Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2017) 25: 2820 – 2830

doi:10.3906/elk-1604-437

Extraction of geometric and prosodic features from human-gait-speech data for behavioral pattern detection: Part I

Raj Kumar PATRA[1,*], **Tilendra Shishir SINHA**[2,3], **Ravi Prakash DUBEY**[1]
[1]Department of Computer Science & Engineering, Dr. C. V. Raman University, Kota, Bilaspur, India
[2]Dr. Sudhir Chandra Sur Degree Engineering College, Surermath, Kolkata, India
[3]Maulana Abul Kalam Azad University of Technology, Salt Lake City, Kolkata, India

Received: 30.04.2016 • **Accepted/Published Online:** 01.11.2016 • **Final Version:** 30.07.2017

Abstract: In this work, we extract prosodic features from subjects while they are talking while walking using human-gait-speech data. These human-gait-speech data are separated into 1D data (human-speech) and 2D data (human-gait) using the adaptive-lifting scheme of wavelet transform. We carry out extraction of prosodic features from human-speech data, such as speech duration, pitch, speaking rate, and speech momentum, using five different natural languages (Hindi, Bengali, Oriya, Chhattisgarhi, and English) for the detection of behavioral patterns. These behavioral patterns form real-valued measured parameters, stored in a knowledge-based model called the human-gait-speech model. Extraction of geometrical features from human-gait data, such as step length, energy or effort, walking speed, and gait momentum, is carried out for the authentication of behavioral patterns. In this paper, the data of 25 subjects of different ages, talking in five different natural languages while walking, are analyzed for the detection of behavioral patterns.

Key words: Adaptive-lifting scheme of wavelet transform, out-of-corpus, blind speech signal separation, modified adaptive vector quantization

1. Introduction

Prosodic features are the real values measured from speech patterns [1] such as pitch, energy, or duration between utterances. In the present work, both human gait (2D data) and human speech (1D) are considered simultaneously for analysis in two different domains. Both these data have been mapped properly using statistical and soft-computing techniques [2–8]. In the central and eastern parts of India, the most commonly spoken languages are Hindi, Bengali, Oriya, and Chhattisgarhi, which contain some phrases with resemblances. The English language has also been adopted. All of these languages have been used as a case-based study for the successful extraction of prosodic features. The entire speech pattern of the five languages is partitioned [7] into considerable isolated subwords with optimal boundaries [9]. A wavelet, which is considered here as any real-valued function of time, has a certain structure and has been detected using the discrete wavelet transform (DWT) method [10]. Hence, lossless compression is performed so that no information is lost. Another focus of this paper is human-gait data for the extraction of geometrical features, including step length, walking speed, and energy or effort. Human-gait images have been studied in the literature, and a proposal has been made for the detection of a set of semantic traits discernible by humans at a distance, outlining their psychological validity [11]. Translation of human-gait biometrics has also been performed for forensic use [12], where features such

*Correspondence: patra.rajkumar@gmail.com

as ankle, knee, and hip locations were measured to match the walking sequences. Similar work with different methods and mechanisms has been carried out for the detection of abnormal footprints using human-gait images [13,14]. Further research has been carried out on tracking people considering video sequences, which is becoming more popular in computer vision research [15]. Still, very little research has been completed on investigating behavioral patterns using human-gait-speech [16–19] data. Consider the schematic diagram that yields a unique detection and authentication system, as depicted in Figure 1.

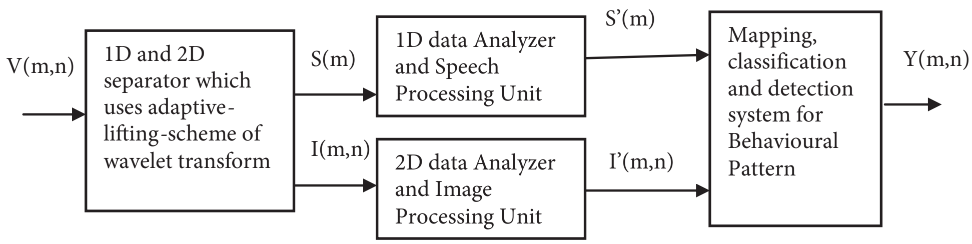

Figure 1. Schematic diagram of detection and authentication system.

