Creating Speech Synthesis Process by Using Pulse Model in Log-Domain Vocoder With Whale Optimized Deep Convolution Neural Networks

Radha N (radhan@ssn.edu.in)  
Sri Sivasubramaniya Nadar College of Engineering

Vijayalakshmi P  
Sri Sivasubramaniya Nadar College of Engineering

Research Article

Keywords: Speech synthesis, vocoder, Pulse model in log-domain vocoder, Mel-generalized cepstrum

DOI: https://doi.org/10.21203/rs.3.rs-436506/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Creating Speech Synthesis process by using Pulse model in log-domain vocoder with whale optimized deep convolution neural networks

Radha N*1, Vijayalakshmi P1

1Sri Sivasubramaniya Nadar College of Engineering
Chennai, India

*Corresponding author: radhan@ssn.edu.in

Abstract

Speech synthesis is an artificial production of human speech by utilizing computer systems called a speech synthesizer. Several statistical parameters are used in previous works to perform the speech synthesis process, but the vocoder is combined only the simpler model. Due to the lack of sequence of modeling, the quality degrading process is reducing the speech synthesis. Therefore, the Pulse model in log-domain vocoder with whale optimized deep convolution recurrent neural network is applied to investing the vocoder in this work. During this analysis, Mel-Generalized Cepstrum (MGC), maximum voice frequency (MVF), and F0 are applied to processing the signal to extracting the features, and the vocoder is generated successfully. The system's effectiveness is then evaluated using experimental results compared to deep neural networks and traditional recurrent networks.

Keywords:

Speech synthesis, vocoder, Pulse model in log-domain vocoder, Mel-generalized cepstrum

1. Background Study

Speech synthesis [1-4] is nothing but artificial reproduction of people's speech with a computer device's help. This speech synthesis process is mainly utilized to convert the text into the audio information applicable for different applications [5] like mobile applications, voice-enabled services, etc. Significantly, the synthesis process is used as assistive technology to vision-impaired individuals. Therefore, in bell laboratory, the Home Dudley voder [6] develops its initial practical voice synthesizer. The created synthesizer quality is determined by computing the similarity of synthesized speech with original human notes or voice. Several operating
systems adapt the speech synthesizer process in the 1990s, which is done by concatenating the recorded speech. The discussed speech synthesis [7,8] is performed by preprocessing and speech synthesis. The preprocessing step helps to remove the ambiguity present in words such as homographs. After eliminating the noise, speech synthesis is performed by converting the text into a series of sounds. Although the text-to-speech synthesis process [9-12] is successfully utilized in different techniques, synthetic quality, flexibility, controls, and robustness are significant problems. In addition to this inadequate acoustic model, capability creates the non-linear relationship between acoustic and linguistic features. To overcome this issue, several researchers analyze the synthesis process for improving the system's overall efficiency. Most researchers utilize the Hidden Markov model (HMM) [13] to create the speech synthesis proc, but it has more computation complexity while analyzing many parameters.

Therefore, deep learning techniques [14-16] are utilized in the speech synthesis process to successfully analyze the high-dimensional data that extracts the high-level information. Due to the effective analysis, learning process, and easily adaptation characteristics, deep learning techniques incorporate the different neural network types. These networks are effectively mapping the linguistic features with the vocoding parameter, and the decisions are carried out. By considering the advantage of the deep learning techniques in the speech synthesis process, various analysts utilize this deep learning concept to create the speech synthesis. Discussing different researcher opinions, ideas, thoughts, and framework for developing the speech synthesis system. (Kolokas N., et al.,2018) [17] creating the recurrent neural network-based keyword to text synthesis system. Initially, keywords are obtained from users; parts of the speech-tagging library are utilized to examine the keywords. From the generated keywords, the network maps the related text for generating the synthesis. The identified textsynthesis is more related to keywords, and the system is developed with minimum cost and complexity. (Al-Radhi M.S., et al., 2017) [18] addressing the discontinuous in traditional vocoders by applying the sequence-to-sequence modeling and convolution neural networks. During this process, gated recurrent networks, bidirectional long-short term memory neural networks, and long-short term memory networks are continuously examined. These neural networks are applied to the Mel-Generalized Cepstrum and maximum voice frequency using the continuous vocoder from the analysis. The process helps generate speech synthesis and reduces the hightime consuming than traditional feed-forward deep learning neural networks.
(Bollepalli, et al., 2019) [19] applying recurrent and long-short term memory neural network for creating the normal-to-Lombard adaptation based speech synthesis process. This process uses the three adaptation methods, such as learning hidden unit contribution, auxiliary features, and fine-tuning. These adaptation models are work along with the neural networks that change the text to the speech synthesis process. The created synthesis system is evaluated using similarity tests and speech intelligibility teststo assess its efficiency. (Anumanchipalli, et al., 2019) [20] developing the speech synthesis system from spoken speech using neural decoding. This process uses the recurrent neural networks for examining the cortical activities and articulatory movements because it helps to investigate the speech acoustics and easy to synthesize the spoken speech. In addition to this, neuroprosthetic techniques are applied in the articulatory decoding process to improving the communication process

