IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 January 2023, accepted 5 February 2023, date of publication 13 February 2023, date of current version 17 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244393

RESEARCH ARTICLE

Pronunciation Scoring With Goodness of Pronunciation and Dynamic Time Warping

KAVITA SHEORAN[1], ARPIT BAJGOTI[1], (Student Member, IEEE), RISHIK GUPTA[1], NISHTHA JATANA[1], GEETIKA DHAND[1], CHARU GUPTA[2], PANKAJ DADHEECH[3], UMAR YAHYA[4], (Member, IEEE), AND NAGENDER ANEJA[5]

[1]Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, New Delhi 110058, India
[2]Department of Computer Science and Engineering, Bhagwan Parhsuram Institute of Technology, New Delhi 110089, India
[3]Department of Computer Science and Engineering, Swami Keshvanand Institute of Technology, Management and Gramothan, Jaipur, Rajasthan 302017, India
[4]Department of Computer Science and Information Technology, Faculty of Science, Islamic University in Uganda, Kampala, Uganda
[5]School of Digital Science, Univerisiti Brunei Darussalam, Gadong BE1410, Brunei Darussalam

Corresponding author: Umar Yahya (umar.yahya@iuiu.ac.ug)

ABSTRACT The current pronunciation scoring based on Goodness of Pronunciation (GOP) uses posterior probabilities of the Acoustic Models. Such algorithms suffer from generalization since they are utilized to determine a score metric for each phoneme rather than on the completeness or comparison with the ideal utterance of the words. This paper proposes a novel method to overcome such limitations by using combined scores of prosodic, fluency, completeness, and accuracy. This is achieved using context-aware GOP in conjugation with dynamic time warping (DTW) matching of the pitch contours of a weighted average of the context tokens found in the audio file that is rich in mispronounced phonemes. The proposed work gives flexibility in tuning the results according to different speech aspects based on a single hyperparameter. The results achieved are encouraging and have been validated on the speechocean762 dataset, where Automatic Speech Recognition (ASR) model has been trained on the Librispeech dataset. The resultant mean error of the proposed approach is 3.38% and the value of the correlation coefficient achieved is 0.652.

INDEX TERMS Pronunciation scoring, goodness of pronunciation, hidden Markov model-deep neural network, dynamic time warping, kaldi, speechocean762.

I. INTRODUCTION

English is a global language with complex grammatical literature. Since this language is used in so many documentaries and information exchanges, a lot of companies and institutions need their employees to be able to speak it. This language's acquisition and understanding have become a common issue, mainly focusing on the utterance of the words, which doesn't impair communication from the perspective of the speakers or the listeners in any way [1]. Teaching this language has many factors that make it difficult for teachers to efficiently teach the learners due to the following reasons:

- Teachers concentrate on improving the grammatical and vocabulary of learners so much that they become proficient in language reading and understanding skills rather than pronunciation [2].
- Learners are often discouraged as their teacher is much more proficient in the language and pronounces it differently [3].
- Continuous monitoring of the learner's pronunciation growth is complicated to achieve.

Computer Aided Pronunciation Teaching (CAPT) overcomes such disadvantages by developing a versatile model that concentrates more on the fluency and prosodic aspects of the speech, allows for easy progress monitoring, and provides automatic pronunciation scoring at any time. To help non-native English speakers, algorithms have been built for quick self-assessment of their performance in phoneme pronunciation, stress, and fluency without relying on a second person. A deep learning approach to generating a result that can be evaluated based on the posterior probabilities of acoustic

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

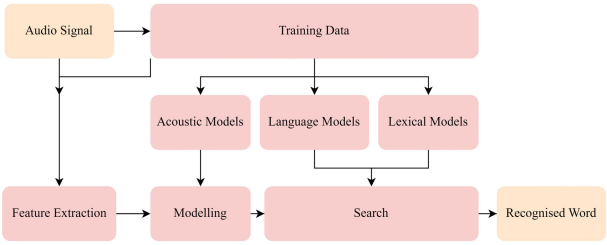


FIGURE 1. Block diagram of ASR System.

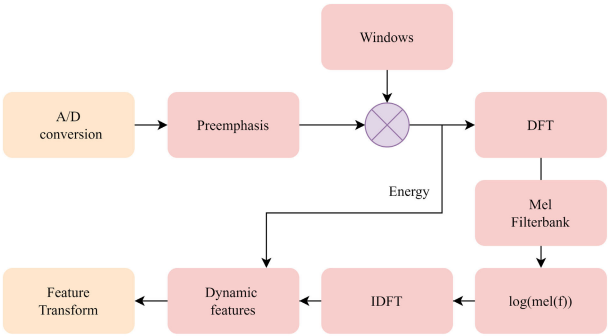


FIGURE 2. Steps involved in MFCC Feature Extraction.

models in ASR systems and the modified probabilities of the result describing the score for each phone of utterance will benefit the learner.

A. GOODNESS OF PRONUNCIATION

A measure of the effectiveness of speech recognition systems, specifically Automatic Speech Recognition (ASR) systems, is called Goodness of Pronunciation (GOP). By putting a number on how well an ASR system can detect and record speech, it can be used to gauge how well it performs. The GOP value is a numeric number between 0 and 1, where 1 denotes complete recognition and 0 denotes no recognition at all. This is also known as pronunciation accuracy, and it is often calculated as the ratio of the number of correctly detected phonemes to the total number of phonemes in the test speech signal. A detailed overview of GOP and its improved versions is described in related works [13].

B. AUTOMATIC SPEECH RECOGNITION SYSTEM

Speech recognition is when a computer or software can take phrases and words from spoken language and turn them into a format that a machine can understand. It is also known as computer voice recognition, speech-to-text, and automatic speech recognition.

Modules for automatic speech recognition consist of five stages.

- Audio Signal Detection
- Feature Extraction
- Acoustic Modelling
- Language and Lexical Modelling
- Training and Recognition

Fig. 1 shows the Automatic Speech Recognition System block diagram. The various blocks used in the ASR model are discussed in detail in the subsections below.

1) FEATURE EXTRACTION

The most crucial step in voice recognition is feature extraction since it separates one speech from another.

