Received: 24 July 2019 | Revised: 5 February 2020 | Accepted: 24 March 2020

DOI: 10.4218/etrij.2019-0364

ORIGINAL ARTICLE

ETRI Journal WILEY

Real-time implementation and performance evaluation of speech classifiers in speech analysis-synthesis

Sandeep Kumar

Department of Electronics and Communication Engineering, National Institute of Technology, Delhi, India

Correspondence
Sandeep Kumar, Department of Electronics and Communication Engineering, National Institute of Technology, Delhi, India.
Email: sandeep@nitdelhi.ac.in

In this work, six voiced/unvoiced speech classifiers based on the autocorrelation function (ACF), average magnitude difference function (AMDF), cepstrum, weighted ACF (WACF), zero crossing rate and energy of the signal (ZCR-E), and neural networks (NNs) have been simulated and implemented in real time using the TMS320C6713 DSP starter kit. These speech classifiers have been integrated into a linear-predictive-coding-based speech analysis-synthesis system and their performance has been compared in terms of the percentage of the voiced/unvoiced classification accuracy, speech quality, and computation time. The results of the percentage of the voiced/unvoiced classification accuracy and speech quality show that the NN-based speech classifier performs better than the ACF-, AMDF-, cepstrum-, WACF- and ZCR-E-based speech classifiers for both clean and noisy environments. The computation time results show that the AMDF-based speech classifier is computationally simple, and thus its computation time is less than that of other speech classifiers, while that of the NN-based speech classifier is greater compared with other classifiers.

KEYWORDS

ACF, AMDF, Cepstrum, neural network, real-time system, speech classification, WACF, ZCR-E

1 | INTRODUCTION

The human speech signal can be broadly classified into two categories: voiced speech and unvoiced speech. Voiced speech is produced as a result of the excitation generated by a periodic vibration of the vocal cords and can be seen as a quasi-periodic signal in the time domain representation. On the other hand, the unvoiced speech, which is non-periodic, consists of random signal-like excitations. A human speech signal also consists of a silence region, in which the signal energy is negligible, and no excitation is supplied to the vocal tract. Therefore, the silence region is assumed to be a subset of the unvoiced speech. Errors in speech analysis predominantly occur due to a voiced part getting wrongly classified as an unvoiced part or vice-versa. Hence, an accurate classification of a speech signal into the voiced and unvoiced speech frames plays an important role in the field of speech processing and its applications in mobile communication such as speech coding, speech analysis-synthesis, and speech recognition. It is used as a pre-processing step in many speech applications and is an important step for pitch detection in any speech analysis-synthesis system.

Accurately classified voiced and unvoiced speech segments can significantly improve the performance of any pitch detector [1,2] which, in turn, results in an improved quality of the synthesized speech signal in a communication system [3].

Several methods of speech classification such as the autocorrelation function (ACF), short-time energy of the signal (E), average magnitude difference function (AMDF), zero crossing rate (ZCR), cepstrum, discrete wavelet transform (DWT), and so on, that make use of acoustic features, have been reported in the literature [1,4–8]. A hybrid approach of speech classification such as the hidden Markov models, Gaussian mixture model or neural network (NN) model, that uses more than one feature, has been also reported [9-21].