From Figure 1, V(m,n) is the video data, which are separated into two types of data: human-speech (1D) and human-gait (2D). Let the human-speech data be S(m) and the human gait data be I(m,n). These are passed through an analyzer separately, resulting in S'(m) and I'(m,n) data. The resultant outputs are mapped and classified, and a decision is made.

1.1. Lifting scheme of wavelet transform and its inverse mechanism

The working principle of the 1D separator is explained through a schematic diagram, as shown in Figure 2.

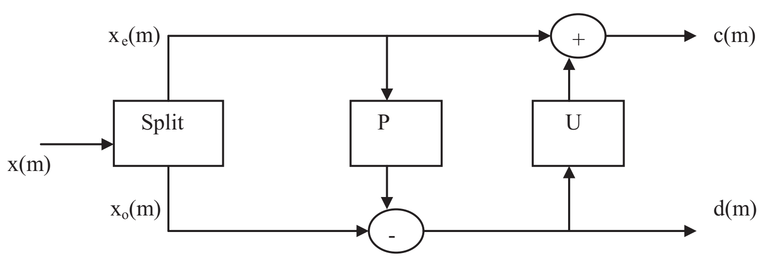

Figure 2. The 1D separator using the adaptive-lifting scheme of wavelet transform.

With reference to Figure 2, the original 1D data, for example x(m), are separated into two disjoint subsets of sample, such as x_e(m) and x_o(m). From these, we generate the detail signal d(m) and another operator called the prediction operator, P. Additionally, the coarser signal, c(m), is generated along with an updation operator, U, which is multiplied by the detail signal and summed up with the even components of the original 1D data [2–5]. The detail signal d(n) is generated using the prediction operator P, as depicted in Figure 2. Thus, mathematically it yields:

$$d(m) = x_o(m) - P(x_e(m)). \tag{1}$$

Similarly, the coarser signal c(m) is generated using the updation operator U. Hence, applied to d(m) and adding the result to x_e(m), it yields:

$$c(m) = x_e(m) + U(d(m)). \tag{2}$$

After substituting Eq. (1) in Eq. (2), it yields:

$$c(m) = x_e(m) + Ux_o(m) - UP(x_e(m)). \tag{3}$$

The lifting scheme parameters, P and U, are computed initially using the simple iterative method of numerical computation. Algorithm 1, depicted below for such computations, has been utilized in the present work.

Algorithm 1 Computation of lifting-scheme parameters U and P.

1: Generate two random numbers for U and P.
2: Calculate U' and P' (by rescaling U and P from –N to +N) and compute S
   (scaling factor) using the following formulae:

$$U' = ((2 \times U) - 1)$$

$$P' = ((2 \times P) - 1)$$

$$S = (U')^2 + (P')^2$$

3: If the value calculated is not within the N circle,
   then repeat Step 2 or do Step 4.
4: Compute standardized normally distributed random variables:

$$U = U' \times (-2log(S)/S)$$

$$P = P' \times (-2log(S)/S)$$

   (scaling factor) using the formulae:

$$U' = ((2 \times U) - 1)$$

$$P' = ((2 \times P) - 1)$$

$$S = (U')^2 + (P')^2$$

Further analysis is performed for checking whether the predicted and updated coefficients are computed properly or not. This has been done using the adaptive inverse lifting scheme of wavelet transform, as shown in Figure 3.

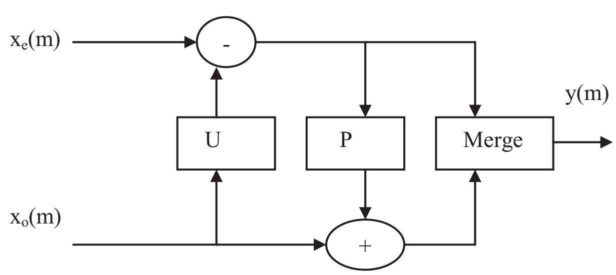

Figure 3. The 1D merger using adaptive inverse lifting scheme of wavelet transform.