For speech recognition, the Mel-frequency cepstral coefficient (MFCC) is the most obvious and often used feature for the extraction method. The logarithmic placement of the frequency bands here resembles the human system reaction more than any other system. They are the result of a Mel-frequency spectral analysis in which the frequency bands of the Mel scale are evenly spaced. The MFCC computation method, which figures out the MFCC vector from each frame, is based on short-term analysis. Using the given formula in eq. 1, it is possible to calculate the MFCC.

$$Mel(f) = 1127 \log(1 + \frac{f}{700}) \tag{1}$$

Fig. 2 lists the steps in the feature extraction process. Firstly, the audio signal is converted from analog to digital format. Preemphasis increases the magnitude of energy in the higher frequency which improves phone detection. Then the signal is broken into multiple pieces(windows) to extract individual phones. The signal is then converted from the time domain to the frequency domain by applying Discrete Fourier Transform. The received signal is mapped according to human perceived levels using the mapping formula given in eq. 1 and logarithm function is applied to the Mel frequency output. The Inverse discrete Fourier transform (IDFT) block performs an inverse transform of the output from the previous step. The MFCC model takes in the first 12 coefficients of the signal after applying the IDFT operations, also taking the energy of the signal as a feature. Along with these 13 features, the MFCC technique will consider the first-order derivative and second-order derivatives of the features thus constituting a total of 39 features [4].

I-Vectors or identity-vectors are a more simplistic speaker recognition model that were presented in [5], they do away with the dichotomy between speaker and channel variability subspaces and describe both in the "total variability space," a common limited low-dimensional space. Particularly in ASR models, they are used to get the uniform dimension of every datapoint of audio feature [6].

2) ACOUSTIC MODELING

The foundational component of the ASR system is acoustic modeling [7]. In acoustic modeling, the relationship between the acoustic data and phonetics is formed. The acoustic model is vital to system performance and oversees the computational burden [8]. The relationship between the fundamental speech units and the acoustic observations is established by training. The system must be trained by using one or more patterns that match speech sounds from the same class to make a pattern

that represents the features of the class. Several models may be used for acoustic modeling. The Hidden Markov Model (HMM) is a popular technique since it is effective for both training and recognition [9].

a: HIDDEN MARKOV ACOUSTIC MODELLING

Hidden Markov Acoustic Model is a finite state Markov model where several output distributions serve as signals for a hidden Markov model. The output distribution model parameters are spectral variability whereas the alteration parameter in the Markov chain models is temporal variability. The capacity to recognize speech depends on these two forms of variability. Compared to a template-based method, hidden Markov modeling is more versatile and has a strong mathematical foundation. HMM makes it simpler to incorporate knowledge sources into structured architecture than the knowledge base method. A drawback of HMM is that it does not offer much insight into the recognition process. Analyzing system failure is done to enhance the performance of the HMM system, but it is a challenging task. However, careful knowledge inclusion has greatly enhanced the HMM-based system [11].

b: HMM-DNN BASED ACOUSTIC MODELLING

A Hidden Markov Model (HMM) and a Deep Neural Network (DNN) are combined to form an HMM-DNN model. The DNN is used to model the intricate patterns in the data, whereas the HMM is used to model the temporal relationships in the data. Due to its ability to combine the benefits of the HMM with the DNN, this model is highly suited for situations involving sequential data with intricate patterns [10] where the HMM-DNN model uses the HMM as prior knowledge to direct the training of a DNN and enhance voice recognition performance [16].

3) LANGUAGE MODELS

The likelihood of a word sequence is calculated by the Language Model (LM) component. By adding linguistic knowledge from big text corpora, LMs help acoustic models be more accurate. Implicitly learned syntactic and semantic criteria are employed by LMs to re-score the acoustic model hypotheses. Using a pronunciation dictionary, a series of phonemes is translated into words to match the phonetic transcriptions produced by the acoustic modeling with the raw text utilized in language models [12].

4) FORCED ALIGNMENT

Automatic speech recognition (ASR) uses a forced alignment technique to line up the words in transcription with the associated audio. A huge dataset of speech and transcripts is used to build an acoustic model as the first step in forced alignment. The acoustic model gains the ability to forecast the possibility that certain sounds will appear at various locations during the spoken stream. Once the model has been trained, it may be used to match the words in a transcription with the audio by identifying the sound sequence that, given the words in the transcription is most likely to have produced the audio.

In this paper, inspired by the posterior calculation of GOP and objective comparison of signals with DTW, we propose a new pronunciation scoring equation that combines both features using a hyparameter. In our approach, context-aware log posterior probabilities extracted by any Acoustic model of an ASR system are averaged together for each utterance of a speaker. Thus, the posterior features only represent the prosodic features of a speech. The DTW then extracts the fluency, and word level accuracy of the speech separately, The scores of both the extracted scores are added by a weighted average to get a total score. Compared with other acoustic model-based scoring, our model gave better results without even biasing or fine-tuning the acoustic model on the pronunciation dataset.

II. RELATED WORK

An algorithm developed by Witt et al. [13] used native acoustics modeling which relied on the Gaussian mixture model-hidden Markov model (GMM-HMM) to define the Goodness of Pronunciation and thus calculated a score from the formulated GOP. Improvements were made to this method by most of the works released afterward either by suggesting modifications to the GOP-based formulation or by enhancing the parameters of the original acoustic models. In the prior work of Zhang et al. [5], the score was calculated using scaled log posteriors instead of posteriors. Using the chosen state sequence acquired from the forward-backward technique, Luo et al. [14] formulated the GOP. The GOP concept put forward by Witt et al. [13] was employed by Wang et al. [15] with erroneous pattern detectors in the process of diagnosing phoneme mispronunciation, but these works only focused on the detection of the phoneme in a frame by forced alignment of phonemes and only took the prosodic score into consideration. Sudhakara et al. [16] introduced context-aware GOP which takes both senone and transition state probabilities into consideration. Ryu et al. [17] inferred that pronunciation scoring must combine phone level as well as articulatory-level diagnoses such as voicing, place of articulation, and manner of articulation on consonants. Lin et al. [18] used the acoustic model and replaced the forced alignment layer with a self-attention layer to get an utterance score based on transfer learning, but the results greatly depend on fine-tuning the scores of datasets. Cheng et al. [19] suggested the problems of GOP-based scoring and proposed an ASR-free scoring method based on I-vector, Normalization Flow (NF), and Discriminative Normalization Flow (DNF) for improvements in GOP and the better performance of the combination of both the approaches (GOP and ASR-free score). Bugdol et al. [20] stated that rather than checking the probability of finding the presence of phonemes in sequential order in an audio file, another approach is to use the Dynamic Time Warping (DTW) algorithm on the vectors of the amplitude or Mel-Frequency Cepstral Coefficients (MFCC) of the utterance in the frame between the ideal file to compare with and

TABLE 1. Comparing the related works.