In the study by Ahmadi and Spanisa [1], a multifeature voiced/unvoiced classification method based on cepstrum, ZCR and short-time energy was presented. It was found that the method is robust to noise. A simple and efficient voiced/unvoiced classification method based on the combination of ZCR and energy (ZCR-E) was presented by Bachu and others [5]. Their method was able to provide good results for speech classification. An improved pitch detection and voiced/unvoiced speech classification method based on the wavelet transform was presented by Janer and others [6]. A significant improvement in the pitch error and error rate of the voiced/unvoiced parts was observed with this method. In the investigation by Atal and others [9], a pattern recognition approach for speech classification was presented which provided satisfactory results. However, their algorithm required training on a specific dataset. Two novel hybrid methods of speech classification were presented by Shah and others [10]. The first method was based on the Mel-frequency cepstral coefficient (MFCC) with a Gaussian mixture model, while the second method was based on the linear predictive coding (LPC) coefficient with a reduced dimensional LPC residual and the Gaussian mixture model. Both the methods were able to give approximately 90% identification accuracy or more. Qi and Hunt [11] presented a speech classifier based on a multilayer feedforward network. Using this method, the rms energy and ZCR were extracted and an accuracy of 96% was achieved. Hassan and others [8] proposed a voiced/unvoiced classification algorithm of noisy speech by extracting the short-time energy and short-time zero crossing rate. The speech signal in the spectrogram image, in which the signal was divided into sub bands, was processed frame by frame and the energy ratio was calculated. A decision on the classification of the signals was taken based on their pattern using an energy ratio pattern matching lookup table. Drugman and others [12] considered voicing detection as a classification problem and pitch contour detection as a regression problem. For voicing detection, they extracted the acoustic features from three domains (time, frequency, and cepstrum). Using the *k*-means clustering algorithm and the multilayer perceptron class of the artificial neural networks, they reduced the voice deduction errors by 20% and 45%, respectively, compared with the other state-of-the-art techniques. Bagavathi and Padma [13] presented a fuzzy c-implies clustering method for classifying voiced and unvoiced activity using MFCC as features and achieved 91.5% classification accuracy. Bendiksen and Steiglitz [14] used an NN as a classifier for voiced/unvoiced speech classification. They extracted six features, namely, the rms energy of signal, the rms energy of the pre-emphasized signal, the normalized autocorrelation coefficient of the signal at unit sample delay, the normalized autocorrelation coefficient of the pre-emphasized signal at unit sample delay, the ratio of the signal energy above 4000 Hz to the signal energy below 2000 Hz, and the product of the signal energy above 4000 Hz to the signal energy below 2000 Hz. They achieved an error rate of 0.4%. Juang and Rabiner [15] discussed about the spectrum representation of speech from the computational (analytical) as well as perceptual viewpoints. This speech representation, in terms of the spectrum, is important if given as an input to the classifier to obtain high accuracy. An automatic speech segmentation using a neural tree network was presented by Sharma and Mammone [16] for the cases in which the number of sub-word acoustic units is either known or unknown a priori. This classifier gave an accuracy of 66.6%. A voiced/unvoiced speech classifier based on the adaptive filtering of the decomposed empirical modes was proposed by Khaldi and others [17]. They used features such as empirical mode decomposition and local statistics of speech and these features were filtered by adaptive center weighted average. It was observed that this proposed classifier gives superior results in terms of the average segmental signal-to-noise ratio (ASSNR) and perceptual evaluation of speech quality (PESQ), compared with the other methods considered. A noise robust voice activity detection system based on an unsupervised method was proposed by Ali and Talha [18], in which the long-term features were computed using the Katz algorithm of fractal dimension. The signal-to-noise ratio (SNR) was calculated at different levels in the presence of various noise sources such as white noise, cars, and babbling. It was observed that the method is reliable in labeling the voiced and unvoiced parts in both clean and noisy environments. In the report by Sun and others [19], a complexity analysis for the voiced/unvoiced speech classification based on the feature of the entropy of phonemes was described. On testing the different single phoneme signals, significant differences were observed in the sample entropy of the voice/unvoiced speech. Hence, it was concluded that the voiced/unvoiced decision can be made based on the measure of their complexity. Struwe [20] presented a voiced/unvoiced speech classification method using LPC as feature and a neural network as a classifier and reported that the proposed method works better in comparison to the other methods. In the study by Park and others [21], an algorithm for automatic speech segmentation in a concatenative text-to-speech synthesis was presented. They proposed the reliable segmentation boundaries of the speech data by applying a number of automatic segmentation machines simultaneously.

A significant improvement in the segmentation accuracy was thus observed.

Usually a threshold value for some acoustic features is used for voiced/unvoiced classification of a speech signal. In such methods, the classification performance generally depends on the choice of the acoustic features and an effective threshold. Owing to its low computational complexity, the time-domain acoustic features are generally used in real-time implementation. However, in the case of the statistical methods, training data having different levels of noise are required. Owing to its multiple features, the hybrid approach offers a significantly accurate voiced/unvoiced classification of speech signals at the cost of computational complexity.