With reference to Figure 3, 1D data are merged after computing the predicted and updated coefficients, and the resultant output y(m) is obtained, which is mapped. The paper is organized as follows: Section 2 illustrates the mathematical modeling for human-gait-speech data, Section 3 provides a brief scientific discussion of the simulated results, and Section 4 gives the concluding remarks and further scope of the work.

2. Mathematical modeling for human-gait-speech data

Let V_D be the video data, which incorporate both human-speech signal (H_S) and human-gait image (H_G) data. Hence, it yields:

$$V_D = H_S + H_G. \tag{4}$$

2.1. Modeling for human speech data segment

Solving for H_S, the human-speech signal is separated into two components: odd and even. This transformation reduces the dimensions of the feature vector from 'n' to d = M − 1, where M is the number of components involved. Suppose for 'N' known samples X_i, N_1 of which belong to component w_1 and N_2 of which belong to component w_2. Consider Y_i to be a linear combination of the features X_i:

$$Y_i = \rho^T X_i. \tag{5}$$

The 'n'-dimensional vector ρ is considered as a line in 'n'-dimensional space. Therefore, Y_i is the projection of X_i on this line (scaled by $\|\rho\|$). Consider μ_i to be the mean of the N_i samples of component w_i in the 'n'-dimensional space. It yields:

$$\mu_i = \frac{1}{N_i}\sum X_i. \tag{6}$$

The mean of the projected points Y_i on the line ρ, $\overline{\mu_i}$, is the projection of μ_i:

$$\overline{\mu_i} = \frac{1}{N_i}\sum Y_i = \frac{1}{N_i}\sum \rho^T X_i = \rho^T \mu_i. \tag{7}$$

The separation of the means on ρ is given by:

$$\left| \overline{\mu_1} - \overline{\mu_2} \right| = \left|\rho^T(\mu_1 - \mu_2)\right|. \tag{8}$$

The separation of two components includes the variance of samples. Therefore, defining the n × n scatter matrix,

$$w_i = \sum_{\beta \in w_i} (X - \mu_i)(X - \overline{\mu_i})^T, \tag{9}$$

where w_i is the covariance of the ith class in n-dimensional space. Thus, the total matrix is:

$$W = w_1 + w_2. \tag{10}$$

Let matrix B represent the dispersion between the means of various components:

$$B = (\mu_1 - \mu_2)\left(\overline{\mu_1} - \overline{\mu_2}\right)^T. \tag{11}$$

Then the power spectrum density is calculated and yields:

$$P_E(f) = \sum_{m=0}^{N-1} C_{xx}(m) W \exp(-j2\pi f m), \tag{12}$$

where $C_{xx}(p)$ is an autocovariance function, W is the total matrix, and *fm* is the modulated frequency. Next, lossless compression is applied:

$$S_c(k) = Re[S(k)] = \sum_{n=0}^{N-1} s(n)\cos\left(\frac{k2\pi n}{N}\right), k = 0, 1, 2, \ldots, N-1. \tag{13}$$

For different natural languages, let $L_{in} = \{L_1, L_2, \ldots, L_n\}$ and $C_{out} = \{C_1, C_2, \ldots, C_m\}$ consisting of 'n' and 'm' input language and output code words, respectively. The values of 'n' and 'm' are of maximum size. In the mapping stage, a sequence of feature vectors, $U = \{U_1, U_2, \ldots, U_u\}$, is compared to a sequence of feature vectors $K_{database} = \{K_1, K_2, \ldots, K_q\}$. Hence, to satisfy the unidirectional associative condition, that is, $K_{out} = K_{in}$, c-means clustering and a genetic algorithm are applied for the positive matching. The utterance with the lowest distortion is chosen. This yields:

$$C_{found} = arg\left\{S\left(U_u, K_q\right)\right\}. \tag{14}$$

Euclidean distance measurement gives the distortion measure:

$$S\left(U, K_i\right) = \frac{1}{Q}\sum_{i=1}^{Q} d\left(u_i, C_{min}^{i,q}\right), \tag{15}$$

where $C_{\min}^{i,q}$ denotes the closest uttered pattern and d(.) is the Euclidean distance. Thus, each human-speech feature vector in U is compared with the code word in the corpus of the system, and the minimum average distance is chosen as the best-fit code word. If the unknown vector is far from the other vectors, it is very difficult to find the desired pattern from the corpus, resulting in an out-of-corpus (OOC) problem. To overcome this problem, a similarity measure is carried out with maximized data values. This yields:

$$S_w\left(U, K_i\right) = \frac{1}{Q}\sum_{i=1}^{q}\frac{1}{d\left(u_i, C_{min}^{i,q}\right)} w\left(C_{min}^{i,q}\right). \tag{16}$$