Reference	Prosodic	Fluency	Completeness	Accuracy	Public Dataset
Witt et al. [13]	✓		✓		✓
Wang et al. [15]	✓		✓		✓
Sudhakara et al. [16]	✓		✓		✓
Ryu et al. [17]	✓	✓		✓	✓
Lin et al. [18]	✓	✓		✓	✓
Cheng et al. [19]		✓	✓	✓	✓
Bugdol et al. [20]	✓	✓		✓	✓
Permanasari et al. [22]		✓		✓	
Miodonska et al. [23]			✓	✓	
Karhila et al. [24]			✓	✓	✓
Chao et al. [35]	✓	✓	✓	✓	✓
Kim et al. [36]	✓	✓			✓

the file to test. The problem arises when the voice of these two speakers differs a lot in pitch and frequency, also the algorithm is quite slow for large audio files. An optimized version of this algorithm was proposed by Salvador et al. [21] called FastDTW which constraints the matching paths calculated by constraining the calculations on a band rather than the whole mesh grid. Permanasari et al. [22] suggested a method to reduce the Euclidian distance in DTW between audio files using the volume and different pronunciation times. Miodonska et al. [23] proposed a combined approach that uses phone-level mispronunciation detection using DTW, but the comparison approach always demands an ideal file to get a metric for distinction. Karhila et al. [24] used the Connectionist temporal classification (CTC) layer model to predict the phone sequence in the audio file and compared it to the ideal base file using Levenstein edit distance to calculate the difference between the sequence and predict the pronunciation score.

The former approaches discussed are based only on the pronunciation scoring based on the likelihood of individual phones in sequential order, hence is limited to phone-level features extraction of the audio file to test [13], [14], [15], [16], [17], [18], [19]. The latter checks for a minimum distance for comparing different length time series or MFCC or LPC for audio comparison [20], [21], [22]. Both approaches compute independent tasks, but the combined effect can be informative for the scoring of the overall quality of speech of a person's audio [23], [24]. A high-level score representing the percentage score is preferred by many speakers i.e., the overall grade they scored in terms of probability or percentage, therefore we integrated both these methods to get a score representing the presence of phones, accent, and prosodic scores and calculating the output as only a probability score. An overview of the methods discussed above is tabulated in table 1.

A. BASIC GOP FORMULATION

The GOP score is defined as the posterior probability of a target phone p, given a sequence of acoustic features O, from among a set of phones Q. The GOP score is heavily based on Bayesian statistics and was developed by Witt et al. [13].

$$GOP(p) = P(p|O) = \frac{P(O|p)P(p)}{\sum_{q \epsilon Q} P(O|q)P(q)} \tag{2}$$

B. REVISED GOP FORMULATION

Eight years after the introduction of the first GOP formulation by Witt et al. [13], in 2008 Zhang et al. [5] came up with the revised approach for GOP formulation which incorporated logarithmic scaling, using the absolute score value and normalization done with a number of frames (NF) for a specific phone.

$$GOP_R(p) = \left| \frac{P(O|p)P(p)}{\max_{q \epsilon Q} P(O^p|p)} \right| \tag{3}$$

C. DEEP NEURAL NETWORK GOP

The GOP devised by Hu et al. [25] replaced the Gaussian mixture model (GMM) with Deep Neural Network (DNN). This was achieved by replacing the use of likelihood probability $P(p|O)$, which was done in the face of GMM. Using a DNN in place of a GMM resulted in the enhanced capability for discrimination in the phoneme models.

However, using the DNN model to get acoustic features for GOP formulation has significant shortcomings, by using Multilayered Perceptron (MLP), structural locality arises since feature distance is not captured because of the fully connected characteristic of the DNN. Long short-term memory (LSTM) networks work better at modeling long-term dependencies of speech signals but due to dependencies between the time-frames being processed in an RNN, parallelization cannot easily be exploited to the same extent as in feedforward networks [38], [40], which greatly increases the training as well as inference time which is not preferred for real-time CAPT. Time Delay Neural Networks(TDNN) have been shown to effectively learn the temporal dynamics of the signal even from short-term feature representations and work better in GOP computing [16].

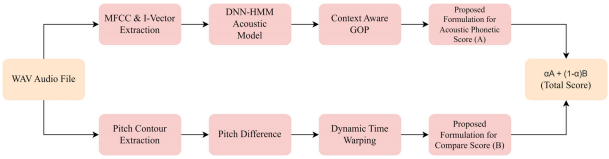


FIGURE 3. Pronunciation Scoring Framework.

III. PROPOSED APPROACH

This section explains the proposed workflow using context-aware GOP and DTW of pitch contours as well as the combined score for CAPT.

A. MODEL AND WORKFLOW

As shown in Fig. 3. The experimental setup is divided into two sections, section I contains the following components: the wav file, MFCC extraction, passing through a trained DNN-HMM Acoustic model, calculating the modified GOP scores, and calculating the posterior score from the proposed formula 7 and 8. Section II contains the same wav file as input, calculating the adjacent difference in pitch contour vectors of the same spoken words from the wav file as well as a comparison file from google text to speech API, then the vectors are passed through optimized DTW from which the compare scores are calculated, finally, the output is calculated as the weighted average of the 2 section scores. A detailed explanation of the two sections is discussed in section C.

Posterior Probability workflow is shown in Fig 4 [16], the DNN-HMM model posterior probabilities and senone probabilities are used to generate an equation for context-aware GOP, and the details are discussed in sections B and C.

B. POSTERIOR PROBABILITIES CALCULATION

The definition of GOP-NN is a bit different from the GOP-GMM. GOP-NN was defined as the log phone posterior ratio between the canonical phone and the one with the highest score [25]. Firstly, we define Log Phone Posterior (LPP):

$$LPP = \log P(p|O; t_s, t_e) \tag{4}$$

Then we define the GOP-NN using LPP:

$$GOP(p) = LPP(p) - \max_{q \epsilon Q} LPP(q) \tag{5}$$

LPP could be calculated as:

$$LPP(p) = \frac{1}{t_e + t_s + 1} \sum_{t=t_s}^{t_e} \log p(p|O_t) \tag{6}$$

$$p(p|O_t) = \sum_{s \epsilon p} p(s|O_t) \tag{7}$$

where s is the senone label, $s|s \in p$ are the states belonging to those triphones whose current phone is p. Normally the classifier-based approach achieves better performance than the GOP-based approach. Being different from other GOP-based methods, an extra supervised training process is needed. The input features for supervised training are phone-level and segmental features. The phone-level feature is defined as:

$$[LPP(p_1), \ldots, LPP(p_m), LPP(p_1|p_i), \ldots, LPP(p_j|p_i)]^T \tag{8}$$

where the Log Posterior Ratio (LPR) between phone p_j and p_i is defined as:

$$LPR(p_j|p_i) = \log \frac{p(p_j|O; t_s, t_e)}{p(p_i|O; t_s, t_e)} \tag{9}$$

This formula calculates the ratio of two phonemes and returns a value less than 0, where a high negative value indicates the audio segment of this phone should be a mispronunciation.

