Periodica Polytechnica Electrical Engineering and Computer Science, 65(3), pp. 262–278, 2021

Extraction of Voiced Regions of Speech from Emotional Speech Signals Using Wavelet-Pitch Method

Lakshmi Srinivas Dendukuri[1*], Shaik Jakeer Hussain[1]

[1] Department of Electronics and Communication Engineering, Vignan's Foundation for Science, Technology and Research, Vadlamudi, Guntur 522213, Andhra Pradesh, India
[*] Corresponding author, e-mail: dls_ece@vignan.ac.in

Received: 02 December 2019, Accepted: 18 July 2020, Published online: 13 July 2021

Abstract

Extraction of voiced regions of speech is one of the latest topics in speech domain for various speech applications. Emotional speech signals contain most of the information in voiced regions of speech. In this particular work, voiced regions of speech are extracted from emotional speech signals using wavelet-pitch method. Daubechies wavelet (Db4) is applied on the speech frames after downsampling the speech signals. Autocorrelation function is performed on the extracted approximation coefficients of each speech frame and corresponding pitch values are obtained. A local threshold is defined on obtained pitch values to extract voiced regions. The threshold values are different for male and female speakers, as male pitch values are low compared to the female pitch values in general. The obtained pitch values are scaled down and are compared with the thresholds to extract the voiced frames. The transition frames between the voiced and unvoiced frames are also extracted if the previous frame is voiced frame, to preserve the emotional content in extracted frames. The extracted frames are reshaped to have desired emotional speech signal. Signal to Noise Ratio (SNR), Normalized Root Mean Square Error (NRMSE) and statistical parameters are used as evaluation metrics. This particular work provides better SNR and Normalized Root Mean Square Error values compared to the zero crossing-energy and residual signal based methods in voiced region extraction. Db4 wavelet provides better results compared to Haar and Db2 wavelets in extracting voiced regions using wavelet-pitch method from emotional speech signals.

Keywords

emotional speech, wavelets, autocorrelation, pitch, thresholding

1 Introduction

Speech is a basic medium of communication for human-human, human- machine interactions. In latest technology, most of the devices are using voice recognition as a feature, for humans to communicate with them. Speech signals not only provide message content in it, but also provide the information of the speaker. This is one of the important aspects of speech, which carries the speaker's information such as age, gender and emotion. Speech based recognition algorithms mainly involves extraction of features from speech signals. The predominant selection of features, involving source features, system features and perceptual features increases the recognition rate.

Speech production process involves various organs as lungs, vocal cords, nasal cavity and lips. All these organs act as time varying filters in production of particular sound to communicate [1]. At first, air is inhaled during normal breathing, and lungs expel the air and are chopped into periodic pulses by the vibration of vocal cords. The periodic pulses of air can enter the vocal cavity or nasal cavity. If the velum is closed, air flow enters the vocal cavity and is filtered by mouth and lips which act as time varying filter for speech production. The speech produced from vocal cavity is usually periodic in nature. If velum is open, air flow enters the nasal cavity and speech sounds produced from nose are usually non-periodic and are random in nature. The complete speech production involves the combination of periodic and non-periodic regions of speech. Speech signals are usually quasi stationary or quasi periodic in nature, having periodicity only for short durations. This property makes speech signal to be divided as voiced regions and unvoiced regions of speech. The speech sounds produced from vocal cavity are periodic in nature and called as voiced regions and the speech sounds produced from nasal cavity are random

Cite this article as: Dendukuri, L. S., Hussain, S. J. "Extraction of Voiced Regions of Speech from Emotional Speech Signals Using Wavelet-Pitch Method", Periodica Polytechnica Electrical Engineering and Computer Science, 65(3), pp. 262–278, 2021. https://doi.org/10.3311/PPee.15373

in nature and called as unvoiced regions of speech. Speech production usually involves voiced regions which are periodic, unvoiced regions which are random and silence regions which are zero values without any sound [2]. Any speech usually comprises of these three components.

2 Related works

Jo and Kim [3] proposed extraction of voiced and unvoiced speech components from speech based on residual signal. The analysis of speech signals is done using pitch and LPC, and is modeled using residual signal. Extraction of voiced speech using residual signal provides poor results in emotional speech signals because modeling of new speech signal based on analysis loses emotional information.

Shaojun et al. [4] proposed voice activity detection using wavelet transform. Voice activity detection is performed by comparing the spectral energy in sub-bands. The spectral energy in sub-bands for voiced region detection method proposed by Jiang Shaojun is implemented on simulated signals, assuming the four frames of noise. In real emotional speech signals, the number of unvoiced frames will be more than four frames and produces very low SNR values. The main parameter, pitch varies from emotion to emotion due to changes in vocal tract. Sub-band energy-based method extracts very few voiced frames in emotions like happy, disgust and surprise compared to the proposed wavelet-pitch method because of considering amplitude parameter, without any frequency parameter as pitch.

Upadhyay and Pachori [5] proposed voiced and unvoiced detection using variational mode decomposition technique. This method separates the fundamental frequency component from the actual speech signal and uses its envelope for voiced region detection. Energy is considered as the thresholding parameter for voiced region detection. This method gives good results in voiced region detection without prior information of pitch. In this method prior information of the number of modes is required.

Erdol et al. [6] proposed a novel method in synthesizing the missing packets of speech based on previous packets short term energy and zero crossing for transmission. The performance metric used in this method is spectral magnitude measure, SNR-M. The disadvantage of this method is, SNR-M goes to infinity which is descriptive measure as the ear is insensitive to phase.

Sun et al. [7] proposed robust voice activity detection by an energy based frame selection algorithm to indicate speech activity at the frame. To extract relevant speaker's speech frames, Voice Activity Detection (VAD) Tags and Automatic Speech Recognition (ASR) transcript Tags are provided by NIST. This approach provides an efficient way to select high quality speech frames and the relevant speaker's voice for speaker recognition.

Haigh and Mason [8] proposed an algorithm for voice activity detection based on the cepstral features. This particular work provides high degree of independence between the noise and the voiced speech using thresholding technique. The noise code book is used for Euclidean distance measure for voice detection.

Jalil et al. [9] proposed short term energy, short term magnitude, zero crossing rate and autocorrelation based voice activity detection. They verified the autocorrelation of voiced speech is also periodic but autocorrelation of unvoiced speech is random in nature. The selection of window and window size for voiced speech are also discussed in this paper. Rectangular window is used for measuring short term magnitude. Voiced regions have higher magnitude compared to unvoiced regions of speech.

Marques and Almeida [10] proposed sinusoidal modeling of speech for voiced region extraction. The harmonic model provides good results but cannot handle the fast variations of pitch. In this paper they discussed the basic sinusoidal modeling of speech and quadrature sinusoidal modeling of speech signals for voiced region extraction.

Wu and Wang [11] proposed a novel method using wavelet transform and Teager energy operator for extraction of voiced regions of speech. The speech signals are decomposed into four sub-bands to have robust envelope of voiced regions, Teager energy operator is applied on wavelet sub-band signals. Autocorrelation function is applied on the Teager energy operated signal for voiced region extraction. Teager energy operator is defined as the difference the square of present sample amplitude to the product of past and future sample amplitudes. In neutral and calm emotional speech, the detail coefficients are related to unvoiced speech and are boosted up considering as partial voiced frames using Teager energy operator. The proposed wavelet-pitch method provides better results compared to Teager energy operator method, by analysing the periodicity of analysis coefficients of speech using pitch parameter.

Asgari et al. [12] proposed an algorithm for voice activity detection based on the spectral entropy of a signal. This paper considers the assumption of both signal and noise entropy magnitudes as Gaussian. The speech signal spectrum is more organised in voiced regions than in unvoiced regions of speech assuming Gaussian in nature. This method provides better results even in low SNR values of speech signal.

Swee et al. [13] proposed a method for voice activity detection in speech signals based on short term energy and this method is used for detection of pitch in the speech frames. Pitch detection results are evaluated for testing the accuracy of the proposed method. In this paper depending on the gender of the speaker, voiced regions of speech are detected. Mean energy of speech signal is considered as metric for gender classification.

Jafer and Mahdi [14] proposed wavelet based voiced and unvoiced region of speech using zero crossing rate and short term energy. This paper performed zero crossing and energy based thresholding in the wavelet domain. The method proposed by Jafer and Mahdi provides poor results on emotional speech like angry and sad. In some voiced frames, the average energy in the wavelet domain is high but the median of zero crossings is also high or average energy in wavelet domain is low and median of zero crossings is low. These voiced frame values do not satisfy both threshold criteria, it is considered as unvoiced frame. The proposed wavelet- pitch method uses the pitch parameter of frames, which checks for periodicity property of voiced regions and uses threshold as percentile, which is more resistant to the outliers of frame pitch values.

Various works have proposed voiced region extraction from clean speech or in noisy environment speech signals [15]. Tanmoy et al. [16] proposed the endpoint detection in speech using wavelet convolution approach. The wavelet convolution speech endpoint detection algorithm uses the wavelet transform to decompose the signal and uses entropy of coefficients for extraction of voiced speech. It uses two thresholds based on the gender of the speaker, because male speaker has higher loudness compared to female speaker. Hamzenejadi et al. [17] proposed extraction of pitch and formant frequencies in the wavelet domain. The speech signal is decomposed to eight scales and zeroing the higher scale coefficients after fifth level of decomposition, produces formant frequencies and zeroing the lower wavelet coefficients up to fifth and above eight scales, provide pitch frequency.

Humans communicate daily in emotions as happy, sad, neutral etc. Extraction of voiced regions from emotional speech signals is little challenging because applying previous methods on emotional speech degrades the emotional content. It involves the modeling of time varying filters such as vocal tract, mouth and lips to vary from emotion to emotion in production of emotional speech signal. Dendukri and Hussain [18] proposed the extraction of feature set from voiced regions of speech for speech emotion recognition application. Wavelet transform is applied on the speech signal, hard and soft thresholding are applied on detail coefficients for the extraction of voiced speech. It provides voiced speech but transition regions between voiced and unvoiced frames are not preserved because of using hard thresholding technique on the detail coefficients. Considering the energy of the signal in sub-bands, using of auto correlation signal and thresholding of detail coefficients, uses the amplitude parameter of the signal. Since emotional speech varies from emotion to emotion in terms pitch as a main parameter because the vocal tract resonates at various fundamental frequencies, that differ from emotion to emotion. For angry and disgust emotions, amplitude is high but major variations are present in pitch. For neutral and happy emotions, calm and sad emotions, fearful and surprise emotions have nearly similar amplitudes but variations are tracked in pitch parameter.