Many speech classification methods have been reported in the literature and their reliability has been tested in terms of their classification accuracy. However, for practical/real-time applications of speech communication, the choice of an appropriate classifier is important for a reliable performance of the system, where, apart from the classification accuracy, some other parameters must be investigated for a proper selection. Therefore, some important parameters such as the speech quality and computational complexity of the complete communication system with different classifiers must be considered for the selection of the speech classifier.

Thus, considering the above-mentioned points, the key contribution of this work is the investigation on the performance of six voiced/unvoiced classification schemes on the basis of their comparison carried out in an LPC-based speech analysis-synthesis system [22,23]. Since a majority of the classification errors occur while classifying the voiced and unvoiced parts, this work focuses on the classification of the voiced/unvoiced parts in the speech only, leaving the silence part untouched. To achieve this goal, complete systems using six voiced/unvoiced classifiers have been simulated and implemented in real time using the TMS320C6713 DSP starter kit in the MATLAB environment. LPC is a simple and commonly used speech analysis-synthesis technique for producing good quality speech signals at low bit rates. It is the base of many speech coding techniques, including the code-excited linear prediction algorithm, that follows the ITU-T G.729 standard [24]. Therefore, an LPC-based speech analysis-synthesis system has been considered in this work in order to evaluate the performance of the speech classifiers. The classifiers chosen for the comparison, which are generally used in real-time applications, are based on: ACF, AMDF, weighted ACF (WACF), ZCR-E, cepstrum, and NN [4–7,10]. Although the performance of different speech classifiers has been compared earlier in terms of the percentage of their classification accuracy [4], their performance comparison after they are integrated into a complete communication system has been rarely found in the literature. In addition, a comparison of the different classifiers implemented in real time has not been reported thus far. Since speech quality and computational complexity are the two major parameters in any practical communication system, the performance of the speech classifiers investigated in this work has been compared in terms of the speech quality through the mean opinion score (MOS) and the PESQ test and the computational complexity through their simulation time and execution time (for real-time implementation), in addition to the percentage of their classification accuracy.

The remainder of the paper has been organized as follows: The details of the six voiced/unvoiced classifiers have been discussed in Section 2. The implementation of the analysis-synthesis system using the different classifiers has been presented in Section 3. Results of the performance comparison have been presented in Section 4 and the conclusions from this work are given in Section 5.

2 | VOICED/ UNVOICED SPEECH CLASSIFIERS

In this section, the steps involved in developing the six voiced/unvoiced classifiers (three single featured and three multi-featured speech classifier) have been presented [1,4–7,10]. The first three single featured speech classifiers are based on ACF, AMDF and cepstrum. The fourth and fifth classifiers, that use two features, are based on WACF (which is a combination of ACF and AMDF) and ZCR-E (which is a combination of the ZCR and short-time energy), respectively. The sixth classifier is based on an NN, which uses some acoustic features. To analyze the speech classifiers, two speech databases (PTDB-TUG and NOIZEUS) [25,26] have been used. A sampling frequency of 8 kHz was used for the analysis. The speech signals were framed in 20 ms chunks (that is, 160 samples) for the analysis using the different classifiers. A subset of these datasets was used to train the NN-based speech classifier. The details of the speech classifiers investigated in this work are described in the sub-sections below:

2.1 | ACF-based speech classifier

The ACF of a speech frame, $x(n)$, can be defined as:

$$F_1(k) = \frac{1}{N} \sum_{n=0}^{N-k-1} x(n)\, x(n+k) \tag{1}$$

where N is the total number of samples in a speech frame and k is the lag number. The ACF of a speech signal consists of large amplitude peaks corresponding to the voiced speech frames and small amplitude peaks corresponding to the unvoiced speech frames. The decision on the voiced/unvoiced part is made by comparing the peak values with respect to a constant threshold.