C. PROPOSED FORMULATION

The theory discussed in equations 4 to 9 set up the basis of the accumulated utterances' phone-level score, followed by the compare score and combined total score of the pronunciation.

1) PHONE-LEVEL FEATURE

The log phone posterior ratio accumulated for a specific test audio file is taken and let p be the vector representing the phone set in series and $LPR(p_j|p_{\max})$ defines the score that is the ratio of canonical phone and the one with the highest score in sequence with vector p as defined in eq. 9. Let $f(p_i) = LPR(p_i|p_{\max})$ then the overall phone score is formulated as:

$$S = \sum_{i=1}^{k} \frac{\sum_{j=1}^{m} f(p) 1\{p_j = p_i\}}{m} \frac{\sum_{j=1}^{m} f(p) 1\{p_j = p_i\}}{\max_{j \epsilon m}(f(p)) \sum_{j=1}^{m} 1\{p_j = p_i\}} \tag{10}$$

where k is the number of phonemes in the language, m is the size of the vector p and $1\{p_i{=}p_j\}$ is an indicator function that returns a value of 1 when the respective phones in the condition are matched. The value S defines the probability of bad pronunciation given the set of LPR scores for an audio file. This probability score is passed through a nonlinear function defined as:

$$p(O) = \frac{2e^{-3S}}{1 + e^{-3S}} \tag{11}$$

where s is the cumulative probability from eq. 10, eq. 11 and gives a probability score of goodness in the overall phone matching, this function doesn't allow the scores to be at the margins.

2) COMPARE SCORE

The audio file is passed through an ASR model to get the words spoken corresponding to the timestamps of each spoken word. Words that contribute to a significant amount regarding the context of the speech are randomly sampled, the sampled signal is represented as:

$$x = [x_1, x_2, \ldots, x_{A-1}, x_A] \tag{12}$$

where A is the sample count of the signal. The vector in eq. 12 is further processed as follows:

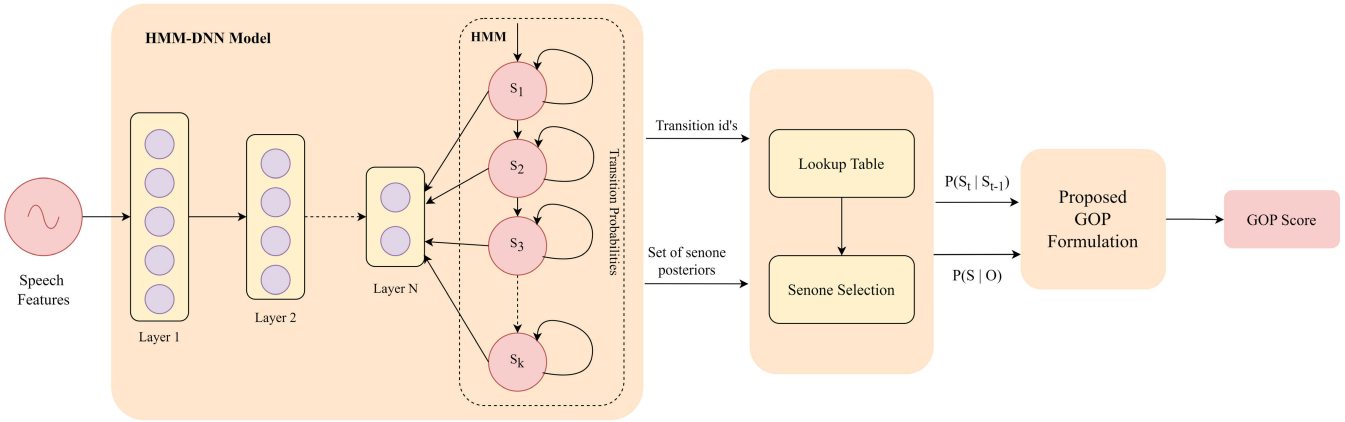


FIGURE 4. Posterior Score: Flow Diagram of GOP calculation using pre-trained weights of HMM-DNN acoustic model [16].

a: PITCH EXTRACTION

The modified Auto-Correlation function-based pitch detection is defined as:

$$r(\tau) = x \otimes x(\tau) = \sum_{i=q}^{q+N-1} x(i)x(i+\tau) \tag{13}$$

where q is the current frame, N is the frame length, τ is the lag index, and $x(i)$ is the current sample of the audio signal.

The main peak in the auto-correlation function is at the zero-lag location ($\tau = 0$). The location of the next peak gives an estimate of the period, and the height gives an indication of the periodicity of the signal, this estimate is given by:

$$\hat{f}_0 = \frac{1}{\tau_{max}}, r(\tau_{max}) = \max_{\tau} r(\tau) \tag{14}$$

where $\tau > 0$.

Applying equations 13 and 14 on the audio signal gives a frame-wise pitch array given by:

$$x = [\hat{f}_1, \hat{f}_2, \ldots, \hat{f_{n-1}}, \hat{f}_n] \tag{15}$$

where n is the number of frames in the ideal speech signal. The auto-Correlation method is used for pitch detection instead of other methods because of its better performance on audio speech signals elaborated in the discussion section. In order to compare the frames of the signals it is often more informative to analyze the overall transition of the pitch track than the absolute value. Therefore delta of the obtained frame-wise fundamental frequencies is taken as follows to obtain a new signal vector as follows:

$$\Delta_k = f_k - f_{k-1} \tag{16}$$

where f_k is any arbitrary frequency in the new array in eq. 14, the new array after applying eq. 15 gives:

$$\Delta x = [\Delta_0, \Delta_1, \ldots, \Delta_{n-1}, \Delta_n] \tag{17}$$

An ideally pronounced speech of the sampled audio Y is then generated using Google Text to Speech (GTTS) library and pitch is extracted similarly to the sampled audio in eq. 12 represented as:

$$\Delta y = [\Delta_0, \Delta_1, \ldots, \Delta_{m-1}, \Delta_m] \tag{18}$$

where m is the no. of frames in the ideal speech signal.