Busso et al. [19] proposed the relation between the pitch and emotions w.r.t the neutral emotion. Gharavian et al. [20] proposed the influence of pitch and pitch slope in emotional speech recognition applications. Breitenstein et al. [21] used acoustic feature to analyse the change in pitch values in perception of emotional speech. The proposed wavelet-pitch method uses the pitch parameter in classification of voiced speech to preserve the emotional content in extracted voiced speech.

In this paper a novel algorithm is proposed based on wavelet-pitch method, for extraction of voiced regions from emotional speech signals and to preserve the transition regions of voiced and unvoiced frames which carry emotional content. In this work, wavelet-pitch method, percentiles of pitch are used as thresholds to extract voiced regions of speech. Logical smoothing is done on all extracted frames to preserve the information in all voiced and transition regions of speech. This particular work helps in extraction of predominant features for speech recognition, mainly in speech emotion recognition applications.

3 Experimental data

As a part of our research work, Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS), recorded by SMART Lab Ryerson University is considered [22]. This particular database contains 24 speakers, out of which 12 are male and 12 are female speakers. Each speaker utters sentences in different emotions. This particular dataset contains Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprise speech signals which are sampled at 48 Khz. The speakers' in the

database uttered two sentences in eight emotions two times in English language. The sentences that are uttered by all male and female speakers are

1. "*Kids are talking by the door.*"
2. "*Dogs are sitting by the door.*".

In this particular work, 1000 emotional speech signals, comprising of eight emotions in two sentences, uttered by 12 male and 12 female speakers' in English language are considered for analysis.

4 Method

Emotional speech signal is first decimated by a factor of 3 to have sampling frequency of 16 Khz. This helps in removing the redundant data in speech signals. The decimated speech signals are normalized to have maximum amplitude in the range of −1 and 1. Speech signals are divided into frames of 20 msec duration, since the speech signals are quasi stationary and analysis can be done in frames. The framing is done with an overlap of 50 % to avoid any loss of information from frame to frame transition.

Discrete wavelet transform is applied on the obtained speech frames, in order to extract the low frequency components and the high frequency components of the speech signal using Daubechies wavelet (Db4) [23]. In this work, one stage decomposition of speech frames is done to extract an approximation coefficient and detail coefficients. Autocorrelation function is applied on the extracted approximation coefficients of each speech frame to calculate the pitch. The obtained pitch values are scaled by a factor of 2, as the pitch values are taken from approximation coefficients of speech, which are also decimated by a factor of 2.

Thresholding technique is applied on the obtained pitch values for extraction of voiced regions of speech. In this particular work, 20^{th} percentile and 60^{th} percentile of the extracted pitch values from speech frames are considered as thresholds Th_1 and Th_2 for male speakers and 35^{th} percentile and 75^{th} percentile of pitch values as thresholds Th_1 and Th_2 for female speakers. Any pitch value of a particular frame which range within 20^{th} percentile and 60^{th} percentile are considered as voiced speech frame for male speakers and which range within 35^{th} and 75^{th} percentile are considered as voiced frames for female speakers. Since male speakers have low pitch values compared to the female speakers, the change in threshold values is considered. Logical addition of the speech frames is done considering the previous voiced frames, to preserve the transition from voiced speech to unvoiced speech which contain emotional content. The extracted speech frames are arranged in an order based on the index of the speech frames. The proposed method is represented in Algorithm 1.

4.1 Decimation

The speech signals taken from data set are sampled at 48 Khz and to reduce the redundancy of data, they are downsampled by a factor 3 to have sampling frequency as 16 Khz.

4.2 Normalization and framing

Many speech samples are represented on different units of scale. To avoid the problems of handling various units of data, speech samples of various emotions are normalized and all are brought to a uniform scale, providing accurate values. Since speech signals are quasi periodic, with small durations of periodic nature, analysis can be done by dividing the entire speech signal into frames with a duration 20–30 msec. The division of frames is done with a 50 % of overlap of previous frame samples. Hamming window is applied to each frame, to make the energy concentrated in the center of the frame.

4.3 Discrete wavelet transform

Analysis of quasi periodic or non-stationary signals using short time Fourier transform provides better results than Fourier transform [24–25]. Since short time Fourier transform has fixed window size, better results can be obtained using wavelet transform, with window size varing in analysis of non-stationary signals. Since speech is quasi-periodic, analysis of speech signals using wavelet transform provides better results than FFT because of dilation and translation property of wavelets. Wavelet transform

Algorithm 1 Wavelet-Pitch method for Voiced Speech Extraction

1: Emotional speech signal is read from speech database.
2: Downsampling the input speech signal by a factor 3 to have sampling frequency as 16 Khz.
3: Dividing signal in frames with 50 % overlap.
4: Applying Discrete Wavelet Transform.
5: Autocorrelation on the approximation coefficient of speech frame.
6: Calculation of pitch from Auto-correlated signal.
7: Calculation of Threshold values from Pitch values.
8: Selection of Voiced frames based on Threshold values.
9: Logical shaping of extracted speech frames based on index of speech frames.
10: Final emotional speech signal comprising of voiced region of speech with emotional content.

provides multi-resolution analysis, with its varying window length in analysis of speech signals. The basis of wavelet function analyses the speech signal by the property of dilation and shifting, providing approximation coefficients and detail coefficients of speech signal. Approximation and detail coefficients are extracted from the original speech signal by decimating with a factor of 2. Approximation coefficients provide most of the low pass information and detail coefficients represent high pass information. The basic mother wavelet is given in Eq. (1):

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right) \quad a \in \Re,\ \tau \in \Re, \tag{1}$$

where $\Re$ indicates the set of real numbers, a is dilation factor and τ is shifting factor. The approximation coefficients and detail coefficients of signal $f(k)$ are given by Eqs. (2) and (3):

$$A_j = \sum_{n=0}^{\infty} f(k)\frac{1}{\sqrt{2^j}}\Phi\left(\frac{n-k^j}{2^j}\right) \tag{2}$$

$$D_j = \sum_{n=0}^{\infty} f(k)\frac{1}{\sqrt{2^j}}\psi\left(\frac{n-k^j}{2^j}\right). \tag{3}$$

The functions $\Phi_{a,\tau}(t), \psi_{a,\tau}(t)$ are called as scaling and wavelet functions to compute the approximation and detail coefficients on the given speech signal.

4.4 Autocorrelation

Autocorrelation function is used to find the repeating of patterns in the given signal [26]. Since voice regions are periodic in nature, autocorrelation is applied on the obtained speech frames to identify the repeating nature, unvoiced frames of speech represent the noise components and its correlation values are random compared to voiced speech. The autocorrelation function is given by Eq. (4):

$$Rxx(n) = \sum_{m=0}^{N-K-1} S[m]S[m+n], \tag{4}$$

where $s(m)$ denotes the speech signal and n is shifting factor.

4.5 Pitch calculation

Pitch is inverse of the time difference between two successive peaks of a voiced speech [27]. Calculation of inverse of time difference between two successive peaks of auto-correlated speech provides the pitch value. Voiced regions of speech have pitch values in the range of 100 Hz to maximum of 1 Khz. Unvoiced regions of speech frames have higher pitch values due to random nature. Pitch frequency calculation is given by Eq. (5):

$$F_0 = \frac{1}{T_0},\ T_0 = \left(t\left(R_{xx}(m_0)\right) - t\left(R_{xx}(m_1)\right)\right), \tag{5}$$

where $t(R_{xx}(m_0))$ is time duration at which first maximum peak occurs and $t(R_{xx}(m_1))$ is the time duration at which second maximum peak occurs in autocorrelation function. The difference between the occurrences of two consecutive peaks provides pitch period which is T_0. Inverse of the pitch period provides pitch frequency F_0.

4.6 Thresholding

Thresholding is the basic technique for segmentation of voiced and unvoiced regions of speech based on zero crossing-energy [28]. In this work, the 20th percentile and the 60th percentile of obtained pitch values are considered as Th_1 and Th_2 for male speakers, 35th and 75th percentile of obtained pitch values as Th_1 and Th_2 for female speakers respectively. The following conditions in Eq. (6) and Eq. (7) are considered for extraction of voiced regions of speech:

$$\text{Frame} = \begin{cases} \text{Voiced if } F_0(i) > Th_1 \text{ and } F_0(i) < Th_2 \\ \text{unvoiced else} \end{cases}, \tag{6}$$

where $F_0(i)$ indicates the pitch of present frame. If the present frame is unvoiced frame, again one more condition is considered based on the previous frame pitch value.

$$\text{Frame} = \begin{cases} \text{Voiced if } F_0(i-1) > Th_1 \text{ and } F_0(i-1) < Th_2 \\ \text{unvoiced else} \end{cases}, \tag{7}$$

where $F_0(i-1)$ indicates the pitch of previous frame. The above conditions provide the consideration of voiced frames based on pitch values of present frame and previous frame. If the present frame pitch values are within the threshold limits, it considered as voiced frame. If present frame is unvoiced, and previous frame pitch values are within threshold limits, the unvoiced frame samples are added to preserve the speech samples in transition region from voiced to unvoiced regions. This helps in preserving the information content at the edges of voiced frames of speech. This thresholding technique requires prior knowledge about the gender of the speaker. The male pitch values are low compared to the female speakers, so this particular work consider higher range of thresholds for female speakers.

4.7 Logical shaping and smoothing

All the voiced speech frames that are extracted from the speech signal are the overlapped versions with 50 % of previous frame samples. To avoid the repetition of data in two consecutive voiced frames, only last half of present frame samples are considered to provide continuity of speech without any redundancy. The extracted speech

frames are arranged in the order based on the index of frames logically, to obtain the speech signal which is fully of voiced regions. If two consecutive frames have pitch values out of range, the second frame is considered as unvoiced speech frame.

5 Results and discussion

The aforementioned algorithm was used in the extraction of voiced regions of speech from the emotional speech signals. Since human- human and machine interactions are done with some kind of emotions, RAVDESS database is considered in this work. Since, vocal cavity and nasal cavity changes effect speech production at different emotions and most of the information content lies in the voiced regions, its extraction is important. The performance of the proposed method is analysed by the standard metrics as Signal to Noise Ratio (SNR), Normalized Root Mean Square Error (NRMSE), spectral entropy and statistical parameters such as mean, standard deviation, mean gradient.

