Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 1182608, 9 pages
https://doi.org/10.1155/2022/1182608

Research Article

Improved Frame-Wise Segmentation of Audio Signals for Smart Hearing Aid Using Particle Swarm Optimization-Based Clustering

Tushar Mehrotra,[1] Neha Shukla,[2,3] Tarunika Chaudhary,[4] Gaurav Kumar Rajput,[1] Majid Altuwairiqi,[5] and Mohd Asif Shah[6]

[1]*College of Computing Sciences & IT, Teerthanker Mahaveer University, Moradabad, India*
[2]*Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad, India*
[3]*Affiliated to Dr. A P J Abdul Kalam Technical University, Lucknow, India*
[4]*Department of Computer Science & Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Delhi-NCR, Modinagar, Ghaziabad, Uttar Pradesh, India*
[5]*College of Computing and Information Technology, Taif University, Taif, Saudi Arabia*
[6]*Kebri Dehar University, Kebri Dehar, Ethiopia*

Correspondence should be addressed to Mohd Asif Shah; drmohdasifshah@kdu.edu.et

Received 23 January 2022; Revised 27 February 2022; Accepted 7 March 2022; Published 5 May 2022

Academic Editor: Vijay Kumar

Labeling speech signals is a critical activity that cannot be overlooked in any of the early phases of designing a system based on speech technology. For this, an efficient particle swarm optimization (PSO)-based clustering algorithm is proposed to classify the speech classes, i.e., voiced, unvoiced, and silence. A sample of 10 signal waves is selected, and their audio features are extracted. The audio signals are then partitioned into frames, and each frame is classified by using the proposed PSO-based clustering algorithm. The performance of the proposed algorithm is evaluated using various performance metrics such as accuracy, sensitivity, and specificity that are examined. Extensive experiments reveal that the proposed algorithm outperforms the competitive algorithms. The average accuracy of the proposed algorithm is 97%, sensitivity is 98%, and specificity is 96%, which depicts that the proposed approach is efficient in detecting and classifying the speech classes.

1. Introduction

The classification of speech into voiced, unvoiced, and silent (V/UV/S) frames is a critical and difficult topic that allows for pitch estimation, automated speech identifier, speaker identification, speech analysis, speech augmentation, and speech signal compression based on whether or not the vocal cords vibrate during the creation of the speech segment [1]. The silent segment in human speech is a period of quiet that may happen at the start of statements, between words/syllables, or after utterances. When the vocal cords aperiodically vibrate, unvoiced segments are generated [2]. When the vocal cords vibrate in a regular pattern, voiced segments are generated. The SUV segmentation is a more difficult classification issue than the two-class voiced activity detection (VAD) and voiced-unvoiced (VU) classifications since it is a three-class problem. The SUV segmentation can be accomplished by combining VAD and V/U segmentations, according to research. This will need previous knowledge of the speech signal's noise statistics, making the classification issue reliant on the noise statistics' correctness. As a result, the SUV segmentation is often viewed as a single issue.

1.1. Preliminaries. The preliminaries of the research are the parameters of the speech that are to be extracted for classification. The five parameters of speech are as follows (https://www.clear.rice.edu/elec532/PROJECTS00/vocode/uv/uvdet.html).

1.2. Zero Crossing. The frequency where the energy inside the signal spectrum is focused is indicated by the zero-

crossing count [3]. Voiced speech is produced by the periodic flow of air just at glottis activation of the vocal tract and has a low zero-crossing count in general. A noise-like source excites the vocal tract at a constriction in the interior of the vocal tract, leading to unvoiced speech with such a high zero-crossing count [4]. Silence is expected to get a lower zero-crossing frequency than unvoiced speech but equal to voiced speech.

1.3. Energy. Log energy E_v is as follows:

$$E_v = 10 \log\left(\lambda + \frac{1}{N}\sum_{m=1}^{M} S_p^2(m)\right). \quad (1)$$

Here, λ represents a tiny positive constant that prevents the log of zero from being computed. E_v shows the energy of spoken data is substantially greater than that of silence. N shows the zero-crossing count.

1.4. Normalized Autocorrelation Coefficient

$$C_1 = \frac{\sum_m^M s_p(m)s_p(m-1)}{\sqrt{\left(\sum_{m=1}^{M} S_p^2(m)\right)\left(\sum_{m=0}^{M-1} S_p^2(m)\right)}}. \quad (2)$$

C_1 defines that at a unit sample delay, the normalized autocorrelation coefficient.

The association between nearby speech samples is this metric. Because of the significant association between adjacent samples of voiced speech waveforms because of the high presence of low-frequency energy in voiced sounds, this value is near to 1 [5]. Unvoiced speech, on the other hand, has a connection that is close to zero.

1.5. The Predictor Coefficient. The initial predictor coefficient in a 12-pole linear predictive coding research using the covariance technique is α_1. At unit sample delay, this value may be shown to be the inverse of the log spectrum's Fourier component. The first LPC coefficient greatly deviates from the spectra of the voiced, unvoiced, and quiet classes [6].