b: DTW MATCHING

Let W be the mesh grid of each point in Δx and Δy given by (i, j), then the FastDTW algorithm by Stan et al. [21] is used to define a warping path and corresponding DTW distance as:

$$D_{\min}(i_k, j_k) = \min_{i_{k-1},\, j_{k-1}} D_{\min}(i_{k-1}, j_{k-1}) + d(k, k-1) \tag{19}$$

where k is an arbitary point in the meshgrid W and d is the Euclidean distance given by:

$$d(i,\ j) = ||f(i) - f(j)||_2 \tag{20}$$

overall path cost:

$$D = \sum_{k}^{d} (i_k,\ j_k) \tag{21}$$

The path cost D is changed to a probability score using a threshold value for comparison i.e., 750, and passed through a smoothing function to change it in the scale of zero to one as:

$$p(d) = e^{\frac{t-d}{t}} \tag{22}$$

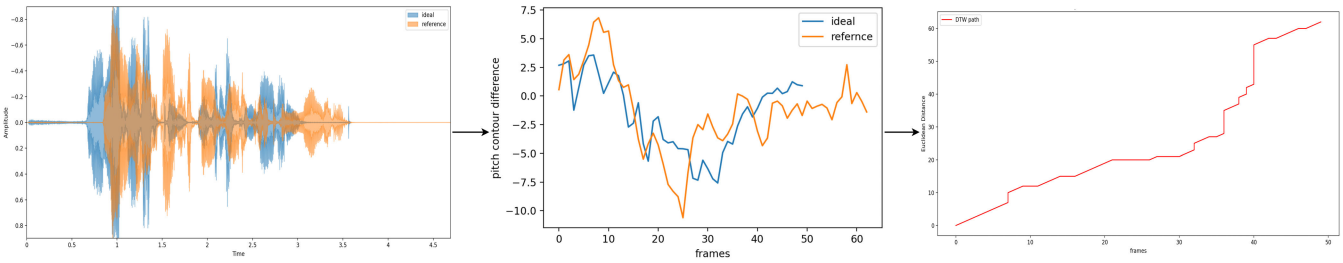
where $t = 750$ (threshold value) The flow diagram of the extraction of compare scores is shown in fig. 5.

3) TOTAL SCORE

Finally, the total score is calculated as a combination of phone feature score and audio pitch difference comparison score as a weighted average where the weights are hyperparameters that can be tuned by the examiner or learner according to the preference of one score over another as:

$$p(t) = \alpha p(o) + (1 - \alpha)\, p(d) \tag{23}$$

Here α is a hyperparameter and $0 < \alpha < 1$.



Waveforms of audio files | Adjacent pitch difference | L2 norm path of DTW

FIGURE 5. Compare Score: DTW of difference of pitch contours of ideal and test waveforms a) Waveform b) Pitch Contour Difference c) DTW with L2 norm metric.

IV. EXPERIMENTS

This section discusses the datasets used for the model and also discusses how the hyperparameter tuning was carried out.

A. DATASET

The speechocean762 speech corpus is designed for pronunciation assessment use, consisting of 5000 English utterances from 250 non-native speakers, where half of the speakers are children [26]. Five experts annotated each of the utterances at the sentence, word, and phoneme levels. The sentence-level scores by experts are used as the proposed equations calculate utterance level mispronunciation whereas high sentence-level scores rely on a consistently "good" word and phoneme pronunciation. The sentence-level scores of the dataset are further divided into Accuracy, Prosody, Fluency, and Completeness. The mean score of completeness and prosody is measured by eq. 11 and sentence-level Accuracy and Fluency are measured by eq. 15. For the Acoustic model, the Librispeech dataset was used which consists of approximately 1000 hours of English speech derived from audiobooks [32].

B. TRAINING SETUP

The Kaldi ASR toolkit contains recipes to aid ASR researchers, WSL2 Ubuntu Terminal Environment is used to set up the dependencies of Kaldi which mostly runs on bash script files. The Kaldi nnet3-chain recipe has some specific properties:

- Fixed transition probabilities are used in the HMM.
- 3 times lower frame rate is used at the output of the neural network, which dramatically reduces the amount of processing necessary during testing and makes real-time decoding considerably simpler.
- TDNN is used as neural nets as it is faster in short sequences and easier to tune.

The ASR system is about 3 times faster at decoding and training time is also reduced, the workflow is depicted in Fig. 5. The recipe is pre-trained on the Librispeech dataset [32] and is used to get the log posterior ratio (LPR) from the utterances of the speechocean762 dataset [26]. The GOP score is calculated using equations 10 and 11 for the extracted LPR from Kaldi.

The LPR score is used to differentiate the words that have incorrect phonemes, such words are extracted at random from the dataset utterances and compared with the same word uttered ideally by google text-to-speech (GTTS), both the files are then converted into their pitch contours and subtracted with their adjacent values, using Praat Parselmouth library in python with the following the pitch settings:

- Audio Sampling Rate: 22050 samples per second.
- Pitch Floor: 75 Hz - below this frequency will not be recruited.
- Pitch Ceiling: 600 Hz - above this frequency will be ignored.
- Silence threshold: 0.03 - frames that do not contain amplitudes above this threshold are probably silent.

The DTW scores with L2 norm are calculated and fed into eq. 15 to obtain compare score. Both results are combined using eq. 16 to get the total score.

V. RESULTS

This section details the evaluation process of the proposed approach on the LibriSpeech and the Speechocean762 dataset. The proposed approach is evaluated on different settings and is further compared with other pronunciation scoring methods.

Fig. 6 shows the different values of Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC) metrics, it can be noticed that as the value of hyperparameter α changes from 0 to 1 the error in the total score decreases and then increases linearly. Similarly, the PCC graph first increases then decreases drastically, this implies that the Posterior Score alone is not effective in conveying the overall Pronunciation score, conversely, the compare score alone cannot determine the same, thus a combination of both these aspects, when taken into consideration, provides a much better result.

Table 2 shows corresponding values of MAE and PCC for different values of hyperparameter α. The maximum value of the total score is obtained at $\alpha = 0.5$, further shifting its value affects the total score.

The graph of deviation for the first 90 speakers with 20 utterances each in comparison with the proposed approach with the general weighted average with hyperparameter as it

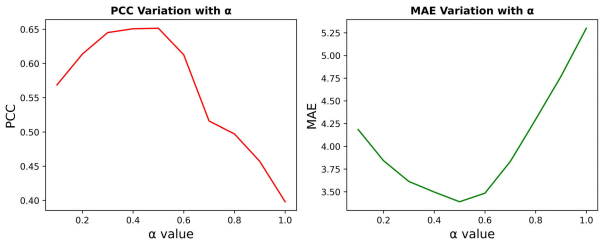


FIGURE 6. Plot of PCC vs α and MAE vs α in range 0 to 1.

TABLE 2. Mean absolute difference of the total score, posterior scores, and compare scores for different values of α.