(Y. Joo et al., 2016)[21] developing the speech synthesis system from bottleneck features using cascading deep learning neural networks. The network consists of bottleneck-to-speech and text-to-bottleneck models to developing the speech synthesis. During this process, the speaker's acoustic and linguistic features are analyzed in bottleneck features. The examined characteristics are analyzed using mapping rules to determine the synthesis speech. The created cascading network-based created system attains higher results compared to traditional structures. (M. Hwang et al., 2020) [22] improving the LPCNet vocoder performance by applying the linear prediction structured mixture density networks. In this process, an autoregressive neural vocoder examines the vocal source and vocal tract components. Then continuous density distribution process is applied to combining linear prediction structure mixture density model with the LPCNet vocoder. This incorporating process creating the persuasive speech synthesis process means the quality of the speech is enhanced effectively. The LPCNet based vocoder performance attains the 4.41 opinion score while creating the text-to-speech structure.

(Y. Zhao et al., 2018) [23] developing the multi-speaker speech synthesis system by applying the wavenet vocoder. The system analyzes the text-to-speech acoustic and natural features and reducing the mismatched characteristics. During this process, the generative adversarial network is utilized as the acoustic model and produces the output used in the wavenet. This process is further enhanced by applying the discretized mixture of logistics loss model that reduces the error rate while creating the text-to-speech synthesis system. (M. Airaksinen et al., 2018)[24]
analyzing different vocoders such as sinusoidal vocoding, glottal, and straight in statistical parametric based speech synthesis. During this process, text-to-speech synthesis split the text into vocoder-related features, and synthesis is done using the share envelope model. Then the wave has been generated by using these vocoders, and the efficiency of the system is analyzed using four voices. From the analysis, glottal vocoder attains significant results compared to other vocoders. Recent studies directly show that a deep learning network successfully creates the speech synthesis process. Even though these approaches fail to meet the accuracy due to the lack of sequence modeling, the quality degrading process reduces speech synthesis. For overcoming this issue in this work, the Pulse model in log-domain vocoder with whale optimized deep convolution recurrent neural network is applied to investing the vocoder in this work. Then the overall contribution of the paper is listed out as follows,

- Maximizing the speech synthesis accuracy by applying a useful Pulse model in log-domain vocoder
- Improving the quality of speech synthesis by extracting the significant features from the signal.
- To reduce the error rate while creating the speech synthesis system

The created whale optimized deep convolution recurrent neural network-based speech synthesis system is implemented using an experimental setup and different performance metrics such as accuracy and error rate to examine the system's efficiency.

The rest of the script is arranged as follows: section 2 analyzes the whale optimized deep convolution recurrent neural network-based speech synthesis system. Section 3 evaluates the proficiency of the speech synthesis process and concludes in section 4.

2. Process of Speech Synthesis

This section discusses the speech synthesis process; the main aim is to transmit the input text into a signal. During this process, the system takes text as the input, which is analyzed continuously; different acoustic and linguistic features are extracted that are changed into the waveform. Based on the discussion, the working process of speech synthesis is illustrated in figure 1.
Figure 1 illustrated that the working process of speech synthesis consists of training and synthesis prediction blocks. The database consists of synthesis information that trains the features based on the linguistic and acoustic model features. From the trained features, testing has been performed, and speech synthesis is identified correctly. During this process, the vocoder placed a crucial role because it helps express the extracted features' waveform. In addition to this, the vocoder maximizes the intelligibility and manipulates the speech that improves the overall speech synthesis. Here, the whale optimized deep convolution recurrent neural network (WODCRNN) worked with the Pulse model in log-domain vocoder is utilized. This vocoder gathers the related speech parameters from the sinusoidal analysis because it is a significant frequency-domain plus based synthesis process. Initially, the spectral envelope \( SE(z) \) and the F0
contour is computed. The \((SE(z))\) value is obtained from the harmonic peak amplitudes of interpolation.

\[
(SE(z)) = I(z)(SE(z))
\]  

In eqn. (1), \((SE(z))\) is defined as a spectral envelope, \(I(z)\) is defined as the flat excitation signal.