1.6. Normalized Prediction Error

$$E_{np} = E_v - 10 \log\left(10^6 + \left|\sum_{j}^{q} \phi(0,j) + \phi(0,0)\right|\right), \quad (3)$$

$$\phi(h,k) = \frac{1}{N}\sum_{m=1}^{M} S_p(m-h)S_p(m-k), \quad (4)$$

where E_v is described above, $\phi(h,k)$ is the (h,k) of the speech samples' covariance matrix, and Ks are the predictor coefficients. This metric quantifies the nonuniformity of the spectrum [7].

The silence, voice, and unvoiced segmentation is a more difficult classification issue than the two-class voiced activity detection and voiced-unvoiced classifications since there is a three-class problem [8]. Previous research has solely looked at segmentation in terms of silent and nonsilent frames or voiced and unvoiced frames. Furthermore, fundamental metrics collected from the speech signal, including the signal's energy [9], zero-crossing rate, and degree of voice periodicity, were employed to achieve the V/UV classification. A single statistic from the speech signal, such as RMS energy or zero-crossing rate, may be used to detect VN/S signals. Because the quantity of any one parameter often overlaps across categories, such a technique may achieve only limited accuracy, especially when the speech is not captured in a high-quality context. For a long time, the V/U/S category has been engaged in defining the periodicity of speech [10]. Due to the fact that vocal fold vibration does not always result in a periodic signal, failure to recognize periodicity for voiced speech might result in a VN/S classification mistake [11].

When it comes to SUV categorization, one of the most significant considerations is the characteristics that must be employed. SUV categorization outperforms the others in terms of LPD-derived cepstrum and mel-frequencey cepstrum coefficients. Calculating the energy of a voice signal, on the other hand, is a very simple operation, with most algorithms relying on fundamental elements such as energy contours and zero crossings [12]. In the prior and current research, unsupervised learning, zero-crossing rate, pattern recognition algorithms, cumulates, autocorrelation algorithms, spectral parameters, and combinations of two or more of these methods have all been employed to construct SUV classification systems. The following are the techniques to classify voice, unvoiced, and silence:

1.6.1. Voiced Speech. When a system's input excitation is a nearly regular impulse sequence, the resultant speech is referred to as voiced speech, as it seems visually periodic (see Figure 1) [13].

1.6.2. Unvoiced Speech. Unvoiced speech occurs when the stimulation is random noise-like, and the resulting speech signal is likewise arbitrary noise-like with no periodicity.

The graphic depicts the nature of natural enthusiasm and the resulting unvoiced words [15]. The unvoiced utterance, as can be observed, will be nonperiodic. The major distinction between voiced and unvoiced speech will be this. The autocorrelation analysis may also detect the nonperiodicity of unvoiced speech (see Figure 2) [16].

1.6.3. Silence. The speech creation process involves the simultaneous development of vocal and unvoiced speech, separated by a silent period [17]. There is no stimulation delivered to the vocal tract during the silent phase; hence, there is no voice production. Silence, on the other hand, is a component of the speech signal. The speech will be incomplete if there is no quiet zone between vocal and unvoiced discourse. Silence, combined with other vocal or unvoiced words, may be used to identify particular sorts of

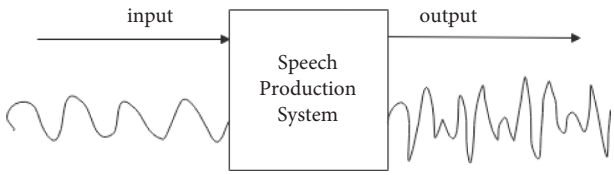

Figure 1: Voice speech [14].

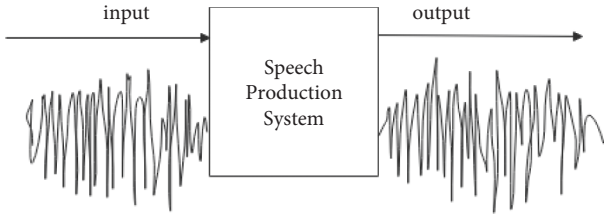

Figure 2: Unvoiced speech [13].

noises [18]. Even if the silent area is negligible in terms of amplitude/energy, its length is critical for comprehensible speech. The frequency where the energy inside the signal spectrum is focused is indicated by the zero-crossing count [3]. Voiced speech is produced by the periodic flow of air just at glottis activation of the vocal tract and has a low zero-crossing count in general. A noise-like source excites the vocal tract at a constriction in the interior of the vocal tract, leading to unvoiced speech with such a high zero-crossing count [4]. Silence is expected to get a lower zero-crossing frequency than unvoiced speech but equal to voiced speech.

This study proposed a novel PSO-based clustering method to categorize speech into three categories: quiet, voice, and unvoiced. These classes are grouped together based on the characteristics that were derived from them. Zero crossing, energy, normalized autocorrelation coefficient, predictor coefficient, and normalized prediction error are the five characteristics that may be extracted from speech. With the use of PSO-based clustering, an audio signal is partitioned into frames and classified according to its class by extracting these characteristics from the signal. Performance criteria such as accuracy, sensitivity, specificity, and confidence intervals are analyzed to show the usefulness of the suggested algorithm.