Dividing Eq. (15) by Eq. (16) yields:

$$\gamma = \text{mapping rate} = \frac{S\left(U, K_i\right)}{S_w\left(U, K_i\right)} = \frac{\text{unweighted}}{\text{weighted}}. \tag{17}$$

The algorithm for computing the weight is depicted in Algorithm 2:

2.2. Modeling for human-gait data segment

Further solving for H_G, let a linear combiner be a function that takes considerable inputs and results in a single value. Let the input sequence be $\{X_1, X_2, \ldots, X_N\}$ and the relevant weight $\{W_1, W_2, W_3, \ldots, W_N\}$, so that the outcome of the linear combiner Y yields:

$$Y = \sum_{i=1}^{N} X_i W_i. \tag{18}$$

Algorithm 2 Computing the weight (S).

```
for each V_I in S do
    for each V_J in C_I do
        sum = 0
        for each V_K and K != I, in S do
          d_min = distancetonearest(V_J,V_K);
          sum = sum + 1 / d_min;
        endfor;
      w(V_IJ) = 1 / sum;
    endfor
  endfor
return weights = w(V_IJ)
```

An activation function, which in this case is treated as a sigmoid function, takes input values between negative and positive infinity. Thus, it becomes:

$$f(Y) = \frac{1}{1+e^{-Y}}. \tag{19}$$

The threshold, which is the set-point, interprets the internal activity of the neuron and is fixed to a negative unity value. For firing, the sum must be higher than its threshold value. The human-gait data have been also used for the computation of additional parameters, which are real-valued, neutral, normalized, and finally optimized.

Consider that 'Z' numbers of frames are being read. Let each frame be FRM_1, FRM_2, FRM_3, FRM_4, $\ldots \mathrm{FRM}_{Z-2}$, FRM_{Z-1}, FRM_Z, adhering to certain guidelines:

- Read FRM_{1L} when the subject is moving from left to right;
- Extract geometrical feature 'step-length', G_{1L};
- Similarly, read FRM_{1R} when the subject is moving from right to left;
- Similarly, extract geometrical feature 'step-length', G_{1R};
- Compute an average 'step-length', $\mathrm{G}_{1avg} = (\mathrm{G}_{1L} + \mathrm{G}_{1R}) / 2$.

Similar computations are carried out for the other frames, resulting in an average 'step-length', measured as real-valued data as follows:

$$\begin{aligned} G_{2avg} &= (G_{2L}+G_{2R})/2, G_{3avg} = (G_{3L}+G_{3R})/2, G_{4avg} \\ &= (G_{4L}+G_{4R})/2, G_{5avg} = (G_{5L}+G_{5R})/2, \ldots, G_{(Z-1)avg} \\ &= (G_{(Z-1)L}+G_{(Z-1)R})/2, G_{Zavg} = (G_{ZL}+G_{ZR})/2. \end{aligned} \tag{20}$$

For the computation of the neutral parameter, let 'N_{odd}' and 'N_{even}' be the number of odd and even frames, respectively. Thus, it further yields:

$$G_{Oddavg} = (G_{1avg} + G_{3avg} + \cdots + G_{(2Z-1)avg})/N_{odd}, \tag{21}$$

$$G_{Evenavg} = (G_{2avg} + G_{4avg} + \cdots + G_{(2Z)avg})/N_{even}. \tag{22}$$

Then the average neutral and normalized parameters for each frame are computed by taking the natural logarithm of Eqs. (18) and (19). As per the boundary condition, the matching of the start and end of the human-gait pattern and the tracing of the optimal route for normal gait have been analyzed. To formulize this, let the human-gait feature vectors β_i(k) and β_j(k) be denoted for human-gait patterns x(t$_i$) and x(t$_j$), respectively. The distance between the above two human-gait feature vectors is defined by:

$$d(c(k)) = d(i(k), j(k)) = \left| \ \beta_i(k) \quad - \quad \beta_j(k) \ \right|. \tag{23}$$