5.1 Evaluation metrics

5.1.1 Signal to Noise Ratio (SNR)

Signal to Noise Ratio is defined as the logarithmic ratio of signal power to that of noise power [29]. In this work, voiced regions are periodic in nature and considered as signal and unvoiced regions are random in nature and are considered as noise. The Signal to Noise Ratio is given by Eq. (8):

$$\mathrm{SNR}\,(dB) = 10\log_{10}\frac{\sum_{k=1}^{N} d^2(k)}{\sum_{k=1}^{N}\left[f(k)-d(k)\right]^2}, \tag{8}$$

where $d(k)$ is the desired emotional speech signal with voiced regions of speech and $f(k)$ is the actual emotional speech signal comprising of voiced and unvoiced regions. As the Signal to Noise Ratio is increased, the speech signal contains mostly of voiced regions of speech.

5.1.2 Normalized Root Mean Square Error (NRMSE)

The Normalized Root Mean Square Error is used to find, the extent of deviation between the original signal and the desired signal [29]. The Normalized Root Mean Square Error is given by Eq. (9):

$$\mathrm{NRMSE} = \sqrt{\frac{\left(f(k)-d(k)\right)^2}{\left(f(k)-u_f\right)^2}}, \tag{9}$$

where u_f is defined as the mean value of the original emotional speech signal.

5.1.3 Spectral entropy

Spectral entropy is the measure of distribution of energy content in signal w.r.t frequency [30]. Since voiced regions of speech are usually periodic in nature with more information, its entropy is high with sharp peaks at particular frequency. The unvoiced regions of speech have less information content and do not contain sharp peaks in frequency domain. The spectral entropy is defined by Eq. (10):

$$E = -\sum_{i=0}^{N-1} h(i)\log 2\left(h(i)\right), \tag{10}$$

where $h(i)$ is considered as the probability density function of power spectral density of speech signal.

5.1.4 Statistical parameters

Statistical parameters like mean, standard deviation and mean gradient are considered as the performance metrics in statistics [31].

Absolut value of mean

Mean of the signal defines the average information content present in the signal. As the mean value is high, the information content is also high. Voiced regions of have higher mean value compared to the unvoiced regions of speech. Mean value of signal $f(k)$ is given by Eq. (11):

$$u_f = \frac{1}{N}\sum_{k=0}^{N-1} f(k), \tag{11}$$

where N indicates the number of samples in the signal.

Standard deviation

Standard deviation defines the spread of signal samples from its mean value. As the standard deviation is higher, the spread is higher with more randomness from mean value and lower standard deviation provides the signal samples to spread around the mean value. It is given by Eq. (12):

$$std = \sqrt{\sum_{k=0}^{N-1}\left(f(k)-u_f\right)^2}. \tag{12}$$

Mean gradient

Mean gradient is measure of slope of the signal samples, which provide variation between two successive samples. This value is low in voiced regions of speech signal compared to unvoiced regions. The formula of mean gradient of signal $f(k)$ is given by Eq. (13):

$$Mg = \frac{1}{N-1}\sum_{k=0}^{N-1}\sqrt{\left(f(k)-f(k-1)\right)^2}, \tag{13}$$

where N defines the number of samples in speech signal.

5.2 Experimental results

The performance metrics are compared with the existing methods in the process of extracting voiced regions of speech.

1. Implementing voiced speech extraction using zero crossing-energy [14].
2. Implementing voiced speech extraction using residual signal [3].

The proposed method provides better results on 893 speech files out of selected 1000 samples. The figures and evaluation metric values of neutral and happy emotions are considered in ease of representation. Neutral emotion is considered in many voiced speech detection algorithms. Happy emotional speech is considered, to compare the proposed method on emotional speech. MATLAB 2014a version is used on laptop with 8 GB of RAM and 1.8 GHz frequency and Intel Core i3-3110M processor for implementing the proposed wavelet-pitch method.

Figs. 1–3 are related to the male speaker 1 and the sentence "*Kids are talking by the door*" is uttered in neutral emotion is indicated in blue color in all the three figures. The extracted voiced speech signal using three methods zero crossing-energy, residual signal and wavelet-pitch are indicated in red color. In Fig. 1, the extracted speech using zero crossing-energy method contains center portions of pure unvoiced speech which are also considered as voiced speech. In Fig. 2, extracted speech using residual signal does not contain the complete of voiced speech. In Fig. 3, the extracted speech using wavelet-pitch method contains most of voiced frames and its transition regions preserving the emotion content. Figs. 4–6 are related to female speaker 1 and the sentence "*Kids are talking by the door*" is uttered in happy emotion is indicated in blue color in all the three figures. The exracted voiced speech signal using three methods zero crossing-energy, residual signal and wavelet-pitch are indicated in red color. In Fig. 4, the extracted speech using zero crossing-energy does not preserve transition regions. In Fig. 5, the extracted speech using residual signal method does not extract full voiced speech regions. In Fig. 6, the extracted speech using wavelet-pitch contains most the voiced regions and preserving transition regions compared to the zero crossing-energy and residual signal methods. The subjective analysis of the proposed method can be done by hearing the new speech signal comprising of voiced regions of speech and emotional content in it. The proposed method preserves most of the emotional content in speech signal by extracting the transition regions.

Table 1 shows the comparison of various methods for voiced speech extraction w.r.t evaluation metrics like SNR and NRMSE.

Table 1 shows that the proposed method provides better results in terms of evaluation metrics for voiced speech extraction from emotional speech signals, compared to existing methods.

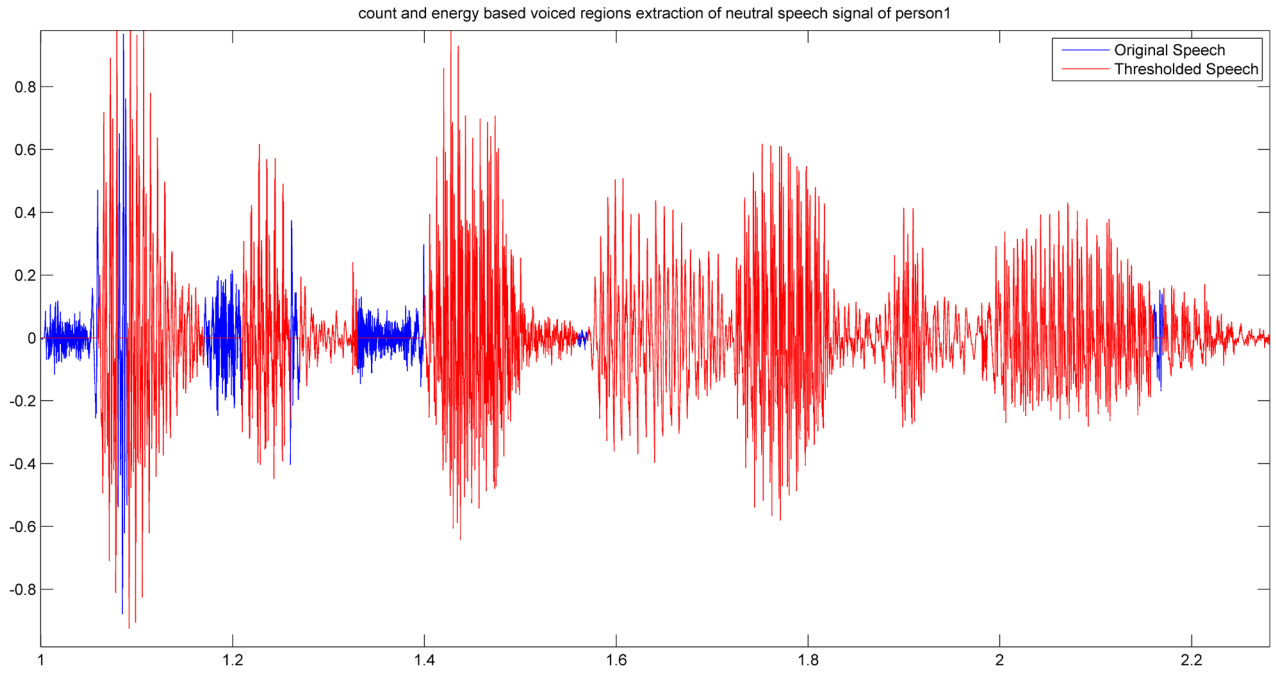

Fig. 1 Male speaker 1 uttered original speech signal with neutral emotion is indicated in blue color. Extracted voiced speech signal by applying zero crossing-energy method for male speaker 1 with neutral emotion, is indicated in red color.

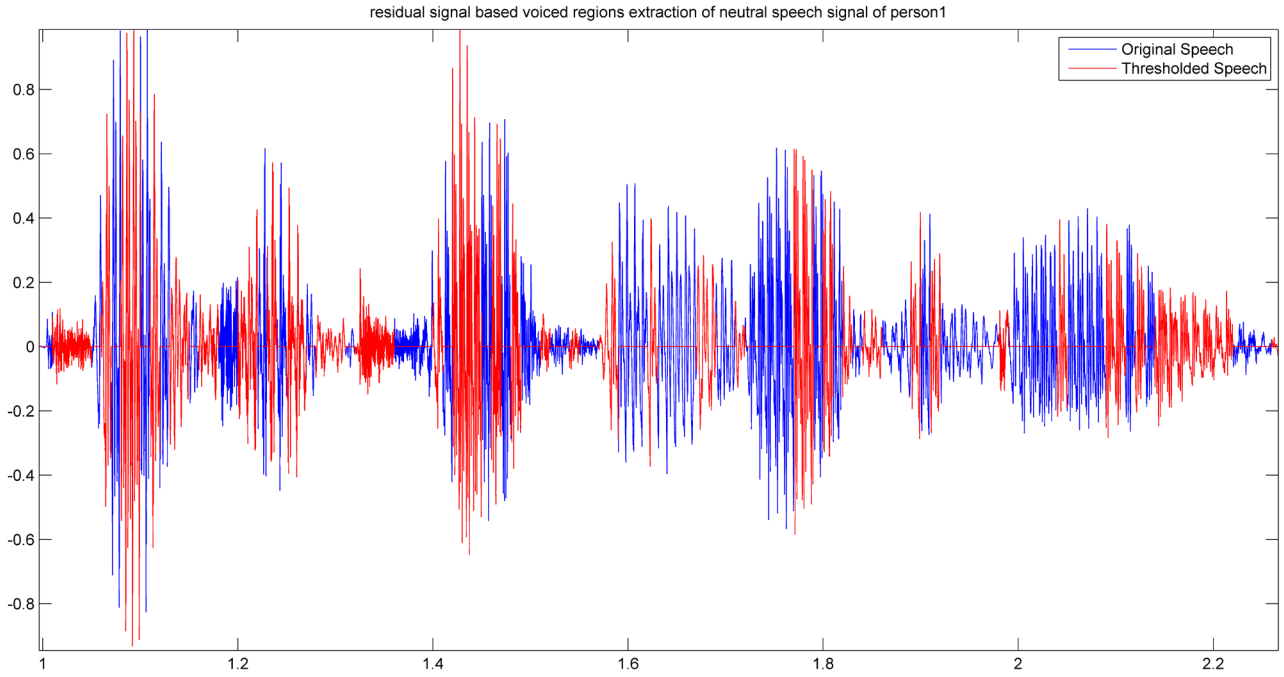

Fig. 2 Male speaker 1 uttered original speech signal with neutral emotion is indicated in blue color. Extracted voiced speech signal by applying residual signal method for male speaker 1 with neutral emotion, is indicated in red color.