Furthermore, this study is organized as follows: Section 2 presents the related work. Section 3 describes the proposed methodology. Section 4 presents the comparative analyses. Finally, Section 5 concludes this study.

2. Related Work

Many of the existing works are developed on SUV classification and segmenting methods. Researchers implemented these segmentations by using different machine learning and clustering techniques. In [19], the author presented a novel approach for segmenting dysarthric speech into silent, unvoiced, and voiced pieces. Short-time energy, zero-crossing rate, and linear prediction error variance are used to solve the segmentation issue in this example. A moving average threshold technique is presented to give completely

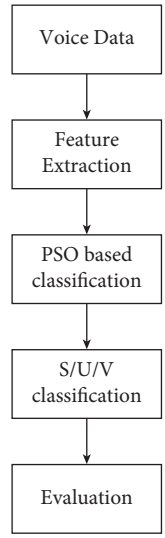

Figure 3: Flowchart of the proposed methodology.

automated "as-fit" major components that can handle highly acute dysarthric speech with changing loudness and ZCRs. The capabilities of the proposed totally automated method are validated using real-world audio signals from healthy and ataxic dysarthric speakers. [20]. As per the findings of the article, the proposed classification strategy not only increased segmentation results but also gave reliable results in low-effort settings.

A voiced-unvoiced-silence classification technique based on a time-frequency description of the measured signal, regarded as a data matrix, is proposed. The study [21] is based on a hierarchical dual-geometry data matrix analysis that makes use of the tight connection between time frames and frequency bins. The method allows for the separation of spoken and quiet frames, and then voiced and unvoiced frames, by progressively learning the associated geometry in two phases. A multilayer feedforward network was used to classify speech into voiced-unvoiced-silence. A maximum-likelihood classifier was used to analyze the network [22]. Using an MFN, a process for classifying speech into voiced, unvoiced, and silent was devised. The network VIUIS classifier is projected to be a valuable tool in this study for speech analysis and for speech-data mixed communication systems.

Using unsupervised learning provided [23] a unique voiced-unvoiced-silence categorization. Using Gaussian mixture models and the expectation-maximization approach, class-dependent statistics such as feature means, covariance matrices, and frequency probabilities of voiced, unvoiced, and silent classes are directly generated from the signal. The NTIMIT database was utilized to evaluate the learning-based categorization, and the dataset illustrated the accuracy of a completely learned classification. To remove noise from speech signals, an improved speech enhancement technique wavelet-based and spectra speech classification is proposed. Using a unique energy-based threshold, the technique splits speech into voiced, unvoiced, and silent sections before applying the wavelet transform. The detailed

```
Initialize signals with a constant value of energy and deploy in a specified area.
Initially, assign 10% of n signals to cluster nodes at random.
For i: 1 to n
Calculate distance (F1) = di, m + 1
Minimum distance = Euclidean distance value from signal i to BSm + 1
If (minimum distance > distance)
Node.id = i
End if
For j: 1 to m (total number of CH)
Calculate distance (F2) = djm (distance from signal j to CHm)
If (minimum distance > distance)
Minimum distance = distance
Node.id = j
End if
Store the distance in an array (A) which maintains values for clusters
A (Node.id).sum = A (Node.id).sum + minimum_distance
A (Node.id).num = A.(Node.id).num + 1
   End for
End for
For k: 1 to m (total number of CH)
Calculate distance (F3) = dk, m + 1
If (minimum distance > distance)
Minimum distance = distance
Cluster.id = k
End if
Store the distance in an array (A) which maintains values for clusters
A (cluster.id).sum = A (cluster.id).sum + minimum_distance
A (cluster.id).num = A(cluster.id).num + 1
End for
Calculate the total energy
Fitness function value for each node
For i: 1 to n
Fitness (clustering) = (0.25 ∗ F1) + (0.25 ∗ F2) + (0.25 ∗ F3) + (0.25 ∗ F4)
End for
Stop
```

Algorithm 1: PSO-based clustering for audio speech signals.

coefficients are thresholded to reduce noise, taking into account the distinctive properties of speech in each of the three domains [24]. For vocal parts, soft thresholding is employed, hard thresholding is used for unvoiced regions, and the wavelet coefficients for quiet zones are set to zero. The proposed technique is tested using white noise-contaminated utterances from the SPEAR collection. The technique generated better results in terms of output SNR, PESQ score, speech waveforms, and spectrograms.

The author presented a digital architecture for classifying noise-free voice segments in an instantaneous V/UV/S manner. The proposed [25] architecture computes two commonly utilized time-domain-based speech metrics, brief energy (STE), and relatively short zero-crossing rate, using the incoming sample of the speech segments. The hardware required to do on-the-fly calculations of the specified parameters is included in an algorithm state machine with such a data path (ASMD). Inside the ASMD, the decision model was implemented as a separate unit. To use the Xilinx ZedBoard Zynq Evaluation and Development Platform XC7Z020CLG484-1, the suggested architecture is prototyped on a ground gate array (FPGA). It has a maximum operating clock frequency of 185 MHz and is fully compatible with prior CORDIC-based window designs.

