sharma, P., Sharma, R., & Alfarraj, O. (2021). Motion-based activities monitoring through biometric sensors using genetic algorithm. Computers, Materials and Continua, 66(3), 2525-2538. doi:10.32604/cmc.2021.012469.