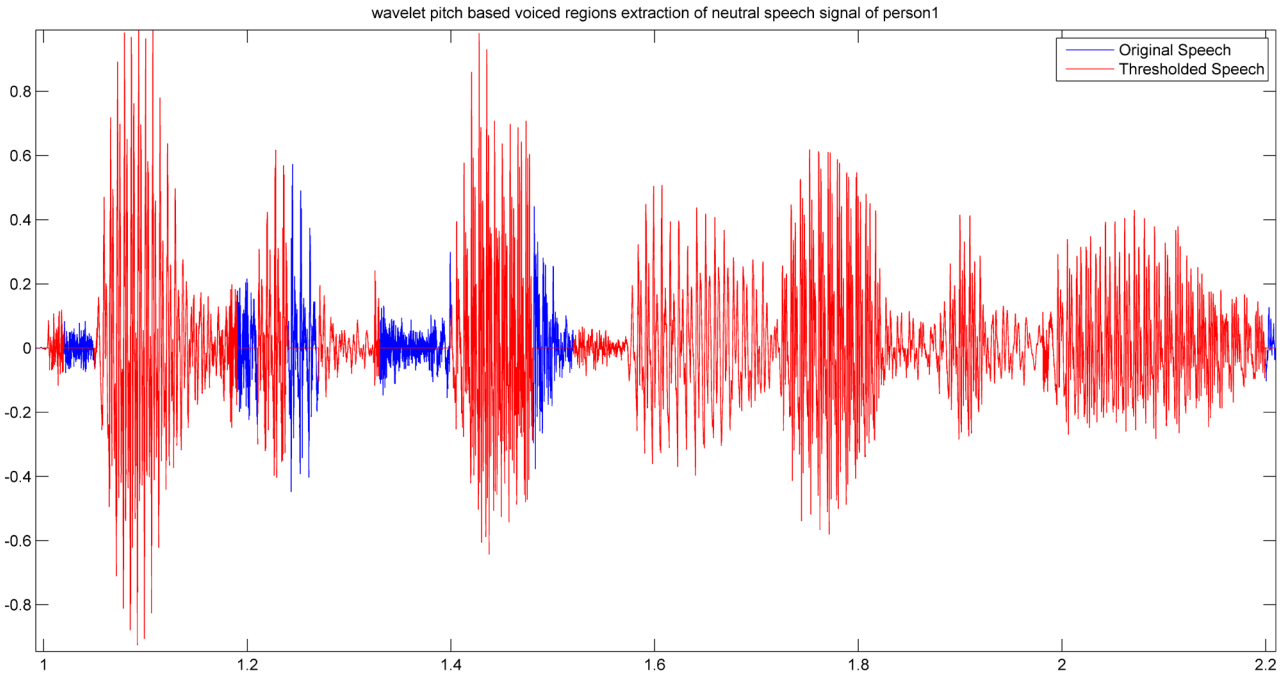

Fig. 3 Male speaker 1 uttered original speech signal with neutral emotion is indicated in blue color. Extracted voiced speech signal by applying proposed wavelet-pitch method for male speaker 1 with neutral emotion, is indicated in red color.

Table 2 shows the performance metrics for voiced speech extraction in terms of entropy and statistical parameters. From Table 2 it is known that extracted voiced speech signal contains less number of unvoiced frames compared to original speech signal.

Figs. 7, 8 show the graphical representation of SNR and NRMSE values of Table 1. From Figs. 7, 8, it is inferred that proposed method has higher SNR and lower NRMSE values compared to zero crossing-energy and residual signal methods in extraction of voiced speech from emotional speech signals.

Figs. 9–12 show the graphical representation of entropy, mean, standard deviation and mean gradient values of Table 2, for comparing w.r.t actual signal metric values.

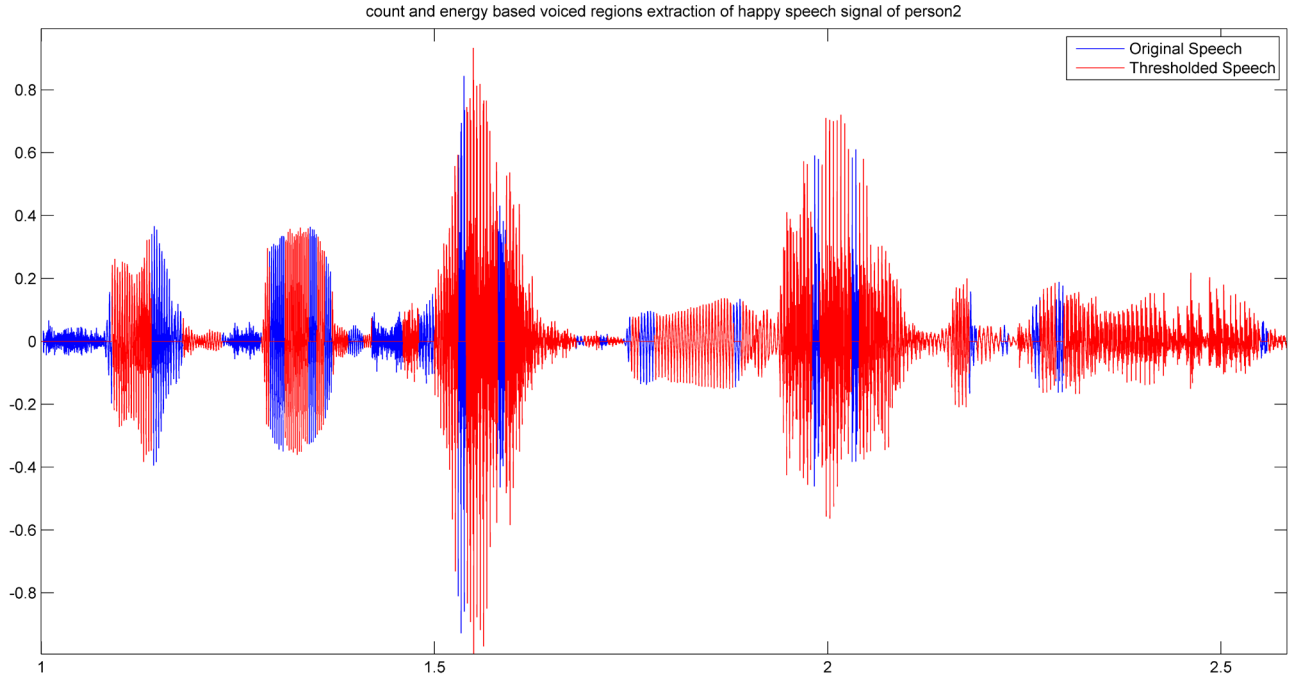

Fig. 4 Female speaker 1 uttered original speech signal with happy emotion is indicated in blue color. Extracted voiced speech signal by applying zero crossing-energy method for female speaker 1 with happy emotion, is indicated in red color.

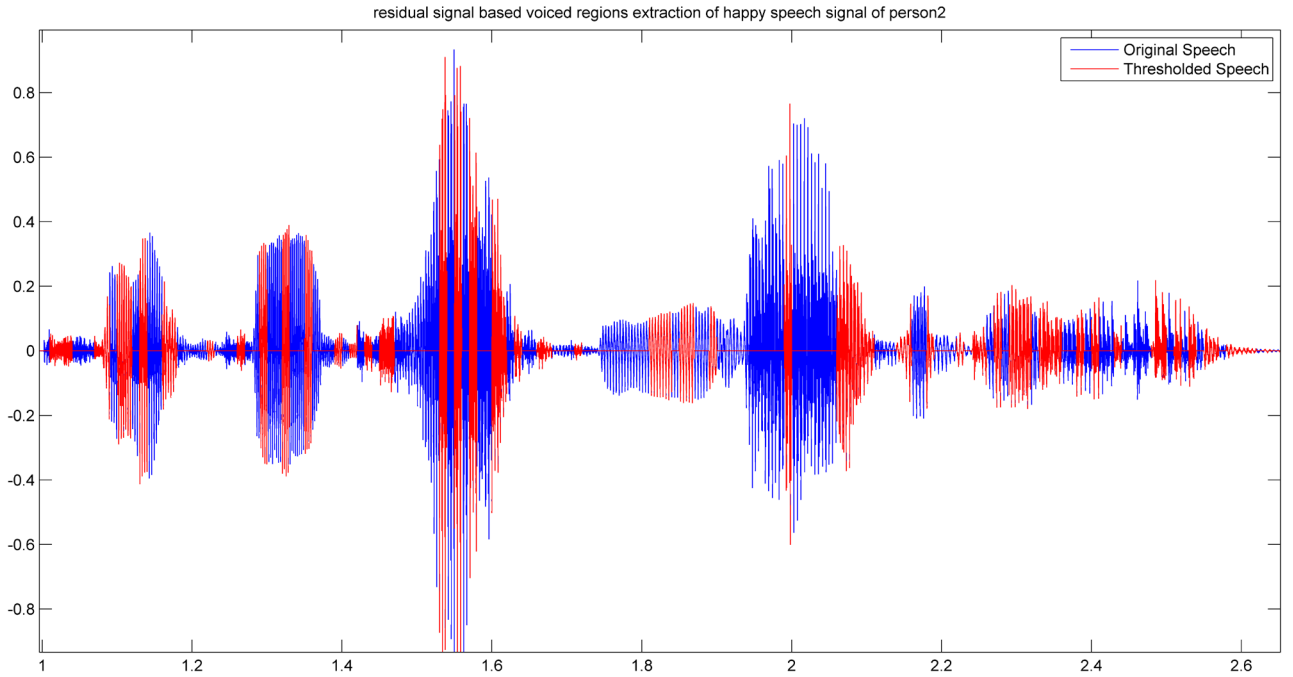

Fig. 5 Female speaker 1 uttered original speech signal with happy emotion is indicated in blue color. Extracted voiced speech signal by applying residual signal method for female speaker 1 with happy emotion, is indicated in red color.