In [26], a hybrid CNN with long short-term memory (LSTM) is used to automatically extract ambient and microphone information from the spoken sound. In the trials, it was also looked at how the usage of voiced and unvoiced chunks of speech affected the accuracy of such environment and microphone classification. The suggested method employs a subset of the KSU-DB corpus that contains three settings, four kinds of recording equipment, 136 speakers, and 3,600 word, phrase, and speech signal recordings. In this work, the CRNN model was established, which incorporates elements of both CNN and RNN models. Speech signals were recorded as spectrograms and sent into the CRNN model as 2D images.

From the literature, it is found that the existing models suffer from various problems such as poor convergence speed [27, 28] and are stuck in local optima [29–31], premature convergence speed [32, 33], gradient vanishing [34, 35], etc. Besides designing the efficient fitness function, there remains a challenge for real-time applications of metaheuristic techniques [31, 36].

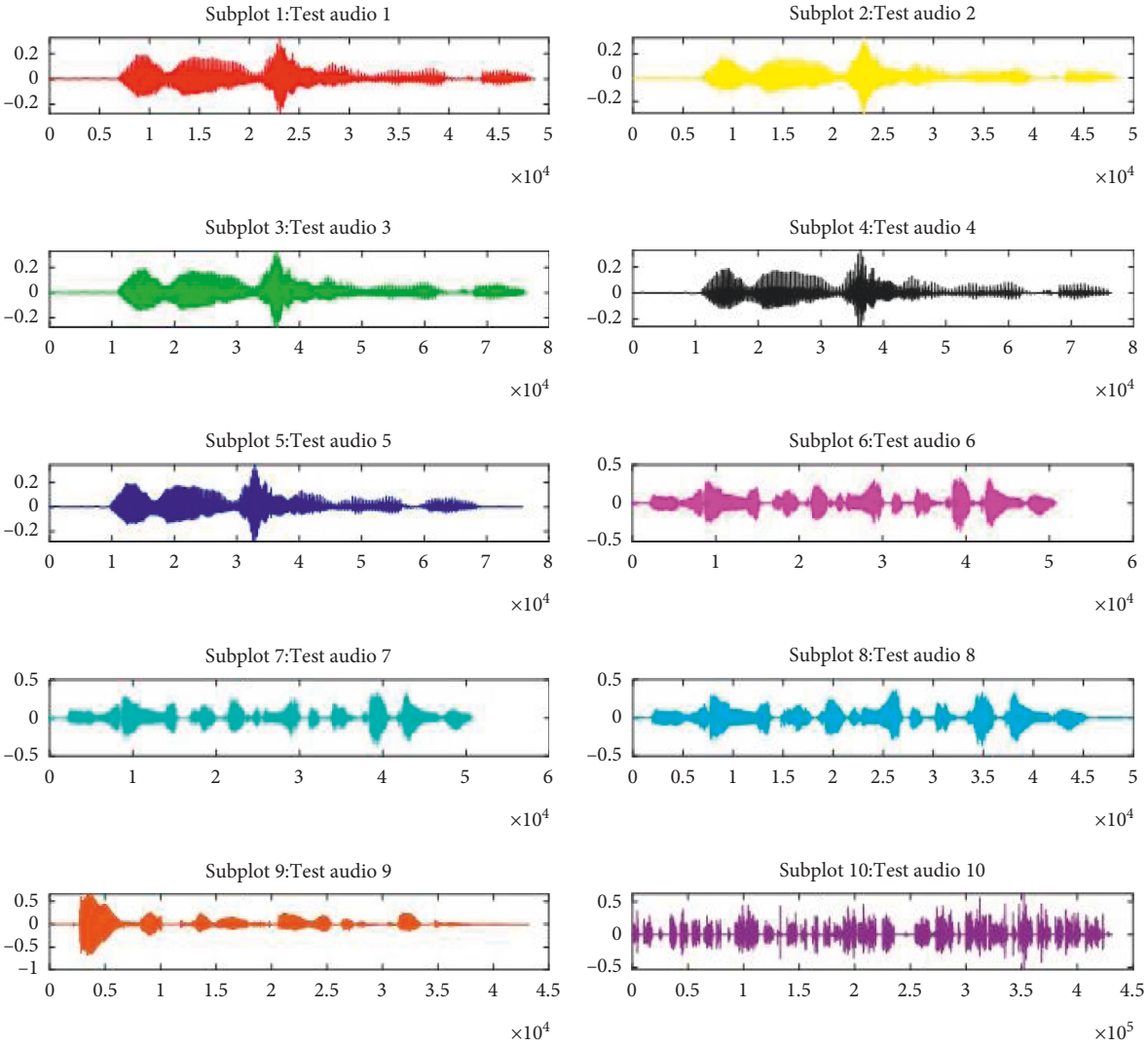

Figure 4: Test audio signals.

2.1. Contributions of the Study. The main contributions of the study are to:

(i) Classify the speech classes as voice, unvoiced, or silence by using particle swarm optimization (PSO)-based classification.

(ii) Measure the efficacy of the proposed algorithm with performance parameters like accuracy, sensitivity, and specificity.