Hence, the dynamic time warping function is computed so that the resultant performance index D(x(t$_i$),x(t$_j$)) is minimized. The performance index is the normalized average weighted distance, which is related as:

$$D(x(t_i), x(t_j)) = \min_w \left[\frac{\sum\limits_{k=1}^{k} d(c(k))\rho(k)}{\sum\limits_{k=1}^{k} \rho(k)} \right], \tag{24}$$

where ρ(k) is the weight that yields I + J; thus, it yields:

$$D(x(t_i), x(t_j)) = \frac{1}{I+J} \min_w \left[\sum_{k=1}^{k} d(c(k))\rho(k) \right]. \tag{25}$$

An optimal solution is computed with a minimum number of divergence values. Let the probability of getting a human-gait feature vector be β, such that it belongs to a class w$_i$, that yields p(β/w$_i$). Additionally, for class w$_j$, it yields p(β/w$_j$). Thus, it becomes:

$$D_{i,j} = (\mu_i - \mu_j)(\mu_i - \mu_j)^T \Sigma^{-1}, \tag{26}$$

where $\mu = \mu_i = \mu_j$ are the expectations and Σ is the covariance.

For the computation of step length, let the source be 'S$_G$' and the destination be 'D$_G$'. Let 'G' steps be the distance target to be achieved. Let the first human-gait frame have left foot (G$_L$) at the back and right foot (G$_R$) at the front, with the coordinates with (x,y) for the first human-gait frame, such that G$_L$(x$_1$,y$_1$) and G$_R$(x$_2$,y$_2$). Thus, it yields:

$$|step - length| = |x_2 - x_1| + |y_2 - y_1|. \tag{27}$$

G$_{act}$ human-gait steps are required to achieve the target. Thus, total steps are calculated as follows:

$$G_{calc} = G_1 + G_2 + G_3 + \cdots + G_n. \tag{28}$$

Furthermore, walking speed or walking rate is calculated as follows, and it yields:

$$\text{walking-speed} = \begin{cases} \text{norm}, & \text{if} \quad G_{act} = G_{calc} \\ \text{fast}, & \text{if} \quad G_{act} < G_{calc} \\ \text{slow}, & \text{if} \quad G_{act} > G_{calc} \end{cases}. \tag{29}$$

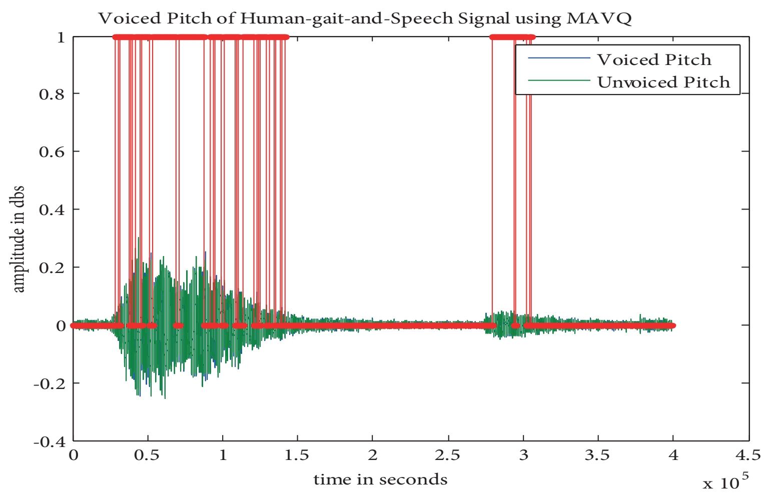

Figure 4. Voiced and unvoiced pitch pattern of 1D data (human speech).

3. Simulated results and discussion

The determination of behavioral trait patterns is a knowledge-intensive process, which must take into account all variable information about human-gait-speech data. In this paper, we have formed a vast corpus called the human-gait-speech model of human-gait-speech data of subjects of varying ages. Figure 4 represents the voiced and unvoiced pattern of human-speech data.