Hyperparameter(α)	MAE of Total Score	PCC of Total Score
0.4	3.49	0.650
0.5	**3.38**	**0.652**
0.6	3.53	0.612

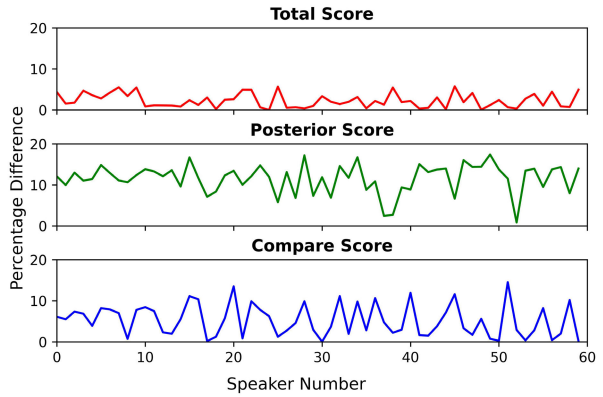


FIGURE 7. Graph of the absolute difference between the Total Score, Posterior Score and Compare Score for $\alpha = 0.5$.

approaches 0.5 is shown in fig. 7. The percentage difference in total score between the proposed method predictions and the dataset is plotted, it can be seen that although the total score gives less difference, the posterior score and compare score individually give a high difference error. To check the individual scores in comparison with the dataset score on the utterance level pronunciation, we sampled 10 different speakers' recordings with their respective total scores from 5 experts from the dataset, their average score, and their corresponding scores generated by our system taking the value of $\alpha = 0.5$ (for general results), as tabulated in table 3.

Table 3 compares the scores given by experts to the predicted scores. It can be clearly seen that the proposed approach matches the average score given by the 5 experts. Expert1 and expert2 scores are lenient whereas expert5 scores are strict but taking an unbiased estimate of pronunciation scores gives us an expected score for a speaker's pronunciation.

The mean difference error for the total score of the ideal expert scores and the proposed approach is shown in fig. 8. The horizontal axis shows the speaker number from the dataset and the vertical axis shows the total score of the speaker in percentage.

TABLE 3. All expert scores and their average score compared with the predicted score.

S no.	E1	E2	E3	E4	E5	Avg	Prediction
1	94.45	84.25	96.26	76.20	85.20	**87.27**	**86.13**
2	95.30	81.55	97.00	71.75	79.90	**85.10**	**78.63**
3	89.55	81.55	84.66	77.85	77.85	**82.92**	**81.24**
4	93.10	85.15	95.95	72.01	77.95	**84.85**	**81.62**
5	90.25	83.55	94.60	70.65	76.88	**83.18**	**84.88**
6	96.15	84.15	98.55	69.30	77.95	**85.16**	**86.90**
7	87.25	78.20	86.85	85.65	78.37	**83.26**	**82.29**
8	95.50	84.20	86.8	86.55	80.21	**86.65**	**81.49**
9	79.80	85.00	88.05	72.80	75.65	**80.26**	**80.29**
10	92.40	83.95	93.22	72.95	76.29	**83.76**	**83.47**

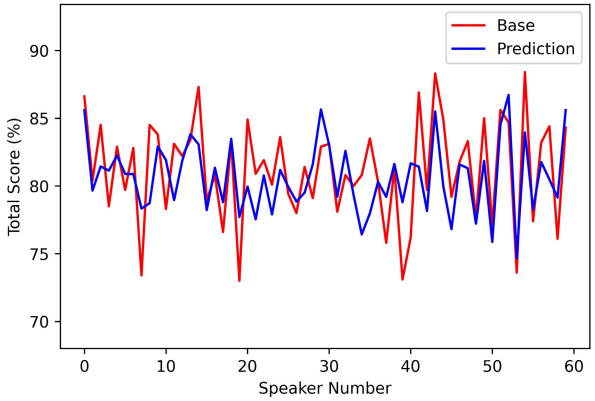


FIGURE 8. Graph of Total Score for expert average scores and model predictions.

TABLE 4. Pearson Correlation Coefficient scores compared with different methods.

Method	Dataset	PCC	Fine Tuned
DNN-HMM [16]	LS, FE Corpus	0.637	Yes
GOP + DNF [19]	WSJ, ERJ	0.667	Yes
PWLD-SVM [24]	WSJ, Spraakbanken	0.610	Yes
HMM-TDNN [28]	LS, TIMIT	0.392	Yes
HuBERT + GOP [35]	KESL	0.650	Yes
Wav2Vec2 Base [36]	SpeechOcean762	0.680	Yes
Proposed Approach	**LS, Speechocean762**	**0.652**	**No**

Table 4 compares the proposed approach with various other approaches using different datasets such as Speechocean762, TIMIT, LibriSpeech, and more. Latest models such as HuBERT [35] and Wav2Vec2 [36] were also compared. It should be noted that our model was not fine-tuned whereas all other models were fine-tuned before getting these results.

VI. DISCUSSION

The results show that there is a deviation of 3.38% from the actual mean total score of the professionals from the speechocean762 dataset. This result is relatively close to the actual labels as leniency is not taken into consideration, but changing the value of α in eq. 4 will fine-tune the predictions

to minimize the mean difference error between the two data vectors, which can be seen from the second row where biasing the total score to compare the score by 10% decreases the error by 0.75%.

A. AUTO-CORRELATION FOR PITCH EXTRACTION

Pitch Extraction is done with the help of the Auto-Correlation function, there are many improved methods for the extraction of the pitch but this method is adaptable to noise as the local maxima generated on the auto-correlation of the speech signal in the time domain is unaffected by the addition of slight noise. The main limitation of using this method is the presence of auto-correlation peaks that exceed the peaks corresponding to the pitch period. As a result, we get 'picking' of peaks, and consecutively incorrect pitch evaluation can occur, but as speech signals don't follow a pattern in a frame unlike musical signals, the limitation is not effective in this case [37].

B. TOTAL SCORE DEPENDENCY ON THE TYPE OF ACOUSTIC MODEL

A large variety of Acoustic models are used in related works, but the proposed method does not focus on a specific type of acoustic model as the model is not fine-tuned on the pronunciation dataset, any trained Acoustic model of an ASR system can be used to calculate the Phone-level features, and although it will decrease the percentage error of the Posterior Score in proposed eq. 11, the total score percentage error will not decrease as shown in Fig. 7. Here DNN-HMM model with TDNN layers is specifically used as its low inference time gives real-time scores in CAPT, although it has low phone score accuracy than other LSTM and transformer-based Acoustic models, it performs better on short utterances.