Furthermore, this study is organized in such a way that Section 2 describes the proposed methodology and Section 3 displays the results. Finally, Section 4 concludes the work.

3. Methodology

In this work, a PSO-based clustering algorithm is proposed to classify the speech classes, i.e., silence, voice, and unvoiced. These classes are clustered on the basis of their extracted features. The five features that are retrieved from the speech are zero crossing, energy, normalized autocorrelation coefficient, predictor coefficient, and normalized prediction error. An audio signal is partitioned into frames and segmented according to its class by extracting these features using PSO-based clustering. The flowchart of the proposed methodology for a brief understanding is presented in Figure 3.

To illustrate the efficiency of the suggested algorithm, performance parameters like accuracy, sensitivity, specificity, and confidence intervals are evaluated.

3.1. Particle Swarm Optimization (PSO). The PSO is a population-based optimization algorithm that is inspired by flocks of birds' social behavior. The PSO is often referred to as an example of evolutionary computation. A swarm of particles moves across the search zone in a PSO system. Each particle represents a potential solution to the issue of optimization [37]. The best place visited by the particle and the particle's position in the particle's vicinity impact the particle's location. When a particle's neighborhood is the whole swarm, the global best particle is the best location in the neighborhood, and the process that follows is known as the

Table 1: Information of audio waves.

Sample	Number of samples	Frequency	No. of frames	Frame length
Audio signal-1	48622	44100	95	512
Audio signal-2	48622	44100	95	512
Audio signal-3	76404	44100	149	512
Audio signal-4	76404	44100	149	512
Audio signal-5	75776	44100	148	512
Audio signal-6	50847	20000	99	512
Audio signal-7	50847	20000	99	512
Audio signal-8	49920	20000	98	512
Audio signal-9	43213	16000	85	512
Audio signal-10	4309544	48000	8417	512

gbest PSO. The technique is known as the West PSO when small neighborhoods are adopted. The optimization problem provides a fitness function that is used to assess the performance of each particle.

The following qualities are represented by each particle in the swarm: u_x = particle's current position. j_x = particle's best personal position. ve_x = particle's present velocity.

Particle x s personal best position is the best position that particle x has visited thus far. The objective function is denoted by the letter f. The particle's personal best at time step ts is then updated as

$$y_i(ts+1) = \begin{cases} j_x(ts) & if\ f(u_x(ts+1)) \geq f(j_x(ts)), \\ u_x(ts+1)\, if\ f(u_x(ts+1)) \geq f(j_x(ts)), \end{cases} \quad (5)$$

*l*best and gbest are two basic techniques to PSO, with the distinction being in the neighborhood topology utilized to trade experience among particles. The best particle in the gbest model is chosen from the whole swarm. If the vector $\hat{y}$ denotes the location of the global best particle, then

$$\hat{j}(ts) \in \{j_0, j_1, ..., j_s\} = \min\{f(j_0(ts)), f(j_1(ts)), ..., f(j_s(ts))\}. \quad (6)$$

Here, s = size of the swarm, and A swarm is partitioned into overlapping zones of particles in the lbest model. This particle is known as the best particle in the neighborhood, and it is defined as

$$\begin{aligned} N_y &= \{j_{x-1}(ts), j_{x-l+1}(ts), ..., j_{x-1}(ts), j_x(ts), j_{x+1}(ts), \\ &\quad ..., j_{x+l-1}(ts), j_{x+1}(ts)\}, \\ \hat{j}_y(ts+1) &\in \left\{N_y \mid f\left(\hat{j}_y(ts+1)\right) = \min\{f(j_x(ts))\}, \forall_{j_x} \in N_y\right\}. \end{aligned} \quad (7)$$

Particle indices are often employed to identify neighborhoods, although topological neighborhoods may also be utilized. The gbest is a special case of *l*best with $l = s$, in which the neighborhood is the whole swarm [38]. Although the lbest PSO has more variety than the gbest PSO, it is also slower. The remainder of the chapter focuses on the gbest PSO, which is quicker.

The velocity ve_x and location u_x are changed as follows for each iteration of gbest:

$$\begin{aligned} ve_x(ts+1) &= wve_i(ts) + c_1 r_1(ts)\left(j_i(ts) - u_i(ts)\right) \\ &\quad + c_1 r_1(ts)\left(\hat{j}(ts) - u_x(ts)\right), \end{aligned} \quad (8)$$

Table 2: Performance parameters.

Sample	Accuracy	Sensitivity	Specificity
Audio signal-1	0.9789	0.9842	0.9684
Audio signal-2	0.9719	0.9789	0.9579
Audio signal-3	0.9955	0.9966	0.9933
Audio signal-4	0.9821	0.9866	0.9732
Audio signal-5	0.9820	0.9865	0.9730
Audio signal-6	0.9798	0.9848	0.9697
Audio signal-7	0.9731	0.9798	0.9596
Audio signal-8	0.9728	0.9796	0.9592
Audio signal-9	0.9765	0.9824	0.9647
Audio signal-10	0.9823	0.9868	0.9735

Table 3: Confidence levels of parameters.