Initially, the enhanced human-speech signal (1D data) is segmented for the detection of voiced/unvoiced speech fragments. On employing the traditional zero-crossing measurement, the segmentation is achieved. However, it is still observed that the performance of the detection of the voiced speech has not improved. Hence, to overcome this problem, another method called modified adaptive vector quantization (MAVQ) is applied. The plot is shown in Figure 5. Comparing the plots shown in Figures 4 and 5, clear and correct voiced speech fragments are detected using the MAVQ method.

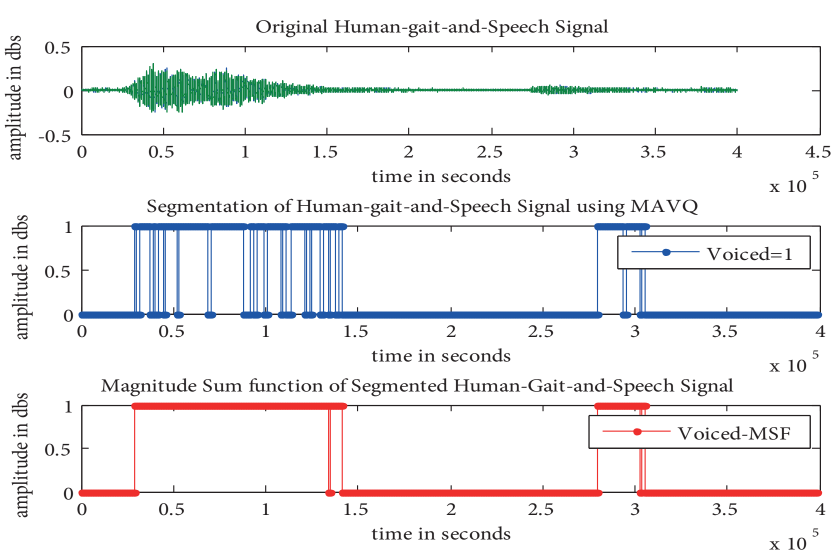

Figure 5. Estimation of pitch of 1D data (human speech).

The envelope of a speech signal is also known as the modulating speech signal. Thus, it is the outer shape of the speech signal. Figures 6 and 7 depict the envelope of the human-speech signal with autocorrelation and autoregression characteristics. Initially, the speech signal is divided into frames of 512 samples without windowing or overlapping. In each frame, a simple loop has been performed to find the maximum and minimum data within the frame. The maximum values found in each frame are linked together to form the top envelope curve. The minimum values in each frame are also linked to produce the lower envelope curve.

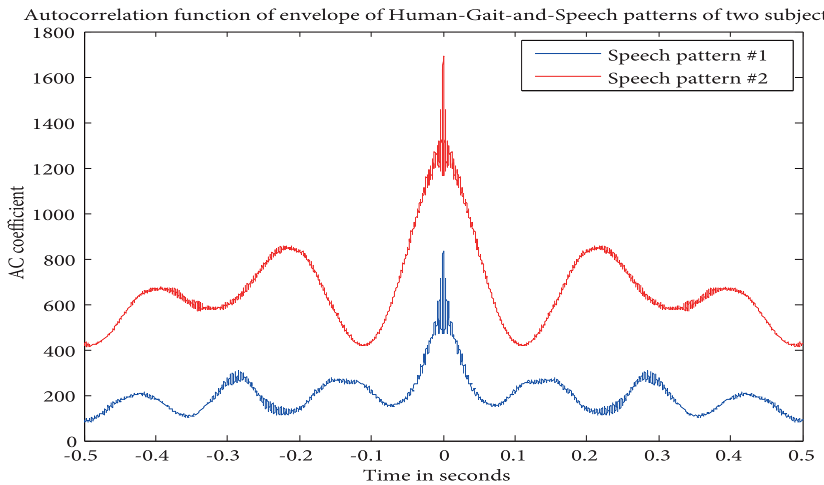

Figure 6. Autocorrelation of the envelope of 1D data (human speech) of different subjects and utterance of same language.