After that, the phase distortion deviation (PDD) value is estimated using harmonic phase distortion value that is computed using eqn. (2)

\[
PDD_i(\omega) = \sqrt{-2 \log \left| \frac{1}{N} \sum_{n} e^{iPDD_i(\omega)}} \right|
\]  

The eqn. (2) is used to determine the phase distortion in each frequency compared with the F0. According to the F0 counter value, the synthesis process is performed by quantizing the binary noise mask value \(M_i(\omega)\). The phase spectrum value is changed by random noise value when

\[
M_i(\omega) = 1
\]

\[
S_i(\omega) = e^{-j\omega t} \cdot V_i(\omega) \cdot N_i(\omega) M_i(\omega)
\]  

In eqn. (3), the phase pulse minimum value is denoted as \(S_i(\omega)\), its length is \(\frac{1}{F0}\).

\(t\) is pitch mark

The spectral envelope minimum phase response value is denoted as \(V_i(\omega)\). \(N_i(\omega)\) is denoted as the Gaussian noise Fourier transform

Based on the eqn. (1 to 3), the synthesis process is performed in the training phase. Hence, the WODCRNN approach based training process requires linguistic and acoustic features. The incoming text input parameters are continuously analyzed to be converted into the sequence of linguistic features. The derived features are processed by introducing the WODCRNN approach for extracting the acoustic features used to synthesize with the vocoder's help. The whale optimized deep convolution recurrent neural network (WODCRNN) process eliminates the difficulties in extracting textual features and minimizes the unwanted analysis in input
examination. More ever, this process improves the overall speech synthesis process. Then the acoustic modeling process is illustrated in figure 2.

**Figure 2: Structure of Optimized convolution deep recurrent network-based speech synthesis working process**

Figure 2 illustrated that the systematic view of the speech synthesis process. Here the MVF, MCG, and F0 features are extracted from the speech signal. The extracted information is fed into the acoustic layer that works with the vocoder for generating the synthesis process. The maximum voice frequency (MVF) and Fundamental frequency (F0) are extracted from the speech. Then the MCG value is derived from the audio signal because it is one type of cepstral illustration. The MCG value is computed by taking the Fourier to transform for a signal, and the overlapping triangular windows applied to map the spectrum powers. Then log value is calculated for mel frequencies and the discrete cosine change for getting the Mel cepstral coefficient (MCG) value. Then the overall MCG value is estimation is computed as follows.

\[ H(z) = S^{-1}_{\gamma} \left( \sum_{m=0}^{M} c_{z}^{\gamma(m)}z^{-m} \right) \]  

(4)
Based on the eqn. (4 and 5), the MCG value is computed from the signal transmitted and the linguistic feature to the vocoder for generating the speech synthesis. Figure 1 shows that the linguistic features and acoustic features are predicted by using the WODCRNN approach. Here recurrent neural networks and convolution neural networks combined to generate the prediction model. This optimized classification model processes the audio signal and provides better results than traditional classifiers such as recurrent and convolution networks. The network consists of bath normalization, exponential linear unit (ELU) activation function, max pooling, and dropout layer. The first three convolution layers are arranged sequentially with respective activation functions.

Furthermore, reshape and permute layers are applied in the network for feature vectors because recurrent network uses the 2d- feature vector and convolution network uses the 3d-feature vector. After applying the reshaping layer to the features, permute layers change the feature vector axes direction. This process was converting the extracted feature vector into a two-dimensional feature vector. At the time of the acoustic model generation process, the network uses the two bi-directional gated recurrent units with 'n' number of recurrent unit cells. Each layer in the gated recurrent unit depends on the number of classes while making the classification process. Due to the future timestamps, bi-directional gated recurrent units are utilized to process the features.