Parameter (%)	Accuracy	Sensitivity	Specificity
90	0.00428	0.00643	0.00643
95	0.00691	0.00518	0.01038

$$u_x(ts+1) = u_x(ts) + v_x(ts+1). \quad (9)$$

Here, w = inertia weight, and c_1, c_2 = acceleration constants. The above equation consists of 3 components, namely, the word inertia refers to the body's ability to remember prior speeds. The influence of the preceding velocity is controlled by the inertia weight: exploration is favored by a high inertia weight, but exploitation is favored by a low inertia weight. The cognitive aspect, $j_x(ts) - u_x$, reflects the particles' knowledge of the optimum solution. The social aspect, $\hat{j}(ts) - u_x(ts)$, symbolizes the swarm's collective conviction on the optimum option. Various social topologies have been studied, with the star topology being the most popular.

To do a classification approach, you must first establish a fitness function. PSO, like other swarm intelligence approaches, has been defined to undertake a search in the space of solutions to maximize outcomes in situations with single and multiple objectives. It has been established that PSO may provide better outcomes in a quicker and less expensive manner than other approaches. It is also possible to parallelize it. Furthermore, it does not take advantage of the gradient of the issue to be optimized. In other words, unlike classic optimization approaches, PSO does not need a differentiable problem. It is becoming more popular as a result of its many benefits such as resilience, efficiency, and simplicity. PSO has been discovered to need less computing

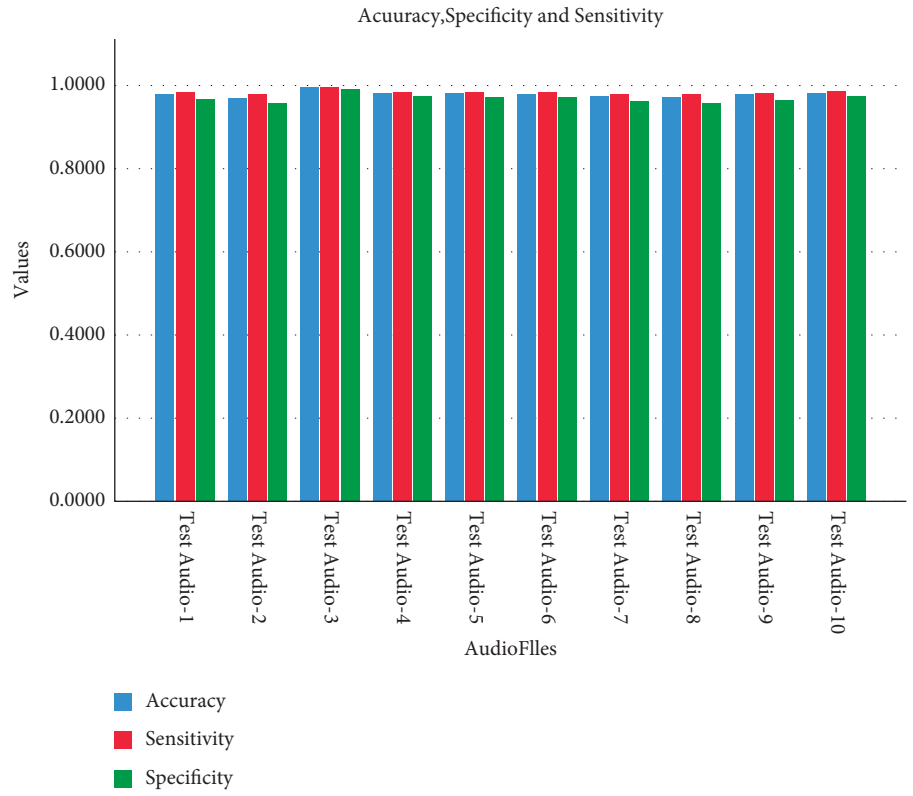

Figure 5: Performance parameters.

effort than other stochastic algorithms. As a result, this is an efficient optimization strategy for classification issues.

The PSO approach repeats the update equations above until the velocity updates are close to zero or until a certain number of iterations have been completed [39]. Particle quality is assessed using a fitness function that determines the optimality of the relevant solution.

This proposed algorithm will be used to cluster audio signals according to their respective classes.

This proposed algorithm 1 will be used to cluster audio signals according to their respective classes. The performance is measured with the following parameters.

3.2. Performance Metrics. The confusion matrix's performance characteristics, such as accuracy, sensitivity, and specificity, are used to assess the suggested algorithm's performance.

Accuracy: It is the fraction of correctly recognized subjects to the total number of subjects.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{10}$$

Sensitivity: Recall, also known as sensitivity, is the proportion of correctly positive labels recognized by our computer.

$$Sensitivity = \frac{TP}{TP + FN}. \tag{11}$$

Specificity: The system has appropriately classified the negative as specificity.

$$Specificity = \frac{TN}{TN + FP}, \tag{12}$$

where TP = true positive, TN = true negative, and FP = false positive.

4. Performance Analyses

The proposed PSO-based clustering method is developed and tested in this section, and the results are examined. Performance measures such as accuracy, sensitivity, and specificity are examined to illustrate the usefulness of the suggested algorithm.