From Figs. 9–12, it is understood that proposed method provides better results than residual signal method, but not than zero crossing-energy method. Since zero crossing-energy method extracts only voiced regions of emotional speech, without preserving transition regions and emotional content, it provides higher values. Residual signal method does not extract complete voiced regions of speech, it provides lower values. The proposed method extracts both voiced and transition regions preserving emotional content in speech signal, providing better values than residual signal method but not than zero crossing-energy method.

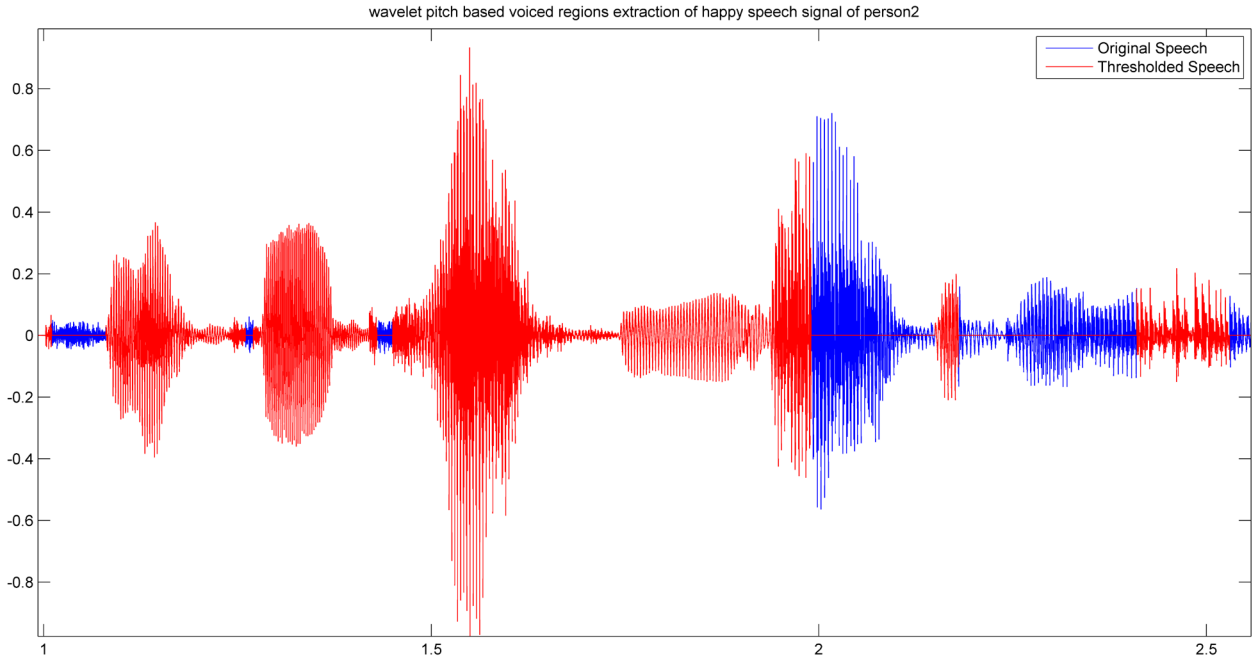

Fig. 6 Female speaker 1 uttered original speech signal with happy emotion is indicated in blue color. Extracted voiced speech signal by applying proposed wavelet-pitch method for female speaker 1 with happy emotion, is indicated in red color.

Table 1 Comparison of various methods in voiced speech extraction

Emotion and Person	Methods	Signal to Noise Ratio	Normalized Root Mean Square Error
Neutral Emotion - male speaker 1	zero crossing-Energy	24.3339	0.3349
	Residual Signal	−0.4699	0.6623
	Wavelet-Pitch	29.2501	0.1333
Happy Emotion - male speaker 1	zero crossing-Energy	15.9587	0.3524
	Residual Signal	2.1793	0.6729
	Wavelet-Pitch	40.5876	0.0489
Neutral Emotion - female speaker 1	zero crossing-Energy	22.4545	0.1865
	Residual Signal	−3.3625	0.8154
	Wavelet-Pitch	40.5091	0.0322
Happy Emotion - female speaker 1	zero crossing-Energy	12.1317	0.4503
	Residual Signal	−3.8642	0.7887
	Wavelet-Pitch	13.9438	0.3644

Table 3 shows the comparison of various wavelets for voiced speech extraction using wavelet-pitch method w.r.t SNR and NRMSE. Table 3 shows that Db4 wavelet provides better results in wavelet-pitch method compared to Haar and Db2 wavelets.

Figs. 13, 14 show the graphical representation of SNR and NRMSE values of Table 3. From Figs. 13, 14, it is inferred that Db4 wavelet provides high SNR and low NRMSE values compared to Haar and Db2 wavelets by extracting voiced and transition regions of emotional speech preserving emotional content.

6 Conclusion

The vocal tract, which acts as a time varying filter, various its shape in production of different emotional speech signals. The vocal tract characteristics are varied form one emotion to another and from one person to another. The vocal tract characteristics in different emotion can be analysed using pitch parameter, which varies from emotion to emotion. Normal voiced speech extraction algorithms provides better results in normal speech, but looses the emotional content in extraction of voiced regions from emotional speech signals. In this paper, voiced regions of

Table 2 Comparison of various methods with statistical parameters in voiced speech extraction

Emotion and Person	Methods	Entropy	Mean	Standard deviation	Mean gradient
Neutral Emotion - male speaker 1	Actual Signal	1.5671	0.188	0.0906	0.0066
	zero crossing-Energy	3.7694	0.564	0.1642	0.0180
	Residual Signal	1.9793	−6.032	0.1058	0.0096
	Wavelet-Pitch	2.5337	2.292	0.1312	0.0116
Happy Emotion - male speaker 1	Actual Signal	1.8631	0.104	0.1158	0.0090
	zero crossing-Energy	3.8327	−5.536	0.1935	0.0205
	Residual Signal	2.1329	11.087	0.1288	0.0109
	Wavelet-Pitch	2.9366	0.820	0.1701	0.0159
Neutral Emotion - female speaker 1	Actual Signal	1.8086	−0.028	0.0831	0.0082
	zero crossing-Energy	3.6469	−6.657	0.1408	0.0189
	Residual Signal	2.2745	0.710	0.0940	0.0118
	Wavelet-Pitch	2.8221	0.080	0.1241	0.0155
Happy Emotion - female speaker 1	Actual Signal	1.7113	0.013	0.0694	0.0076
	zero crossing-Energy	3.4250	0.674	0.1167	0.0184
	Residual Signal	1.6207	−0.636	0.0679	0.0071
	Wavelet-Pitch	2.4125	0.978	0.0993	0.0135

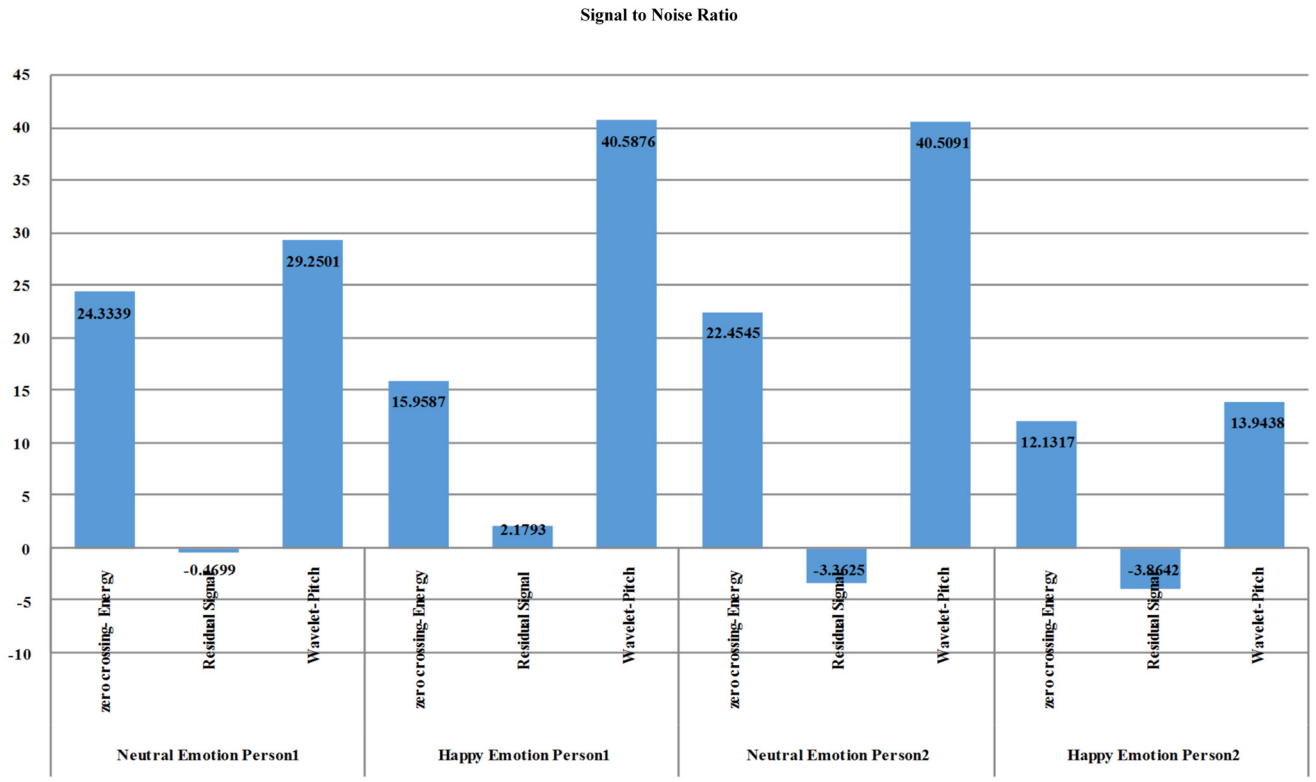

Fig. 7 Comparison of zero crossing- energy, residual signal and wavelet-pitch methods w.r.t SNR.