The timestamp process was incorporating the time dimensional features in an optimized manner. The bi-directional recurrent network's obtained output is fed into the fully connected layers for creating the useful acoustic prediction model. During this process, the neural network performance's efficiency is enhanced by using the whale optimization technique. The algorithm works according to whale behavior while optimizing network performance. The whale utilizes the prey encircle, bubble-net attacking phase, and prey searching phase for selecting their food. Initially, the prey encircle is computed as follows.

\[
D = \left| CX^* (t) - X (t) \right| 
\]  

(6)
In the above eqn. (6 and 7), the prey current position iteration is defined as \( t \).

Coefficient vectors are denoted as \( \hat{\lambda} \) and \( C \).

The current optimal solution position is denoted as \( \hat{X}^* \), \( \hat{X} \) is mentioned as a position vector.

Absolute value is defined as \(|.|\). The coefficient vectors are computing as follows

\[
\hat{\lambda} = 2 \cdot \hat{\lambda}^0 - \hat{\lambda} \tag{8}
\]

\[
C = 2 \cdot \hat{\lambda} \tag{9}
\]

For every iteration \( b \) value is chosen from 0 to 2. A random vector is defined as \( \hat{b} \) has values from 0 to 1. Based on the position values, the whale prey encircle is identified, and the optimized one is selected according to the bubble net attacking. The mathematical derivation of this process is described in eqn. (10)

\[
\hat{X}(t+1) = \begin{cases} 
(\hat{X}^*(t) - \hat{\lambda} \cdot \hat{B}) & \text{ if } p < 0.5 \\
\hat{B} e^{b \cdot \cos(2\pi l)} + \hat{X}^*(t) & \text{ if } p \geq 0.5 
\end{cases} \tag{10}
\]

In eqn. (10), the position updating process is defined as \( \hat{B} e^{b \cdot \cos(2\pi l)} + \hat{X}^*(t) \), the constant value is \( b \), the random number is represented as \( l \), having a value from -1 to 1. \( p \) is also the random value having a value from 0 to 1. After that, the best prey has been searched according to eqn. (11 and 12)

\[
\hat{B} = |C \cdot \hat{X}_{rand} - \hat{X}| \tag{11}
\]

\[
\hat{X}(t+1)\hat{X}_{rand} - \hat{\lambda} \cdot \hat{B} \tag{12}
\]

The random position vector is defined as \( \hat{X}_{rand} \). Based on the above process, the optimized weight value is selected for the network, which reduces the error value while predicting
linguistic and acoustic features. The generated features are fed into the vocoder for creating the speech synthesis that is achieved with the help of a trained speech synthesizer. Then the efficiency of the system is evaluated using experimental results and discussions.

3. Results and Discussions

The excellence of whale optimized deep convolution recurrent neural network (WODCRNN) based speech synthesis process is evaluated in this section. The system uses the Pulse model in the log-domain vocoder to generate the speech synthesis from MCG, MVF, and F0 counter features. The derived features are more helpful in getting the sequence speech from the input speech. The discussed system is implemented using the implementation tool, and the efficiency is examined using the CMU-ARCTIC database [24]. This database is created by Carnegie Mellon University language technologies institute. The developed database is more helpful in making the speech synthesis research process. The dataset has 1150 utterances selected from project Gutenberg, including female (slt) and male (bdl) speakers speech information. Around 1132 sentences are used for the speech synthesis process, in which 80% of data is utilized for training and 20% for testing purposes. The discussed optimized deep convolution recurrent network is developed using Merlin speech toolkit 3 that is open source.

The same neural network structure is utilized for both linguistic and acoustic predictive model generation. The network performance improved by optimizing the network weight and bias value using a whale optimization algorithm. This optimization algorithm reduces the error rate, and stochastic gradient function is used further to minimize the mean square error (MSE) value while doing speech synthesis in the training process. Then the spectral enhancement is performed using the speech signal-processing toolkit, and delta features are extracted. The extracted features are normalized by applying min-max (linguistic features) and mean-variance normalization process (acoustic features). Based on the general guidelines, the speech features are trained continuously in creating speech synthesis. The developed system proficiency is evaluated using the following evaluations.