C. PROBLEM WITH FINE-TUNING MODEL ON DATASET

The graphs plotted in Fig 7 depict that without fine-tuning the Posterior Scores, which are purely dependent on the output of the Acoustic model show a percentage difference of up to 15%, this indicates that the models used in related works [16], [19], [24] produced results which were biased due to some extra feature extraction in the model to fit the scores of the specific dataset used, and are not generalized. Adding the compare score solves the high prosodic score dependency as the mispronounced words are compared with their ideal utterance to generate a different score which deals with the stress, Accuracy, and Fluency of the spoken word. The methods used in this approach use a single hyper-parameter α to tune the total score of the dataset which resolves the issue of biasing. The compare score percentage error can also be improved by changing the threshold value in eq. 22 from a constant scalar value to another hyper-parameter that is proportional to the number of phonemes in the word to maximize the compare score's contribution.

VII. CONCLUSION

Considering the basic GOP approach and distinct length vector comparison using DTW, we derive two formulas to calculate the pronunciation score of an audio file. The first calculates the accumulated posterior probabilities of the acoustic model, and the second randomly clips some of the words spoken in it to get vectors to compare with the ideal pronunciation of those specific words using DTW. These two scores are combined together using a weighted average. The results are evaluated with the total scores of the speachocean762 dataset, each scored by 5 professionals. The suggested method significantly resembles the average expert score from the dataset and offers results that can be tailored for both beginners and experienced learners. A limitation of the proposed approach is that the result of the prosodic score cannot be tuned according to the level of the learner and is entirely dependent on the accuracy of the acoustic model used. Future plans include providing more flexibility to the prosodic score and concentrating more on constructing a corpus for gender and region-based score calculation.

REFERENCES

[1] A. P. Gilakjani, "English pronunciation instruction: A literature review," *Int. J. Res. English Educ.*, vol. 1, no. 1, pp. 1–6, 2016.

[2] J. Harmer, *The Practice of English Language Teaching*. Harlow, U.K.: Pearson Longman, 2007.

[3] A. P. Gilakjani and R. Rahimy, "Using computer-assisted pronunciation teaching (CAPT) in English pronunciation instruction: A study on the impact and the teacher's role," *Educ. Inf. Technol.*, vol. 25, no. 2, pp. 1129–1159, Mar. 2020.

[4] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Mar. 2018, pp. 379–383.

[5] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for Mandarin," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 5077–5080.

[6] N. S. Ibrahim and D. A. Ramli, "I-vector extraction for speaker recognition based on dimensionality reduction," *Proc. Comput. Sci.*, vol. 126, pp. 1534–1540, Jan. 2018.

[7] R. K. Aggarwal, "Improving Hindi speech recognition using filter bank optimization and acoustic model refinement," Doctor Philosophy, Dept. Comput. Eng., Nat. Inst. Technol. Kurukshetra, Kurukshetra, India, 2012.

[8] A. Kumar, M. Dua, and T. Choudhary, "Continuous Hindi speech recognition using monophone based acoustic modeling," *Int. J. Comput. Appl.*, vol. 24, pp. 1–5, Jan. 2014.

[9] G. Gaurav, D. Deiv, G. Sharma, and M. Bhattacharya, "Development of application specific continuous speech recognition system in Hindi," *J. Signal Inf. Process.*, vol. 3 no. 3, pp. 394–401, 2012, doi: 10.4236/jsip.2012.33052.

[10] M. Afify, F. Liu, H. Jiang, and O. Siohan, "A new verification-based fast-match for large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 546–553, Jul. 2005.

[11] H. Wang, J. Xu, H. Ge, and Y. Wang, "Design and implementation of an English pronunciation scoring system for pupils based on DNN-HMM," in *Proc. 10th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Aug. 2019, pp. 348–352.

[12] Y.-C. Tam, Y. Lei, J. Zheng, and W. Wang, "ASR error detection using recurrent neural network language model and complementary ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 2312–2316.

[13] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, nos. 2–3, pp. 95–108, Feb. 2000.

[14] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation," in *Proc. Interspeech*, Sep. 2009, pp. 1–4.
[15] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 5049–5052.
[16] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities," in *Proc. Interspeech*, Sep. 2019, pp. 954–958.
[17] H. Ryu and M. Chung, "Mispronunciation diagnosis of L2 english at articulatory level using articulatory goodness-of-pronunciation features," in *Proc. 7th ISCA Workshop Speech Lang. Technol. Educ. (SLaTE)*, Aug. 2017, pp. 65–70.
[18] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment," in *Proc. Interspeech*, Aug. 2021, pp. 4438–4442.
[19] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "ASR-free pronunciation assessment," 2020, *arXiv:2005.11902*.
[20] M. Bugdol, Z. Segiet, and M. Krcichwost, "Pronunciation error detection using dynamic time warping algorithm," in *Information Technologies in Biomedicine*, vol. 4. Cham, Switzerland: Springer, 2014, pp. 345–354.
[21] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, pp. 561–580, Oct. 2007.
[22] Y. Permanasari, E. H. Harahap, and E. P. Ali, "Speech recognition using dynamic time warping (DTW)," *J. Phys., Conf. Ser.*, vol. 1366, no. 1, Nov. 2019, Art. no. 012091.
[23] Z. Miodonska, M. D. Bugdol, and M. Krecichwost, "Dynamic time warping in phoneme modeling for fast pronunciation error detection," *Comput. Biol. Med.*, vol. 69, pp. 277–285, Feb. 2016.
[24] R. Karhila, A.-R. Smolander, S. Ylinen, and M. Kurimo, "Transparent pronunciation scoring using articulatorily weighted phoneme edit distance," 2019, *arXiv:1905.02639*.
[25] W. Hu, Y. Qian, and F. K. Soong, "An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech," in *Proc. SLaTE*, 2015, pp. 71–76.
[26] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "Speechocean762: An open-source non-native English speech corpus for pronunciation assessment," 2021, *arXiv:2104.01378*.
[27] A.-L. Georgescu, H. Cucu, and C. Burileanu, "Kaldi-based DNN architectures for speech recognition in Romanian," in *Proc. Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Oct. 2019, pp. 1–6.
[28] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," 2020, *arXiv:2008.08647*.
[29] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7262–7266.
[30] M. Sancinetti, J. Vidal, C. Bonomi, and L. Ferrer, "A transfer learning based approach for pronunciation scoring," 2021, *arXiv:2111.00976*.
[31] S. Kim, M. L. Seltzer, J. Li, and R. Zhao, "Improved training for online end-to-end speech recognition systems," 2017, *arXiv:1711.02212*.
[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[33] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," 2018, *arXiv:1807.01738*.
[34] X. Han and T. Huwan, "The modular design of an english pronunciation level evaluation system based on machine learning," *Secur. Commun. Netw.*, vol. 2022, pp. 1–11, Jun. 2022.
[35] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 575–582.
[36] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," 2022, *arXiv:2204.03863*.
[37] L. Sukhostat and Y. Imamverdiyev, "A comparative analysis of pitch detection methods under the influence of different noise conditions," *J. Voice*, vol. 29, no. 4, pp. 410–417, Jul. 2015.
[38] F. Wu, L. P. García-Perera, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network," in *Proc. Interspeech*, Sep. 2019, pp. 1–5.
[39] W. Ying, "English pronunciation recognition and detection based on HMM-DNN," in *Proc. 11th Int. Conf. Measuring Technol. Mechatronics Autom. (ICMTMA)*, Apr. 2019, pp. 648–652.
[40] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, Sep. 2015, pp. 1–5.