emotional speech signals are extracted by using wavelet-pitch method, which extracts voiced regions and transition regions between voiced and unvoiced regions, preserving the emotional content. The speech is divided into frames with 50 % overlap. Hamming window is applied to make, energy concentrated at the center of the frame. Discrete wavelet transform is applied using Daubechies Db4 wavelet on the frames and autocorrelation function is applied on approximation coefficients to extract the pitch values. The pitch values are compared based on thresholding technique. Percentile of pitch is considered as thresholding parameter because it is more resistive to the outliers

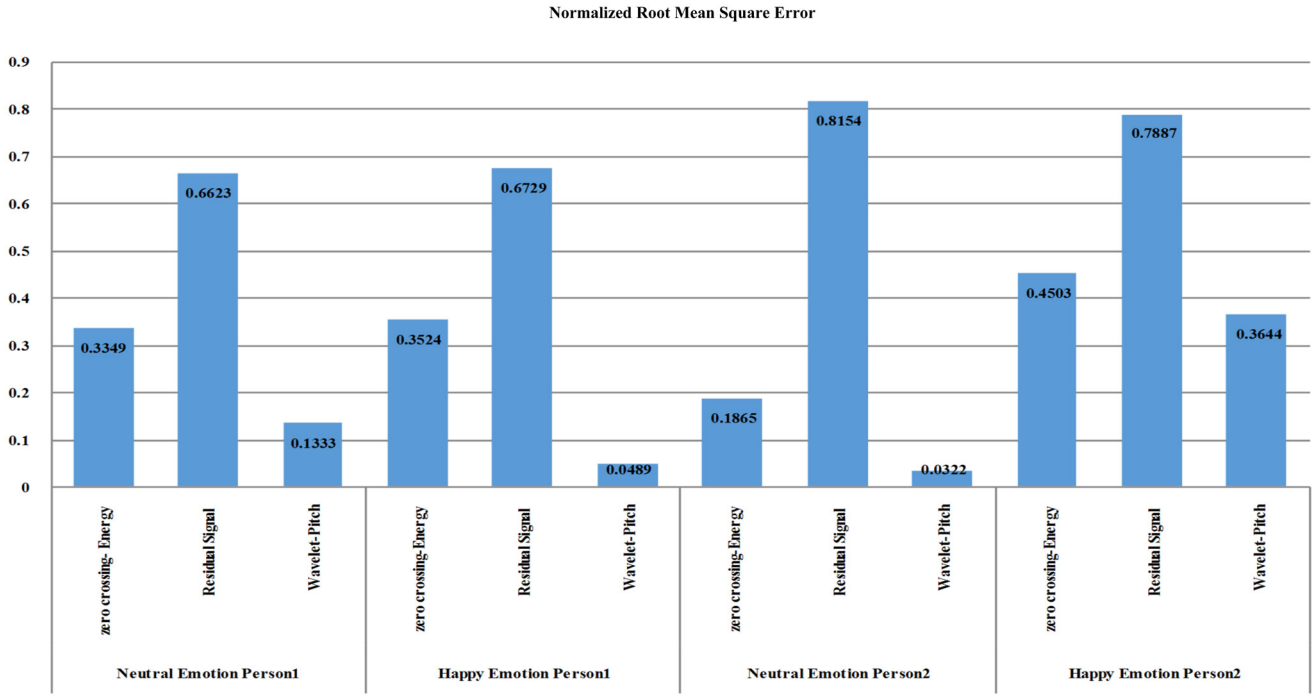

Fig. 8 Comparison of zero crossing-energy, residual signal and wavelet-pitch methods w.r.t NRMSE.

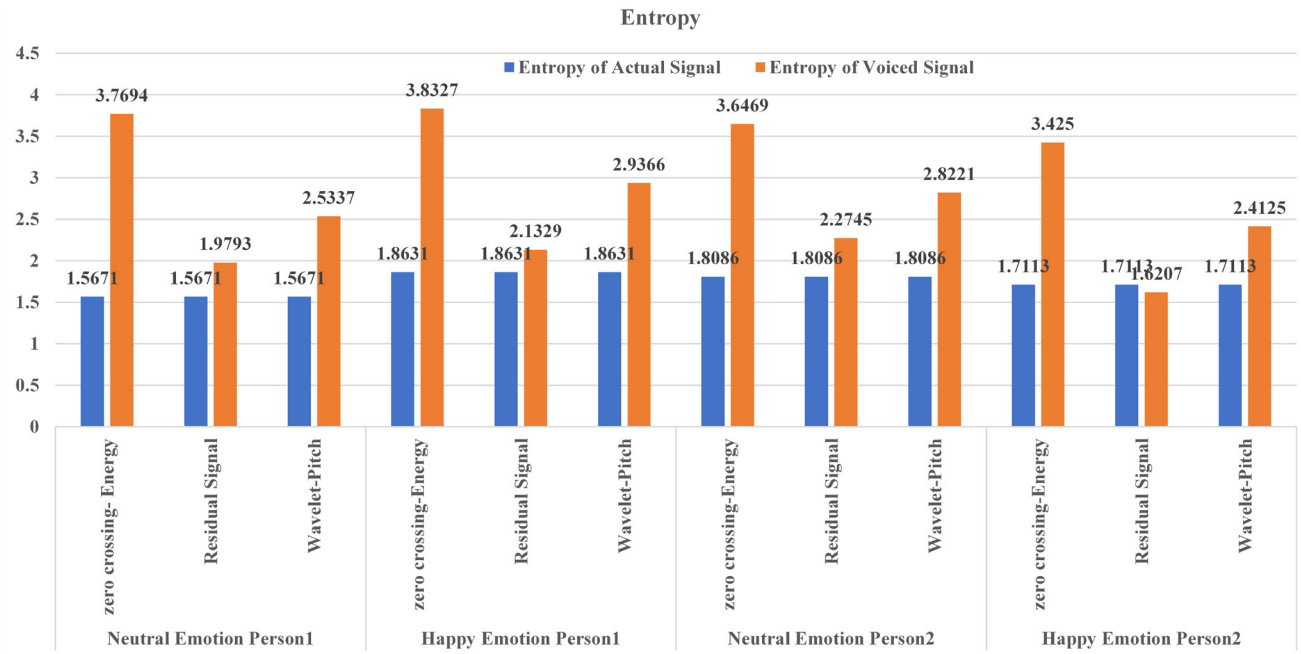

Fig. 9 Comparison of zero crossing-energy, residual signal and wavelet-pitch methods based extracted voiced speech entropy w.r.t entropy of original signal.

of frame pitch values. The threshold values are varied to male and female speaker as the male pitch values are lower the female pitch values. Based on thresholding, voiced regions of speech are extracted from emotional speech signal. The proposed method is compared with the zero crossing-energy based extraction algorithm and residual signal based extraction algorithm. The evaluation metrics like signal to noise ratio, Normalized Root Mean Square Error, entropy, mean, standard deviation and mean gradient metrics are calculated and results shows that, the proposed algorithm provides better signal to noise ratio and low Normalized Root Mean Square Error values, better entropy, mean, standard deviation and mean gradient values compared to existing methods. Daubechies Db4 wavelet provides better signal to noise ratio and Normalized Root Mean Square Error values compared to Haar and

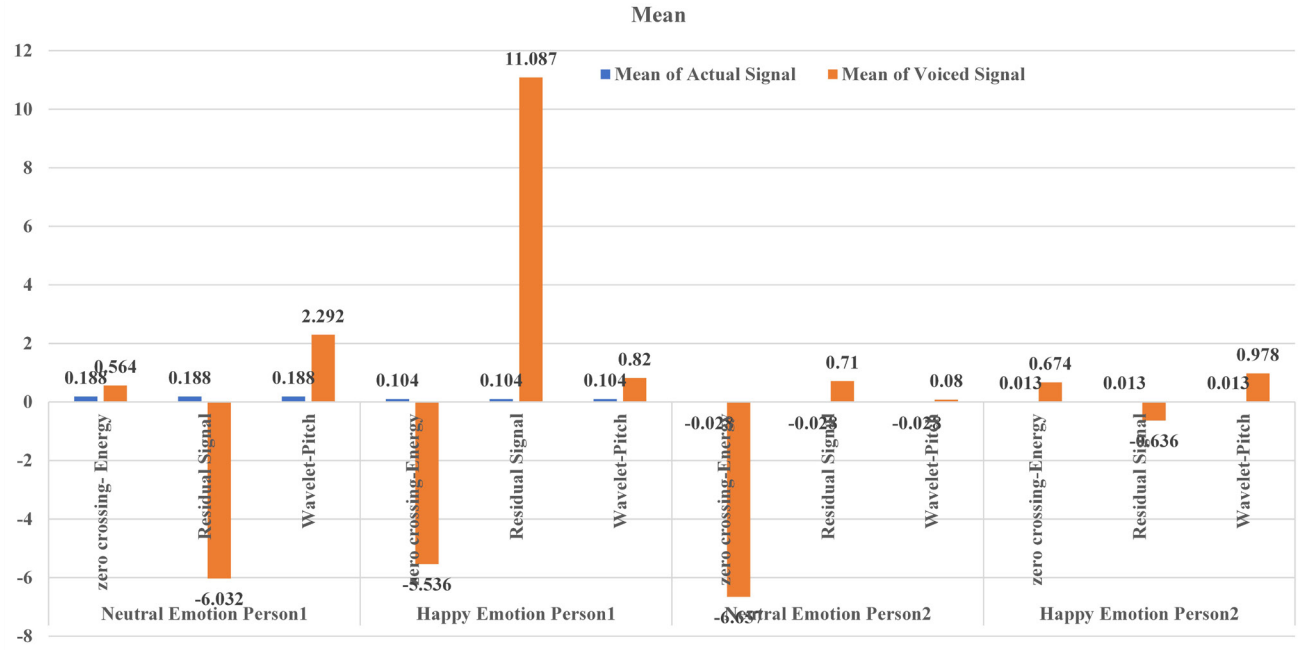

Fig. 10 Comparison of zero crossing-energy, residual signal and wavelet-pitch methods based extracted voiced speech mean w.r.t mean of original signal.