Evaluations

The discussed whale optimized deep convolution recurrent neural network (WODCRNN) based speech synthesis process effectiveness is evaluated in this section. During this process, efficiency
is determined in terms of a subjective and objective process. First, the vocoder performance is assessed in this section because it converts the extracted features into the speech synthesis. The created WODCRNN approach efficiency is estimated because it only determines both linguistic and acoustic models' predictive models.

**Evaluation in terms of Objective**

This section determines how effectively whale optimized deep convolution recurrent neural network (WODCRNN) obtains the significant results while working with the vocoder. The efficiency is determined in the training set using the mean square error rate. The obtained results are compared with the traditional techniques such as deep learning networks (DNN), recurrent networks (RNN), long-short term memory network (LSTM), bi-directional long-short term memory network (BLSTM), and gated recurrent units (GRU). The obtained error rate is illustrated in table 1.

| Methods  | Mean Square Error (MSE) | Root mean square error (RMSE) | Voice/ unvoiced error (VUV) |
|----------|------------------------|-------------------------------|-----------------------------|
| DNN      | 1.47                   | 1.28                          | 1.29                        |
| RNN      | 1.83                   | 1.358                         | 1.32                        |
| LSTM     | 1.37                   | 1.38                          | 1.28                        |
| BLSTM    | 1.28                   | 1.23                          | 1.13                        |
| GRU      | 1.37                   | 1.39                          | 1.25                        |
| WODCRNN  | **1.02**               | **1.034**                     | **1.054**                   |

Table 1 demonstrated that the whale optimized deep convolution recurrent neural network (WODCRNN) algorithm ensures minimum training error rate. The effective incorporation of the WODCRNN method with the Pulse model in the log-domain vocoder creates speech synthesis with a minimum error rate. The obtained error values are lower compared to the other traditional techniques like deep learning networks (DNN), recurrent networks (RNN), long-short term memory network (LSTM), bi-directional long-short term memory network (BLSTM), and gated...
recurrent units (GRU). Based on the discussion, the respective graphical analysis is illustrated in figure 3. From the results, the whale optimized deep convolution recurrent neural network (WODCRNN) with Pulse model in the log-domain (PML) vocoder approach predicts the speech synthesis with minimum deviations. In addition to this, it clearly states that the method utilizes the MCG, MVF, and F0 counter features effectively. Even though the baseline methods achieve the best results, the WODCRNN approach has provided better results.

Figure 3: Training Validation Error

In addition to the training validation error, objective evaluation of different vocoder techniques is examined, and the respective results are illustrated in table 2.
Table 2: Objective evaluation of vocoders

| Vocoders       | Mel Cepstral deviations (MCD) | Root mean square error (RMSE) | Correlation errors (CE) | Mean Square Error (MSE) | Voice/unvoiced error (VUV) |
|----------------|-------------------------------|------------------------------|-------------------------|-------------------------|-----------------------------|
| STRAIGHT       | 4.302                         | 1.234                        | 1.23                    | 1.9                     | 1.26                        |
| GlotDNN        | 4.134                         | 1.29                         | 1.13                    | 1.24                    | 1.23                        |
| GlottHMM       | 4.298                         | 1.32                         | 1.24                    | 1.31                    | 1.1                         |
| WODCRNN-PML    | 4.023                         | 1.13                         | 1.024                   | 1.04                    | 1.03                        |

Table 2 demonstrated that the whale optimized deep convolution recurrent neural network (WODCRNN) with Pulse model in the log-domain (PML) vocoder algorithm achieves significant results while training the input speech. The vocoder generates the speech

\[ PDD(\omega) = \sqrt{-2 \log \left( \frac{1}{N} \sum_{n} e^{i(pD_{n}(\omega))} \right)} \]

effectively by updating the spectrum phase with minimum pulse value

\[ S(\omega) = e^{-j\omega t} \cdot V(\omega) \cdot N(\omega) \cdot M(\omega). \]

This process creates the speech synthesis successfully, but the error rate or correlation error values are reduced by successful updating of the whale optimal prey solution selection process

\[ \dot{X}(t+1) = \begin{cases} \frac{1}{2} \dot{X}(t) - \frac{1}{2} \cdot \frac{\rho}{\dot{P}} & \text{if } p < 0.5 \\ \frac{\rho}{\dot{P}} \cdot e^{j\nu} \cdot \cos(2\pi t) + \dot{X}(t) & \text{if } p \geq 0.5 \end{cases} \]