Figure 4 shows the wave signals of chosen 10 test audio samples.

Table 1 depicts the information of the chosen data samples. For implementation, 10 test audio samples are taken and their sample size, frequency, frame size, and frame length are extracted and displayed. The number of frames is generated by partitioning the audio signal. These frames are clustered according to their respective classes.

Table 2 exhibits the audio samples' performance metrics such as accuracy, sensitivity, and specificity. The average accuracy of the 10 test samples is 0.9794, or 97%, the average sensitivity value is 0.9846, or 98%, and the average specificity value is 0.9692, or 96%. These numerical values of the metrics are visualized in Figure 4. It shows the accuracy, specificity, and sensitivity of the chosen 10 test audio signals.

From Table 3, it is observable that the performance parameters vary over its mean value with confidence intervals of 90% and 95%. Accuracy varies by the values of 0.00428 and 0.00691, respectively. Sensitivity varies by the values of 0.00643 and 0.00518 and specificity with a confidence intervals of 90% and 95% by the values of 0.00643 and 0.01038, respectively. Figure 5 shows the performance parameters for the tested audio signals on the basis of accuracy, specificity, and sensitivity.

5. Conclusions

In this work, MATLAB 2020a is used for the implementation. The proposed algorithm particle swarm optimization (PSO)-based clustering algorithm is used to classify the three speech classes. These classes are silence, voice, and unvoiced. These classes are clustered based on extracted features. The five features that are retrieved from the speech are zero crossing, energy, normalized autocorrelation coefficient, predictor coefficient, and normalized prediction error. A sample of 10 audio signals is chosen for the implementation. Each audio wave is partitioned into frames, and each frame is clustered into either voice, unvoiced, or silence. In order to demonstrate the effectiveness of the proposed algorithm, performance parameters like accuracy, sensitivity, specificity, and confidence intervals are evaluated. The average accuracy of the audio samples is 97%, sensitivity is 98%, and specificity is 96%, which demonstrates that the proposed algorithm is highly accurate in clustering the speech classes. The average accuracy of the audio samples is 97%, sensitivity is 98%, and specificity is 96%, which demonstrates that the proposed algorithm is highly accurate in clustering the speech classes. This allows the smart hearing aid to distinguish between silence, voice, and unvoiced sounds.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. N. Mohammed and A. K. Hassan, "Automatic voice activity detection using fuzzy-neuro classifier," *Journal of Engineering Science & Technology*, vol. 15, no. 5, pp. 2854–2870, 2020.

[2] C. Montacié and M.-J. Caraty, "Phonetic, frame clustering and intelligibility analyses for the interspeech 2020 compare challenge," in *Proceedings of the Interspeech 2020*, vol. 2020, pp. 2062–2066, Shanghai, China, 2020.

[3] L. A. L. Er, *A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features*, IEEE, 2020.

[4] M. J. Al Dujaili, A. Ebrahimi-Moghadam, A. Fatlawi, and A. Fatlawi, "Speech emotion recognition based on SVM and KNN classifications fusion," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, pp. 1259–1264, 2021.

[5] H. Aouani and Y. B. Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251–260, 2020.

[6] S. Kumar, "Real-time implementation and performance evaluation of speech classifiers in speech analysis-synthesis," *ETRI Journal*, vol. 43, no. 1, pp. 82–94, 2021.

[7] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 45–55, 2020.

[8] K. Strenilkov, J. Debladis, J. Salles et al., "A study of voice and non-voice processing in Prader-Willi syndrome," *Orphanet Journal of Rare Diseases*, vol. 15, no. 1, pp. 1–12, 2020.

[9] J. S. Latha and G. U. Rani, "Devanagari script using energy of speech signal," *Journal of Science and Technology*, vol. 6, no. 1, pp. 167–171, 2021.

[10] N. H. Tandel, H. B. Prajapati, and V. K. Dabhi, "Voice recognition and voice comparison using machine learning techniques: a survey," in *Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 459–465, Coimbatore, India, 6-7 March 2020.

[11] E. Kalhor and B. Bakhtiari, "Speaker independent feature selection for speech emotion recognition : a multi-task approach," *Multimedia Tools and Applications*, vol. 80, pp. 1–20, 2020.

[12] R. A. Sharon and H. A. Murthy, "The 'Sound of Silence' in EEG - cognitive voice activity detection," 2020, https://arxiv.org/abs/2010.05497.

[13] B. M. S. Rani, A. J. Rani, T. Ravi, and M. D. Sree, "Basic fundamental recognition of voiced unvoiced and silence region of A speech," *International Journal of Engineering and Advanced Technology*, vol. 2, pp. 2249–8958, 2014.

[14] N. Agrawal, A. Jain, and A. Agarwal, "Simulation of network on chip for 3D router architecture," *International Journal of Recent Technology and Engineering*, vol. 8, no. 1C2, pp. 58–62, 2019.

[15] S. Id, *Voice_classification.docx*, 2021.

[16] A. Jain, A. Kumar, and S. Sharma, "Comparative design and analysis of mesh, torus and ring NoC," *Procedia Computer Science*, vol. 48pp. 330–337, C, 2015.