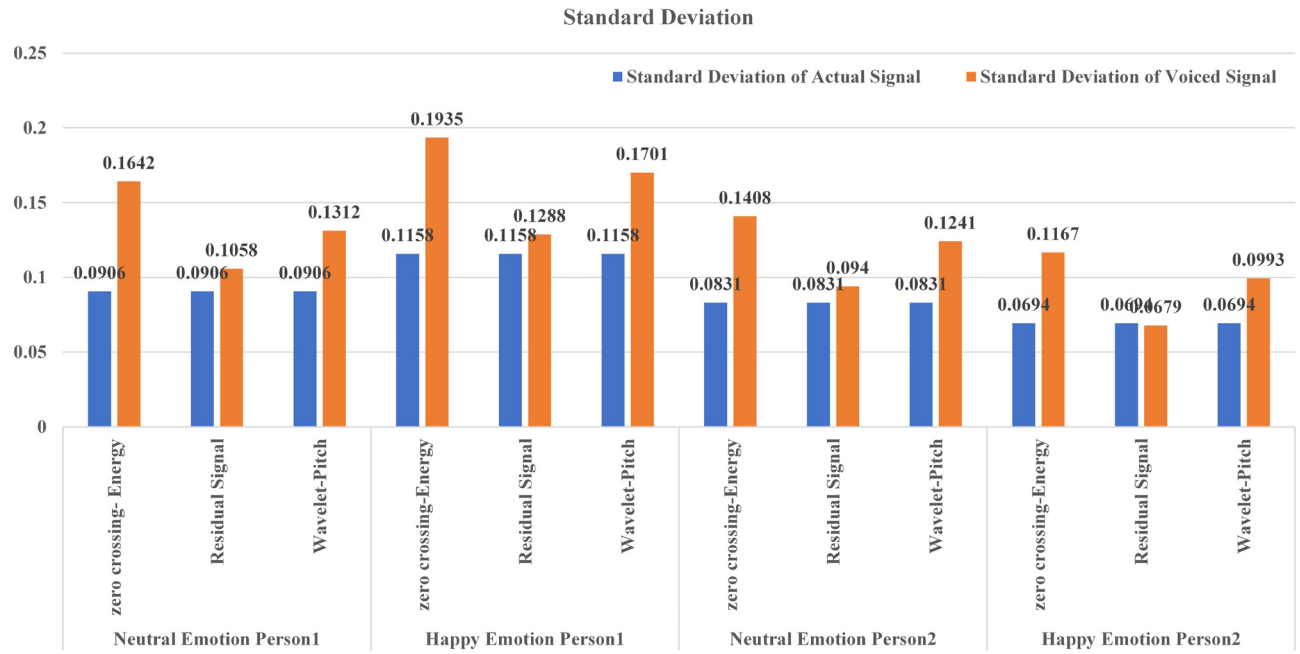

Fig. 11 Comparison of zero crossing-energy, residual signal and wavelet-pitch methods based extracted voiced speech standard deviation w.r.t standard deviation of original signal.

Db2 wavelets. This particular work can be used in pre-processing stage for better accuracy in speech recognition and mainly in speech emotion recognition applications.

Acknowledgement

The Authors would like to thank Vignan's Foundation for Science, Technology and Research for providing required facilities to carry out this research.

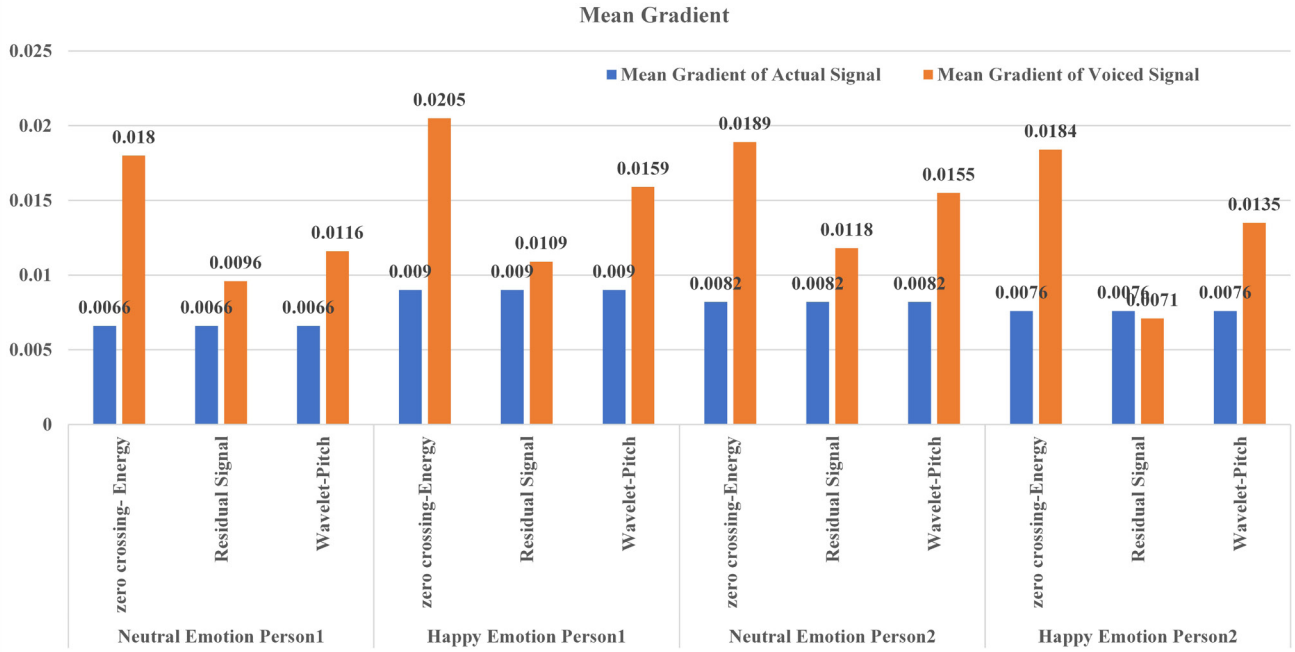

Fig. 12 Comparison of zero crossing-energy, residual signal and wavelet-pitch methods based extracted voiced speech mean gradient w.r.t mean gradient of original signal.

Table 3 Comparison of various wavelets with SNR and NRMSE

Emotion and Person	Wavelet	Signal to Noise Ratio	Normalized Root Mean Square Error
Neutral Person1	Haar	27.8411	0.1443
	Db2	27.3083	0.1454
	Db4	29.2501	0.1333
Happy Person 1	Haar	39.3887	0.0723
	Db2	37.5368	0.0741
	Db4	40.5876	0.0489
Neutral Person 2	Haar	39.8241	0.0334
	Db2	39.6153	0.0335
	Db4	40.5091	0.0322
Happy Person 2	Haar	17.9181	0.3179
	Db2	13.9579	0.3644
	Db4	13.9438	0.3644

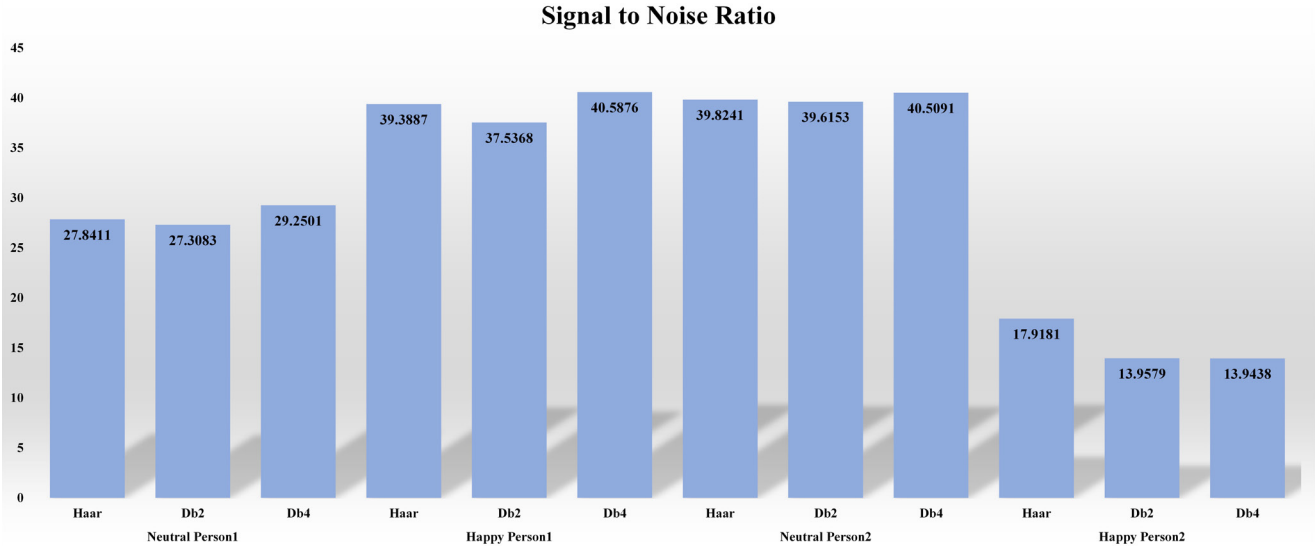

Fig. 13 Comparison of Haar, Db2 and Db4 wavelets in wavelet-pitch based voiced speech extraction for male speaker 1 and female speaker 1 in neutral and happy emotions w.r.t signal to noise ratio parameter.