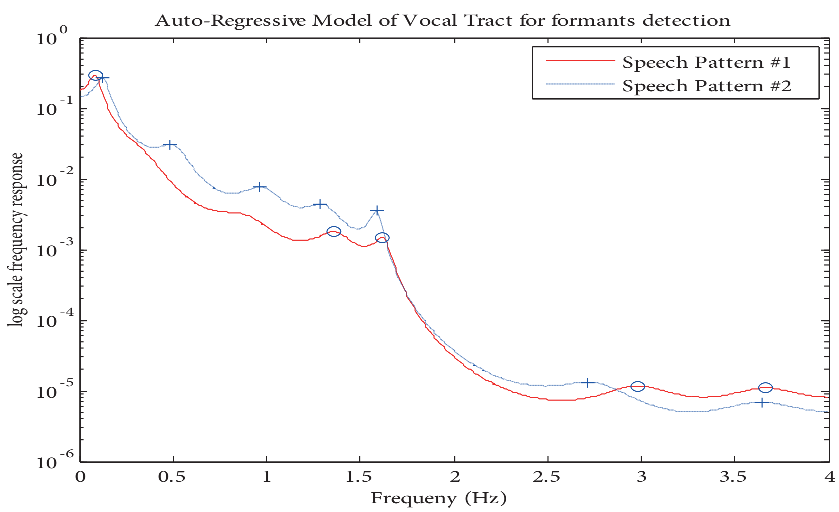

Figure 7. Autoregression of the envelope of 1D data (human speech) of different subjects and utterance of same language.

Figure 8 shows the energy count or the effort applied by the subject considering human-gait-speech data with the utterance of the same language while walking on a flat surface. Scientifically determining the start and end points of the speech signal is not an easy task, because the energy at the start of the speech signal is very low. Thus, in conjunction with the energy values and zero-crossing rate, the detection of the start and end points of the speech signal has been done in the present work. Algorithm 3, as given below, is used for calculating the energy values of the human-gait-speech pattern.

Algorithm 3 Computing energy values of human-gait-speech data.

1: Read the human-gait-speech signal, say, X(m,n)
2: Find the length of the frame, say, N = length(X(m,n))
3: Assign k = 0 as counter for the frame of samples
4: While k $<$ N do
5: energy_values $\leftarrow$ X(k,p) $\times$ X(k,p)
6: k $\leftarrow$ k + 1
7: End while
8: Return energy_values.

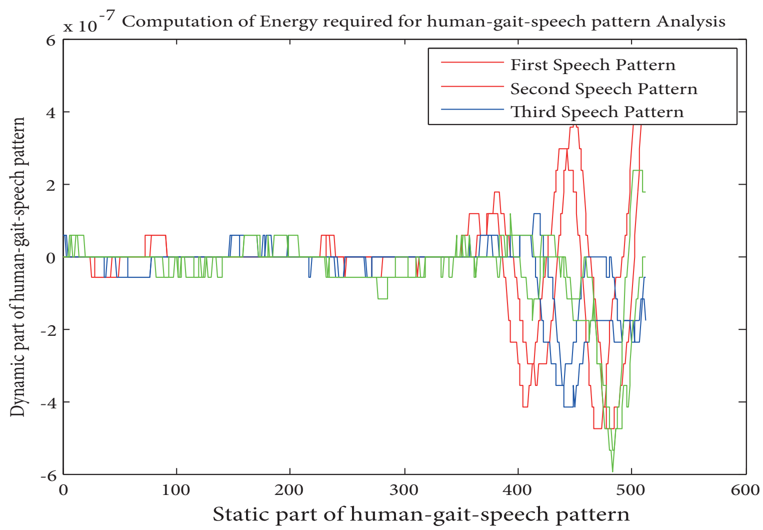

Figure 8. The 2D energy data (human gait) of different subjects and utterance of same language while walking on a flat surface.

4. Concluding remarks and further scope of the work

The present research was carried out for the formation of a corpus that will be helpful for the detection of behavioral patterns through human-gait-speech data after extraction of geometric and prosodic features. First, an input human-gait image was enhanced and compressed for distortion removal with lossless information. Second, it was segmented for contour detection. Third, the relevant geometric features were selected and, based on the geometric features, relevant parameters were extracted. Fourth, using an artificial neural network, the feature vectors were utilized for framing a corpus or knowledge-based model. In the second part of the research, an unknown human-gait-speech video stream was considered. The unknown human-gait-speech video stream was split into two components: audio and video of the subject. The gait video was converted into frames and then enhanced and compressed for distortion removal. Then it was segmented for extracting relevant geometric and prosodic features and parameter values.