The steps involved in the speech classification using this classifier are as follows:

(i) Take the speech signal to be analyzed.
(ii) Take the first frame of the speech signal.
(iii) Compute the ACF values for the speech frame using (1).
(iv) Obtain the highest peak value, T_{P1}, from the ACF values.
(v) Compare T_{P1} with a constant threshold value, T_{HR1}. If $T_{\mathrm{P1}} > T_{\mathrm{HR1}}$, then the speech frame is classified as a voiced frame, else it is classified as an unvoiced frame.
(vi) Repeat steps (i) to (v) for all speech frames.

2.2 | AMDF-based speech classifier

The AMDF of a speech frame, $x(n)$, can be defined as:

$$F_2(k) = \frac{1}{N} \sum_{n=0}^{N-k-1} |x(n) - x(n+k)| \tag{2}$$

where N is the total number of samples in a speech frame and k is the lag number. AMDF consists of several local minimum amplitude peaks in a voiced speech frame. These amplitude values are used to make the decision on a voiced/unvoiced frame. The following are the steps involved in using the AMDF-based classifier for speech classification:

(i) Take the speech signal to be analyzed.
(ii) Take the first frame of the speech signal.
(iii) Compute the AMDF values for the speech frame using (2).
(iv) Obtain the global minimum value, $T_{\min}$, from the AMDF values.
(v) Compare $T_{\min}$ with a constant threshold value, T_{HR2}. If $T_{\min} \le T_{\mathrm{HR2}}$ than the speech frame is classified as a voiced frame, else it is classified as an unvoiced frame.
(vi) Repeat steps (i) to (v) for all speech frames.

2.3 | Cepstrum-based speech classifier

The cepstrum is defined as the inverse discrete Fourier transform of the log magnitude of the discrete Fourier transform of a signal. The cepstrum of a speech frame, $x(n)$, can be obtained using the following equation:

$$C(n) = \sum_{n=0}^{N-1} \log\left(\left| \sum_{n=0}^{N-1} x(n)\, e^{-j\frac{2\pi kn}{N}} \right| \right) e^{j\frac{2\pi kn}{N}}. \tag{3}$$

The cepstrum of the unvoiced speech contains smaller magnitude cepstral peaks as compared to those for the voiced speech. Therefore, the speech can be classified by identifying the magnitude of the cepstral peak. The steps involved in using the cepstrum-based classifier for speech classification are as follows:

(i) Take the signal to be analyzed.
(ii) Take the first frame of the speech signal.
(iii) Compute the cepstrum for the speech frame.
(iv) Compute the threshold value, C_{T}.
(v) Compare the magnitude of the cepstral peaks, C_i, (where i is the number of cepstral peaks in a frame) for the speech frame. If $C_i > C_{\mathrm{T}}$, then the speech frame is classified as a voiced frame, else it is classified as an unvoiced speech frame.
(vi) Repeat steps (i) to (v) for all speech frames.

The threshold value, C_{T}, chosen was the median value of the magnitude of the cepstral peaks. This threshold was calculated and updated for every utterance in the speech signal.

2.4 | WACF-based speech classifier

This classifier is based on the combination of features of both ACF and AMDF. The WACF is defined as.

$$F_3(k) = \frac{F_1(k)}{\left(F_2(k) + \alpha\right)}, \tag{4}$$

where $F_1(k)$ is ACF (defined by (1)), and $F_2(k)$ is AMDF (defined by (2)) of speech frame $x(n)$. k is the lag number and α is a fixed number ($\alpha < 0$) used for avoiding the condition of $F_2(k) = 0$ at $k = 0$. The steps involved in using this classifier for speech classification are as follows:

(i) Take the speech signal to be analyzed.
(ii) Take the first frame of the speech signal.
(iii) Compute the WACF values for the speech frame using (4).
(iv) Obtain the highest peak value, T_{P2}, from the WACF values.
(v) Compare $T_{\mathrm{P2}}$with a constant threshold value, T_{HR3}. If $T_{\mathrm{P2}} > T_{\mathrm{HR3}}$, then the speech frame is classified as a voiced frame, else it is classified as an unvoiced frame.
(vi) Repeat steps (i) to (v) for all speech frames.