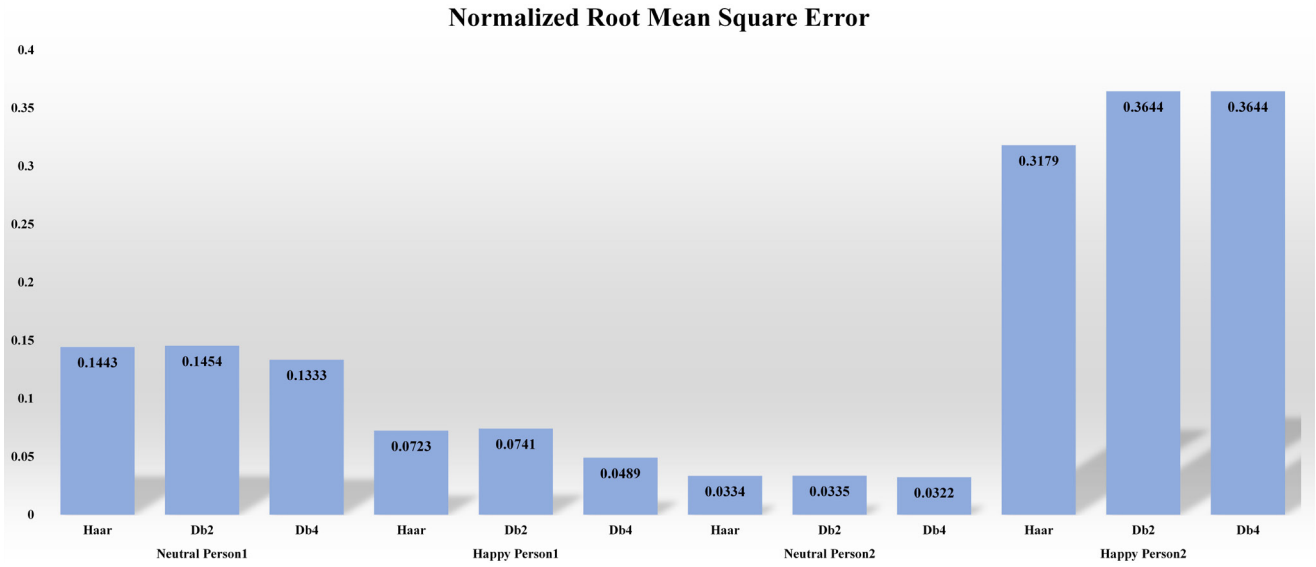

Fig. 14 Comparison of Haar, Db2 and Db4 wavelets in wavelet-pitch based voiced speech extraction for male speaker 1 and female speaker 1 in neutral and happy emotions w.r.t signal to Normalized Root Mean Square Error parameter.

References

[1] Lacroix, A. "Speech production-physics, models and prospective applications", In: International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces, Pula, Croatia, 2001, p. 3. https://doi.org/10.1109/ISPA.2001.938596

[2] Graf, S., Herbig, T., Buck, M., Schmidt, G. "Features for voice activity detection: a comparative analysis", EURASIP Journal on Advances in Signal Processing, 2015, Article number: 91, 2015. https://doi.org/10.1186/s13634-015-0277-z

[3] Jo, C.-W., Kim, J.-H. "Segregation of voiced and unvoiced components from residual of speech signal", Journal of Central South University, 19(2), pp. 496–503, 2012. https://doi.org/10.1007/s11771-012-1031-4

[4] Shaojun, J., Haitao, G., Fuliang, Y. "A new algorithm for voice activity detection based on wavelet transform", In: 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, China, 2004, pp. 222–225. https://doi.org/10.1109/ISIMP.2004.1434040

[5] Upadhyay, A., Pachori, R. B. "Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition", Journal of the Franklin Institute, 352(7), pp. 2679–2707, 2015. https://doi.org/10.1016/j.jfranklin.2015.04.001

[6] Erdol, N., Castelluccia, C., Zilouchian, A. "Recovery of missing speech packets using the short-time energy and zero-crossing measurements", IEEE Transactions on Speech and Audio Processing, 1(3), pp. 295–303, 1993. https://doi.org/10.1109/89.232613

[7] Sun, H., Ma, B., Li, H. "Frame selection of interview channel for NIST speaker recognition evaluation", In: 2010 7th International Symposium on Chinese Spoken Language Processing, Tainan, Taiwan, 2010, pp. 305–308. https://doi.org/10.1109/ISCSLP.2010.5684886

[8] Haigh, J. A., Mason, J. S. "Robust voice activity detection using cepstral features", In: IEEE Region 10 International Conference on Computers, Communications and Automation, Beijing, China, 1993, pp. 321–324. https://doi.org/10.1109/TENCON.1993.327987

[9] Jalil, M., Butt, F. A., Malik, A. "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals", In: 2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering, Konya, Turkey, 2013, pp. 208–212. https://doi.org/10.1109/TAEECE.2013.6557272

[10] Marques, J. S., Almeida, L. B. "Sinusoidal Modeling of Voiced and Unvoiced Speech", In: 1989 First European Conference on Speech Communication and Technology, Paris, France, 1989, pp. 2203–2206. [online] Available at: https://www.isca-speech.org/archive/eurospeech_1989/e89_2203.html [Accessed: 12 November 2019]

[11] Wu, B. F., Wang, K. C. "Voice Activity Detection Based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator", International Journal of Computational Linguistics and Chinese Language Processing, 11(1), pp. 87–100, 2006. [online] Available at: https://www.aclweb.org/anthology/O06-2006.pdf [Accessed: 15 November 2019]

[12] Asgari, M., Sayadian, A., Farhadloo, M., AbouieMehrizi, E. "Voice Activity Detection Using Entropy in Spectrum Domain", In: 2008 Australasian Telecommunication Networks and Applications Conference, Adelaide, Australia, 2008, pp. 407–410. https://doi.org/10.1109/ATNAC.2008.4783359

[13] Swee, T. T., Salleh, S. H. S., Jamaludin, M. R. "Speech pitch detection using short-time energy", In: International Conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia, 2010, pp. 1–6. https://doi.org/10.1109/ICCCE.2010.5556836

[14] Jafer, E., Mahdi, A. E. "Wavelet-based voiced/unvoiced classification algorithm", In: 2003 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications, Zagreb, Croatia, 2003, pp. 667–672. https://doi.org/10.1109/VIPMC.2003.1220540

[15] Rani, B. M. S., Rani, A. J., Ravi, T., Sree, M., D. "Basic Fundamental Recognition of Voiced Unvoiced and Silence Region of a Speech", International Journal of Engineering and Advanced Technology, 4(2), pp. 83–86, 2014. [online] Available at: https://tarjomefa.com/wp-content/uploads/2016/09/5351-English.pdf [Accessed: 09 November 2019]

[16] Tanmoy, R., Marwala, T., Chakraverty, S. "Precise detection of speech endpoints dynamically: A wavelet convolution based approach", Communications in Nonlinear Science and Numerical Simulation, 67, pp. 162–175, 2019. https://doi.org/10.1016/j.cnsns.2018.07.008

[17] Hamzenejadi, S., Goki, S. A. Y. H., Ghazvini, M. "Extraction of Speech Pitch and Formant Frequencies using Discrete Wavelet Transform", In: 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems, Bojnord, Iran, 2019, pp. 1–5. https://doi.org/10.1109/CFIS.2019.8692150

[18] Dendukuri, L. S., Hussain, S. J. "Enhanced Feature Set Calculation from Emotional Speech Signals", In: 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking, Vellore, India, 2019, pp. 1–5. https://doi.org/10.1109/ViTECoN.2019.8899416

[19] Busso, C., Bulut, M., Lee, S., Narayanan, S. "Fundamental Frequency Analysis for Speech Emotion Processing", In: Hancil, S. (ed.) The Role of Prosody in Affective Speech, Peter Lang Publishing Group, Berlin, Germany, 2009, pp. 309–337. [online] Available at: https://sail.usc.edu/publications/files/Busso_2009_2.pdf [Accessed: 19 April 2020]

[20] Gharavian, D., Sheikhan, M., Janipour, M. "Pitch in Emotional Speech and Eemotional Speech Recognition Using Pitch Frequency", Majlesi Journal of Electrical Engineering, 4(1), pp. 19–24, 2010. https://doi.org/10.1234/mjee.v4i1.159

[21] Breitenstein, C., Van Lancker, D., Daum, I. "The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample", Cognition and Emotion, 15(1), pp. 57–79, 2001. https://doi.org/10.1080/02699930126095

[22] Livingstone, S. R., Russo, F. A. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English", PloS One, 13(5), Article number: e0196391, 2018. https://doi.org/10.1371/journal.pone.0196391

[23] Daubechies, I. "The wavelet transform, time-frequency localization and signal analysis", IEEE Transactions on Information Theory, 36(5), pp. 961–1005, 1990. https://doi.org/10.1109/18.57199

[24] Eshaghi, M., Mollaei, M. R. K. "Voice activity detection based on using wavelet packet", Digital Signal Processing, 20(4), pp. 1102–1115, 2010. https://doi.org/10.1016/j.dsp.2009.11.008

[25] Wang, K.-C. "Robust Voice Activity Detection Based on Discrete Wavelet Transform", In: 20th Conference on Computational Linguistics and Speech Processing, Taipei, Taiwan, 2008, pp. 216–228. [online] Available at: https://www.aclweb.org/anthology/O08-1016.pdf [Accessed: 10 November 2019]

[26] Ghaemmaghami, H., Baker, B., Vogt, R., Sridharan, S. "Noise Robust Voice Activity Detection Using Ffeatures Extracted from the Time-Domain Autocorrelation Function", In: Hirose, K., Nakamura, S., Kaboyashi, T. (eds.) Proceedings of the 11th Annual Conference of the International Speech Communication Association, International Speech Communication Association, CD Rom, 2010, pp. 3118–3121. [online] Available at: https://eprints.qut.edu.au/40656/ [Accessed: 10 November 2019]

[27] Drugman, T., Alwan, A. "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics", In: 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), Florence, Italy, 2011, pp. 1973–1976. [online] Available at: http://www.seas.ucla.edu/spapl/paper/IS110135.pdf [Accessed: 09 November 2019]

[28] Bachur, R. G., Kopparthi, S., Adapa, B., Barkana, B. D. "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal", In: Proceedings of the American Society for Engineering Education Zone Conference, United States Military Academy, West Point, NY, USA, 2008, pp. 1–7. [online] Available at: https://www.asee.org/documents/zones/zone1/2008/student/ASEE12008_0044_paper.pdf [Accessed: 08 November 2019]

[29] Baili, J., Lahouar, S., Hergli, M., Al-Qadi, I. L., Besbes, K. "GPR signal de-noising by discrete wavelet transform", Ndt & E International, 42(8), pp. 696–703, 2009. https://doi.org/10.1016/j.ndteint.2009.06.003

[30] Zhang, Y., Wei, S., Long, Y., Liu, C. "Performance Analysis of Multiscale Entropy for the Assessment of ECG Signal Quality", Journal of Electrical and Computer Engineering, 2015, Article ID: 563915, 2015. https://doi.org/10.1155/2015/563915

[31] Al-Hashemy, B. A. R., Taha, S. M. R. "Voiced-unvoiced-silence classification of speech signals based on statistical approaches", Applied Acoustics, 25(3), pp. 169–179, 1988. https://doi.org/10.1016/0003-682X(88)90092-8