The present work may be further applied to a human–computer interaction environment for the detection of human brain signals and their analysis, along with the human-gait-speech pattern, which will be a trimodal biometric security system. Future research may be carried out for the detection of various health problems related to breathing, walking, and speaking, as well as neurological and heart problems. Finally, the present work can be applied to promoting global multimodal biometrics-based security systems.

References

[1] O'Shaughnessy D. Speech Communications: Human and Machine. 2nd ed. Telangana, India: Universities Press, 1987.

[2] Goldberg DE. Genetic Algorithms in Search, Optimization and Machine Learning. 1st ed. Boston, MA, USA: Addison-Wesley, 1989.

[3] Jang JSR, Sun CT, Mizutani E. Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. 1st ed. Upper Saddle River, NJ, USA: Pearson Education, 1990.

[4] Hassoun MH. Fundamentals of Artificial Neural Networks. Cambridge, MA, USA: MIT Press, 1995.

[5] Cohen A. Biomedical Signal Processing Volume 2. Compression and Automatic Recognition. 1st ed. Boca Raton, FL, USA: CRC Press, 1986.

[6] Tolvi J. Genetic algorithms for outlier detection and variable selection in linear regression models. Soft Comput 2004; 8: 527-533.

[7] Sun H, Lam K, Chung S, Dong W, Gu M, Sun J. Efficient vector quantization using genetic algorithm. Neural Comput Appl 2005; 14: 203-211.

[8] Bação F, Lobo V, Painho M. Applying genetic algorithms to zone design. Soft Comput 2005; 9: 341-348.

[9] Sinha TS, Sanyal G, Mukherji A. Some aspects of modelling and simulation for the recognition of speech and speaker of Bengali language using proposed GSAMTSS algorithm. International Journal on Cybernetics & Informatics 2006: 69-75.

[10] Ghanbari Y, Karami MR. Spectral subtraction in the wavelet domain for speech enhancement. International Journal on Softwares & Information Technology 2004; 1: 26-29.

[11] Samangooei S, Nixon MS. Performing content-based retrieval of humans using gait biometrics. Semantic Multimedia 2008; 5392: 105-120.

[12] Bouchrika I, Goffredo M, Carter J, Nixon MS. On using gait in forensic biometrics. J Forensic Sci 2011; 56: 882-889.

[13] Sinha TS, Patra RK, Raja R. A comprehensive analysis of human gait for abnormal foot recognition using neuro-genetic approach. International Journal of Tomography & Statistics 2011; 16: 56-73.

[14] Wang C, Zhang J, Wang L, Pu J. Human identification using temporal information preserving gait template. IEEE T Pattern Anal 2012; 34: 2164-2176.

[15] Kao E, VanMaasdam P, Sheppard J. Image-based tracking with particle swarms and probabilistic data association. In: IEEE 2008 Swarm Intelligence Symposium; 21–23 September 2008; St. Louis, MO, USA. New York, NY, USA: IEEE. pp 1-8.

[16] Sinha TS, Chakravarty D, Patra RK, Raja R. Modelling and simulation for the recognition of physiological and behavioral traits through human gait and face images. In: Al-Asmari AK, editor. Discrete Wavelet Transforms: A Compendium of New Approaches and Recent Applications. Rijeka, Croatia: InTech, 2013. pp. 95-125.

[17] Cantiniaux S, Vaugoyeau M, Robert D, Horrelou-Pitek C, Mancini J, Witias T, Azulay JP. Comparative analysis of gait and speech in Parkinson's disease: hypokinetic or dysrhythmic disorders? J Neurol Neurosurg Psychiatry 2009; 81: 177-184.

[18] Janssen D, Schollhorn WI, Lubienetkzi J, Folling K, Davis K. Recognition of emotions in gait patterns by means of artificial neural nets. J Nonverbal Behav 2008; 32: 79-82.

[19] Roether CL, Omlor L, Christensen A, Giese MA. Critical features for the perception of emotion from gait. J Vis 2009; 9: 1-32.