KAVITA SHEORAN has been working at the Maharaja Surajmal Institute of Technology, since 2004, where she is currently an Associate Professor at the Department of Computer Science and Engineering. She has published several papers in SCI, Scopus, and Elsevier-indexed journals. Her current interests include software engineering and soft computing. She has chaired several sessions of international conferences and published several patents. She is a Life Member of the Indian Society for Technical Education (ISTE) and the Computer Society of India (CSI).

ARPIT BAJGOTI (Student Member, IEEE) was born in Delhi, India, in 2002. He is currently pursuing the B.Tech. degree in computer science and engineering with the Maharaja Surajmal Institute of Technology. His research interests include audio processing, ASR systems, and computer vision.

RISHIK GUPTA was born in Delhi, India, in 2002. He is currently pursuing the B.Tech. degree in computer science and engineering with the Maharaja Surajmal Institute of Technology. His research interests include speech detection and computer vision.

NISHTHA JATANA received the B.Tech. degree from the Computer Science and Engineering Department, the M.Tech. degree in computer technology and applications, and the Ph.D. degree in software testing from Guru Gobind Singh Indraprastha University, in 2021. She is currently an Assistant Professor at the Computer Science and Engineering Department, Maharaja Surajmal Institute of Technology. She has 12 years of teaching and research experience. She has 19 research publications, including eight journals, out of which three are SCIE-indexed journal articles, one book, and seven conference papers, published in various ESCI/Scopus publications. Her research interests include software engineering, meta-heuristic approaches, and network security.

GEETIKA DHAND has been associated with the Maharaja Surajmal Institute, since 2004, where she is currently an Associate Professor at the Department of Computer Science and Engineering. She has published several articles in SCI, Scopus, and Elsevier indexed journals. She has chaired several sessions of international conferences. She has published several patents. Her current research interests include wireless sensor networks, computer networks, and database management systems. She is a Life Member of the Indian Society for Technical Education (ISTE).

CHARU GUPTA graduated the B.E. degree in computer science and engineering. She received the M.Tech. degree (Hons.) in computer science and engineering from the JSS Academy of Technical Education, Noida, and the Doctoral degree from the Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. She is currently serving as an Associate Professor at the Bhagwan Parshuram Institute of Technology (Affiliated to GGSIPU, Dwarka), Delhi, with a teaching experience of more than ten years. She is a Research Associate with Nokia in the ''Invent with Nokia'' venture. She has to her credit various sessions at national and international conferences. She is a Faculty Coordinator (Delhi Section) of Free and Open Source Cell (FOSS cell) from the International Centre for Free and Open Source Software (ICFOSS), Goverment of Kerala, India. She is also the Faculty Coordinator of the e-Yantra Laboratory setup initiative (eLSI) in collaboration with IIT Bombay. She has also served as a Coordinator for Software Engineering for GGS Indraprastha University (IPU) B.Tech. syllabus designing committee. She has a research publications in various national and international journals/conferences of repute (SCIE/ESCI/SCOPUS). She is also a member of the *Encyclopedia of Neutrosophic Researchers* (Vol. 3, Book edited by Prof. Florentin Smarandache, University of New Mexico). She has been the co-convenor of various webinar series and a Faculty Coordinator of the National and International Conferences at BPIT. Her research interests include natural language processing, neutrosophic logic and its applications, time series analysis and forecasting, and evolutionary computation. She is a reviewer, a section editor, and an editorial member of many international journals/conferences. She is the lead editor and the co-editor in various book series by Wiley, De Gruyter, Nova publishing, and many more.

PANKAJ DADHEECH is currently working as an Associate Professor and the Deputy Head at the Department of Computer Science and Engineering (NBA Accredited), Swami Keshvanand Institute of Technology, Management and Gramothan (SKIT), Jaipur, Rajasthan, India (Accredited by NAAC A++ Grade). He has more than 17 years of experience in teaching. He has been appointed as a Ph.D. Research Supervisor at the Department of Computer Science and Engineering, SKIT (Recognized Research Centre of Rajasthan Technical University, Kota). He has also guided various M.Tech. Research Scholars. He has published 22 patents and granted five patents at Intellectual Property India, Office of the Controller General of Patents, Design and Trade Marks, Department of Industrial Policy and Promotion, Ministry of Commerce and Industry, Government of India. He has also published and granted five Australian patents at Commissioner of Patents, Intellectual Property Australia, Australian Government. He has also published and granted one German patent, one South African patent, and one USA patent. He has also registered and granted two research copyrights at Registrar of Copyrights, Copyright Office, Department for Promotion of Industry and Internal Trade, Ministry of Commerce and Industry, Government of India. He has presented 60 papers in various national and international conferences. He has 67 publications in various international and national journals. He has published seven books and 18 book chapters. His research interests include high performance computing, cloud computing, information security, big data analytics, intellectual property right, and the Internet of Things. He is a member of many professional organizations, such as the IEEE Computer Society, CSI, ACM, IAENG, and ISTE. He has chaired technical sessions in various international conferences and contributed as a resource person in various FDP's, workshops, STTP's, and conferences. He is also acting as a guest editor of the various reputed journal publishing houses and conference proceedings and the Bentham Ambassador of Bentham Science Publisher.

UMAR YAHYA (Member, IEEE) received the B.Sc. degree from the Islamic University of Technology (IUT), Bangladesh, in 2011, and the M.Sc. and Ph.D. degrees in computer science from Universiti Brunei Darussalam (UBD), Brunei, in 2014 and 2019, respectively. He is currently a Senior Lecturer of computer science at the Islamic University in Uganda (IUIU), Uganda. His research interests include intelligent systems, computational biomechanics, applied machine learning, the Internet of Things, and wearable computing.

NAGENDER ANEJA received the M.E. degree in computer technology and applications from the Delhi College of Engineering, and the Ph.D. degree in computer engineering from the J. C. Bose University of Science and Technology, YMCA. He is currently working as an Assistant Professor at the School of Digital Science, Universiti Brunei Darussalam. He is also working in deep learning, computer vision, and natural language processing. He is also the Founder of ResearchID Company.