2.5 | ZCR-E-based speech classifier

The ZCR for a speech frame $x(n)$ can be obtained using the following equation:

$$F_{\mathrm{ZCR}} = \frac{1}{2} \sum_{n=1}^{N} |\mathrm{sgn}(x(n)) - \mathrm{sgn}(x(n-1))|. \tag{5}$$

The average energy of a speech frame, $x(n)$, can be obtained by:

$$E = \frac{1}{N} \sum_{n=0}^{N-1} |x(n)|^2. \tag{6}$$

Here, N is the total number of samples in the speech frame. The ZCR has the property that the speech frame is likely to be unvoiced if ZCR exceeds a given threshold. Otherwise, the frame is likely to be a voiced speech frame. On the other hand, the energy for a voiced speech frame is quite high as compared to that for an unvoiced speech frame. These two features have been used in the ZCR-E-based speech classifiers. The steps involved in using this algorithm for speech classification are as follows:

(i) Take the speech signal to be analyzed.
(ii) Take the first frame of the speech signal.
(iii) Compute the energy, E, for the speech frame using (6).
(iv) Compare E with a constant threshold value, E_{THR}. If $E \leq E_{\mathrm{THR}}$, then the speech frame is classified as an unvoiced frame.
(v) If $E > E_{\mathrm{THR}}$, then obtain ZCR (F_{ZCR}) using (5) for the speech frame and compare it with constant threshold, T_{ZCR}. If $F_{\mathrm{ZCR}} \leq T_{\mathrm{ZCR}}$, then the speech frame is classified as a voiced frame, else it is classified as an unvoiced frame.

The threshold values, T_{HR1}, $T_{\mathrm{HR2,}}$ T_{HR3}, and T_{ZCR}, for the entire frame of each utterance are taken as the median values of T_{P1}, T_{min}, T_{P2}, and F_{ZCR}, respectively. These thresholds were calculated and updated for every utterance in the database. The short-time energy threshold of $E_{\mathrm{THR}} = 0.05$ was considered in the ZCR-E speech classifier.

2.6 | NN-based speech classifier

The steps involved in using the NN-based speech classifier for speech classification are as follows:

(i) Take the speech signal to be analyzed.
(ii) Take the first frame of the speech signal.
(iii) Compute the feature vector using waveform analysis and linear predictive (LP) analysis.
(iv) Train the network with the training samples.
(v) Compute the performance of the network classifier using the dataset.

In this work, an NN with a single hidden layer having 20 nodes (15-20-2 network architecture) has been used. A feature vector was obtained for each frame of 20 ms which consists of 13 cepstral coefficients and two waveform parameters (rms energy and ZCR). The energy prediction error and 12 LP coefficients were used to derive the cepstral coefficients. The LP coefficients were obtained by using windowing, autocorrelation, and pre-emphasis. A generalized delta rule for the back propagation of error (with a learning rate of 0.9) was used to train the network. A subset of data was randomly selected from the speech database (that is, the database which has been used for the performance evaluation) to train the network. The voiced/unvoiced frame classification of the input signals was made after the network training was completed. The output vector is a binary decision of the voiced/unvoiced frame, where the first and second outputs indicate the voiced frame and unvoiced frame, respectively. An output vector coded as [1, 0] indicates a voiced frame, while the vector [0, 1] indicates an unvoiced frame.

3 | IMPLEMENTATION OF ANALYSIS-SYNTHESIS SYSTEM USING DIFFERENT SPEECH CLASSIFIERS

All six speech classifiers have been developed using the SIMULINK and integrated in the LPC-based speech analysis-synthesis system [22,23] which is based on the source