The obtained error values are lower compared to the other traditional techniques like deep learning networks (DNN), recurrent networks (RNN), long-short term memory network (LSTM), bi-directional long-short term memory network (BLSTM), and gated recurrent units (GRU). The discussion clearly states that the training model for linguistic and acoustic models is created with the minimum error value. The minimum error value leads to maximize the overall efficiency of the speech synthesis. Therefore, the obtained accuracy values are illustrated in table 3.
Table 3: Efficiency of the training set with vocoders based speech synthesis

| Methods   | Precision | Recall | Mathew correlation coefficient (MCC) | F-measure |
|-----------|-----------|--------|--------------------------------------|-----------|
| DNN       | 93.89     | 93.28  | 94.77                                | 93.7      |
| RNN       | 94.28     | 94.19  | 94.82                                | 94.0      |
| LSTM      | 95.19     | 95.28  | 95.60                                | 95.28     |
| BLSTM     | 96.1      | 96.48  | 95.98                                | 96.29     |
| GRU       | 96.89     | 97.20  | 96.29                                | 97.28     |
| WODCRNN   | 98.98     | 99.02  | 99.13                                | 99.34     |

Table 3 demonstrated that the whale optimized deep convolution recurrent neural network (WODCRNN) algorithm attains maximum speech synthesis accuracy. The effective incorporation of the WODCRNN method with the Pulse model in the log-domain vocoder creates speech synthesis with minimum deviation and maximum synthesis accuracy. The obtained accuracy values are maximum compared to the other traditional techniques like deep learning networks (DNN), recurrent networks (RNN), long-short term memory network (LSTM), bi-directional long short term memory network (BLSTM), and gated recurrent units (GRU). Based on the discussion, the respective graphical analysis is illustrated in figure 4.
The above results are clearly illustrated that the whale optimized deep convolution recurrent neural network (WODCRNN) algorithm attains the significant results on training speech data. The predictive training model helps to analyze the testing speech from the user. Then the obtained testing evaluation is determined using subjective analysis.

**Evaluation in terms of subjective**

This section discusses the excellence of the whale optimized deep convolution recurrent neural network (WODCRNN) algorithm with a subjective manner. As discussed earlier, the speech data is collected from the CMU-ARCTIC database, both male and female participants. Therefore, the testing efficiency is determined using different participants in respective baseline methods.
Table 4: Testing Efficiency on subjective evaluation

| Participants | Precision | Recall | Mathew correlation coefficient (MCC) | F-measure |
|--------------|-----------|--------|-------------------------------------|-----------|
| 10           | 99.34     | 99.18  | 99.28                               | 99.21     |
| 30           | 99.18     | 99.38  | 99.23                               | 99.25     |
| 50           | 98.90     | 99.45  | 99.28                               | 99.36     |
| 70           | 99.23     | 99.32  | 99.38                               | 99.57     |
| 90           | 99.41     | 99.38  | 99.32                               | 99.21     |
| 110          | 99.29     | 99.39  | 99.41                               | 99.25     |
| 130          | 99.19     | 99.27  | 99.36                               | 99.31     |
| 150          | 99.28     | 99.37  | 99.27                               | 99.41     |
| 170          | 99.38     | 99.47  | 99.28                               | 99.45     |
| 190          | 99.39     | 99.28  | 99.18                               | 99.27     |
| 210          | 99.28     | 99.38  | 99.28                               | 99.3      |
| 230          | 99.37     | 99.28  | 99.29                               | 99.42     |

Table 4 clearly states that the introduced whale optimized deep convolution recurrent neural network (WODCRNN) attains the maximum testing accuracy on 20% of testing data in the CMU-ARCTIC database. The obtained accuracy is higher compared to other methods because of the effective optimization of network performance using

\[
\hat{X}(t+1) = \begin{cases} 
X^*(t) - \frac{\Delta t}{\mathcal{D}} \mathcal{B} & \text{if } p < 0.5 \\
\mathcal{D} e^{bi} \cdot \cos(2\pi t) + X^*(t) & \text{if } p \geq 0.5 
\end{cases}
\]

Whale prey searching process. This optimization process minimizes the unwanted search of features and creates the predictive model by utilizing the speech features like F0, MCG, and MVF features. Therefore, the vocoder creates speech synthesis without creating any complexity. Thus the WODCRNN system recognizes the speech synthesis with 99.27% precision, 99.34% recall, 99.29% of MCC, and 99.33% F-measure value.
4. Conclusion