[17] D. Ghai, H. K. Gianey, A. Jain, and R. S. Uppal, "Quantum and dual-tree complex wavelet transform-based image watermarking," *International Journal of Modern Physics B*, vol. 34, no. 04, pp. 2050009–2050011, 2020.

[18] H. M. Chandrashekar, K. S. Pavithra, V. Karjigi, and N. Sreedevi, "Region based prediction and score combination for automatic intelligibility assessment of dysarthric speech," in *Proceedings of the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 407–412, Greater Noida, India, 19-20 Feb. 2021.

[19] T. Ijitona, H. Yue, J. Soraghan, and A. Lowit, "Improved silence-unvoiced-voiced (SUV) segmentation for dysarthric speech signals using linear prediction error variance," in *Proceedings of the 2020 5th Int. Conf. Comput. Commun. Syst. ICCCS*, pp. 685–690, Shanghai, China, 15-18 May 2020.

[20] A. Jain, A. K. Gahlot, R. Dwivedi, A. Kumar, and S. K. Sharma, "Fat tree NoC design and synthesis," *Advances in Intelligent Systems and Computing*, vol. 624, pp. 1749–1756, 2018.

[21] A. Jain, R. Kumar Dwivedi, H. Alshazly, A. Kumar, S. Bourouis, and M. Kaur, "Design and simulation of ring network-on-chip for different configured nodes," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 4085–4100, 2022.

[22] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 250–255, 1993.

[23] H. Deng and D. O'Shaughnessy, "Voiced-unvoiced-silence speech sound classification based on unsupervised learning," in *Proceedings of the 2007 IEEE International Conference on Multimedia and Expo*, pp. 176–179, Beijing, China, 2-5 July 2007.

[24] M. Greenwood and A. Kinghorn, "SUVing: automatic silence/unvoiced/voiced classification of speech," pp. 3–6, Coursework Dep, 1999, http://staffwww.dcs.shef.ac.uk/xzprxz/M.Greenwood/uni/speech1.pdf Available:.

[25] N. S. S. Srinivas, N. Sugan, L. S. Kumar, and M. Kumar Nath, "Digital architecture for instantaneous V/UV/S classification of noise free speech segments," in *Proceedings of the 2020 24th International Symposium on VLSI Design and Test (VDAT)*, Bhubaneswar, India, 23-25 July 2020.

[26] M. A. Qamhan, H. Altaheri, A. H. Meftah, G. Muhammad, and Y. A. Alotaibi, "Digital audio forensics: microphone and environment classification using deep learning," *IEEE Access*, vol. 9, pp. 62719–62733, 2021.

[27] S. Bai, S. Song, S. Liang, J. Wang, Bo Li, and E. Neretin, "UAV maneuvering decision-making algorithm based on twin delayed deep deterministic policy gradient algorithm," *Journal of Artificial Intelligence and Technology*, vol. 2, no. 1, pp. 16–22, 2022.

[28] P. K. Singh, "Data with non-Euclidean geometry and its characterization," *Journal of Artificial Intelligence and Technology*, vol. 2, no. 1, pp. 3–8, 2022.

[29] M. Kaur and D. Singh, "Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 2483–2493, 2021.

[30] X. Zhang and G. Wang, "Stud pose detection based on photometric stereo and lightweight YOLOv4," *Journal of Artificial Intelligence and Technology*, vol. 2, no. 1, pp. 32–37, 2022.

[31] T. Singh, N. Saxena, M. Khurana, D. Singh, M. Abdalla, and H. Alshazly, "Data clustering using moth-flame optimization algorithm," *Sensors*, vol. 21, no. 12, p. 4086, 2021.

[32] R. Chen, D. Pu, Y. Tong, and M. Wu, "Image-denoising algorithm based on improved K-singular value decomposition and atom optimization," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 1, pp. 117–127, 2022.

[33] K. Yadav, M. Yadav, and S. Saini, "Stock values predictions using deep learning based hybrid models," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 1, pp. 107–116, 2022.

[34] H. Kaushik, D. Singh, M. Kaur, H. Alshazly, A. Zaguia, and H. Hamam, "Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models," *IEEE Access*, vol. 9, pp. 108276–108292, 2021.

[35] J. Zhang, G. Ye, Z. Tu et al., "A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 1, pp. 46–55, 2020.

[36] A. S. Kini, A. N. Gopal Reddy, M. Kaur et al., "Ensemble deep learning and internet of things-based automated COVID-19 diagnosis framework," *Contrast Media and Molecular Imaging*, vol. 2022, pp. 1–10, 2022.

[37] J. Cai, H. Wei, H. Yang, and X. Zhao, "A novel clustering algorithm based on DPC and PSO," *IEEE Access*, vol. 8, pp. 88200–88214, 2020.

[38] S. Paul, S. De, and S. Dey, "A novel approach of data clustering using an improved particle swarm optimization based K-means clustering algorithm," in *Proceedings of the 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2-4 July 2020.

[39] C. Velázquez, L. Cagnina, and M. Errecalde, *A PSO-Based Clustering Approach Assisted by Initial Clustering Information*, 2012.