Thus, the paper analyzing the whale optimized deep convolution recurrent neural network (WODCRNN) with Pulse model in the log-domain vocoder speech synthesis process. In this process, speech data is analyzed continuously using a predictive model for creating the training process. The training model generates the linguistic and acoustic model feature frame. The trained features are stored in the database. The CMU-ARCTIC database-related speaker speech information is collected and analyzed by a created optimized model that predicts the speech synthesis from the F0 counter, MCG, and MVF feature generations. The method attains the maximum synthesis recognition accuracy during this process because of the vocoder training and weight and bias updating process. Finally, the system's efficiency is determined using implementation results in which the system creates speech synthesis with 99.3%. In the future, the speech synthesis process is enhanced by applying a meta-heuristic based speech feature selection process.

Funding: No funding
Conflicts of interest/Competing interests: No conflict and Competing Interest
Availability of data and material: Available in paper
Code availability: Available in paper
Ethics approval: Not applicable
Consent to participate: Not applicable
Consent for publication: Not applicable

Reference

1. Vainio M. (2014) Phonetics and Machine Learning: Hierarchical Modelling of Prosody in Statistical Speech Synthesis. In: Besacier L., Dediu AH., Martín-Vide C. (eds) Statistical Language and Speech Processing. SLSP 2014. Lecture Notes in Computer Science, vol 8791. Springer, Cham. https://doi.org/10.1007/978-3-319-11397-5_3
2. Vainio, M., Suni, A., Aalto, D.: Emphasis, word prominence, and continuous wavelet transform in the control of HMM based synthesis. In: Speech Prosody in Speech Synthesis - Modeling, Realizing, Converting Prosody for High Quality and Flexible Speech Synthesis, Prosody, Phonology and Phonetics. Springer (2015)
3. Bird J.J., Ekárt A., Faria D.R. (2020) Phoneme Aware Speech Synthesis via Fine Tune Transfer Learning with a Tacotron Spectrogram Prediction Network. In: Ju Z., Yang L., Yang C., Gegov A., Zhou D. (eds) Advances in Computational Intelligence Systems.
4. Manogaran, G., Shakeel, P. M., Fouad, H., Nam, Y., Baskar, S., Chilamkurti, N., & Sundarasekar, R. (2019). Wearable IoT smart-log patch: An edge computing-based Bayesian deep learning network system for multi access physical monitoring system. Sensors, 19(13), 3030. https://doi.org/10.3390/s19133030

5. Y. Huang, L. He, W. Wei, W. Gale, J. Li and Y. Gong, "Using Personalized Speech Synthesis and Neural Language Generator for Rapid Speaker Adaptation," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7399-7403, doi: 10.1109/ICASSP40776.2020.9053104.

6. "Where "HAL" First Spoke (Bell Labs Speech Synthesis website)". Bell Labs. Archived from the original on 2000-04-07. Retrieved 2010-02-17.

7. Baskar, S., Shakeel, P. M., Kumar, R., Burhanuddin, M. A., & Sampath, R. (2020). A dynamic and interoperable communication framework for controlling the operations of wearable sensors in smart healthcare applications. Computer Communications, 149, 17-26. https://doi.org/10.1016/j.comcom.2019.10.004

8. Mac DK., Tran DD. (2015) Modeling Vietnamese Speech Prosody: A Step-by-Step Approach Towards an Expressive Speech Synthesis System. In: Li XL., Cao T., Lim EP., Zhou ZH., Ho TB., Cheung D. (eds) Trends and Applications in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, vol 9441. Springer, Cham. https://doi.org/10.1007/978-3-319-25660-3_23

9. Shiga Y., Ni J., Tachibana K., Okamoto T. (2020) Text-to-Speech Synthesis. In: Kidawara Y., Sumita E., Kawai H. (eds) Speech-to-Speech Translation. SpringerBriefs in Computer Science. Springer, Singapore. https://doi.org/10.1007/978-981-15-0595-9_3

10. Arik, S.O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., Shoeybi, M.: Deep voice: real-time neural text-to-speech. In Proceedings of ICML, pp. 195–204 (2017)

11. Hinterleitner F., Norrenbrock C., Möller S., Heute U. (2014) Text-To-Speech Synthesis. In: Möller S., Raake A. (eds) Quality of Experience. T-Labs Series in Telecommunication Services. Springer, Cham. https://doi.org/10.1007/978-3-319-02681-7_13
12. Panda, S.P., Nayak, A.K. An efficient model for text-to-speech synthesis in Indian languages. Int J Speech Technol 18, 305–315 (2015). https://doi.org/10.1007/s10772-015-9271-y

13. Manikandan, S & Chinnadurai, M 2019, ‘Intelligent and Deep Learning Approach OT Measure E-Learning Content in Online Distance Education’, The Online Journal of Distance Education and e-Learning, vol.7, issue 3, July 2019, ISSN: 2147-6454

14. Takaki S., Yamagishi J. (2016) Constructing a Deep Neural Network Based Spectral Model for Statistical Speech Synthesis. In: Esposito A. et al. (eds) Recent Advances in Nonlinear Speech Processing. Smart Innovation, Systems and Technologies, vol 48. Springer, Cham. https://doi.org/10.1007/978-3-319-28109-4_12

15. Fernandez, R., Rendel, A., Ramabhadran, B., Hoory, R.: Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In: Proceedings of Interspeech, pp. 2268–2272 (2014)

16. Manikandan, S, Chinnadurai, M, "Effective Energy Adaptive and Consumption in Wireless Sensor Network Using Distributed Source Coding and Sampling Techniques"., Wireless Personal Communication (2021), Springer, https://doi.org/10.1007/s11277-021-08081-3

17. Kolokas N., Drosou A., Tzovaras D. (2018) Keywords-To-Text Synthesis Using Recurrent Neural Network. In: Iliadis L., Maglogiannis I., Plagianakos V. (eds) Artificial Intelligence Applications and Innovations. AIAI 2018. IFIP Advances in Information and Communication Technology, vol 519. Springer, Cham. https://doi.org/10.1007/978-3-319-92007-8_8

18. Al-Radhi M.S., Csapó T.G., Németh G. (2017) Deep Recurrent Neural Networks in Speech Synthesis Using a Continuous Vocoder. In: Karpov A., Potapova R., Mporas I. (eds) Speech and Computer. SPECOM 2017. Lecture Notes in Computer Science, vol 10458. Springer, Cham. https://doi.org/10.1007/978-3-319-66429-3_27

19. Bollepalli, Bajibabu, Lauri Juvela, Manu Airaksinen, Cassia Valentini-Botinhao, and PaavoAlku. "Normal-to-Lombard adaptation of speech synthesis using long short-term memory recurrent neural networks." Speech Communication 110 (2019): 64-75.

20. Manikandan S, Chinnadurai M, Thiruvenkatasuresh M.P, Sivakumar M. (2020). "Prediction of Human Motion Detection in Video Surveillance Environment Using
Tensor Flow", International Journal of Advanced Science and Technology, 29(05), 2791 - 2798. Vol. 29 No. 05 (2020); Vol. 29 No. 05 (2020)

21. Y. Joo, W. Jun and H. Kang, "Efficient deep neural networks for speech synthesis using bottleneck features," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820721.

22. M. Hwang, E. Song, R. Yamamoto, F. Soong and H. Kang, "Improving LPCNET-Based Text-to-Speech with Linear Prediction-Structured Mixture Density Network," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7219-7223, doi: 10.1109/ICASSP40776.2020.9053704.

23. Y. Zhao, S. Takaki, H. Luong, J. Yamagishi, D. Saito and N. Minematsu, "Wasserstein GAN and Waveform Loss-Based Acoustic Model Training for Multi-Speaker Text-to-Speech Synthesis Systems Using a WaveNet Vocoder," in IEEE Access, vol. 6, pp. 60478-60488, 2018, doi: 10.1109/ACCESS.2018.2872060.

24. M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi and P. Alku, "A Comparison Between STRAIGHT, Glottal, and Sinusoidal Vocoding in Statistical Parametric Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 9, pp. 1658-1670, Sept. 2018, doi: 10.1109/TASLP.2018.2835720.

25. http://www.festvox.org/cmu_arctic/
Figures

Figure 1

Speech Synthesis working process
Figure 2

Structure of Optimized convolution deep recurrent network-based speech synthesis working process
Figure 3

Training Validation Error
Figure 4

Training Efficiency