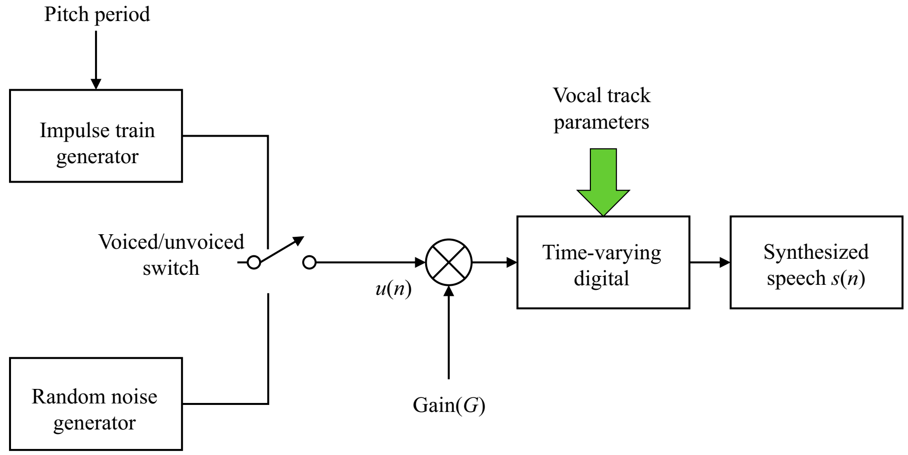

FIGURE 1 Block diagram of the speech production model

filter model of speech production. A block diagram of the speech production model is presented in Figure 1. Here, the excitation is presented as an impulse train and random noise sequence for the voiced and unvoiced speech respectively.

In the LPC analysis-synthesis system, the vocal tract is modeled as an all-poll infinite impulse response filter and the transfer function is given by:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a(k) z^{-k}}, \tag{7}$$

where, $S(z)$ and $U(z)$ are the z-transform of the synthesized speech signal s(n) and excitation signal $u(n)$, respectively. G represents the gain of the filter and $a(k)$ represents the filter coefficients which are calculated by:

$$\mathbf{a} = -\mathbf{R}^{-1} r_n, \tag{8}$$

where $\mathbf{R}$ is a $p \times p$ matrix and $\mathbf{r}_n$ is the $p \times 1$ matrix of the speech signal. The gain of the filter is calculated as:

$$G^2 = r_n(0) - \sum_{k=1}^{p} a(k) r_n(k). \tag{9}$$

Six separate analysis-synthesis models have been developed using the speech classifiers described in Section 2. The autocorrelation method has been used for extracting the pitch from the speech signals in all the cases. For the analysis, the speech signals were divided into frames of length 20 ms. Using linear predictive analysis [22,23], the filter parameters, namely, the gain, G, the filter coefficients, $a(k)$, the voiced/unvoiced frame classification, and the pitch period, were determined from the speech signal. Prediction order 15 was used for the LPC analysis. At the synthesis stage, an impulse train corresponding to the estimated pitch period of the voiced frame was generated. Random-noise-like excitation was used for the unvoiced frame. Finally, the speech signal was reconstructed using a proper excitation signal, gain, and filter coefficient.

The simulation models were created in SIMULINK (in MATLAB) using the blocks available in its library. The blocks which were not available in the library were created using the *Embedded Matlab function* utility. The simulation models were modified to build real-time models. The TMS320C6713 DSP starter kit (DSK) was used for real-time implementation and a*C6713DSK target preference* block (available in SIMULINK) was used to configure the simulation model. The sampling and framing of the input speech signal was achieved by the *analog-to-digital* (*ADC*) block, while a *digital-to-analog* (*DAC*) block was used at the output port of the system for obtaining a continuous signal from the digital signal. The *rate transition* blocks were used (inserted between the two blocks which were operating at different sampling rates) to ensure deterministic data transfer in real time. A *buffer* block at the output port just before the *DAC* block was used for obtaining continuous speech output (samples) without any loss of the sample.

4 | PERFORMANCE COMPARISON RESULTS

The performance of the speech classifiers has been compared in terms of the percentage of their classification accuracy, speech quality (MOS and PESQ), [27,28] and execution time [23] by applying them to the PTDB-TUG [25] speech database (clean speeches) and NOIZEUS (noisy speeches with different SNRs) [26]. The PTDB-TUG database consists of recordings of 20 English speakers (10 male and 10 female) reading phonetically rich sentences from the TIMIT database. A subset of this database, consisting of 40 speech files (20 by males and 20 by females) was used in the investigation. White noise was added to the PTDB-TUG database in order to obtain different levels of SNRs (15 dB, 10 dB, 5 dB, 0 dB and –5 dB). The NOIZEUS database consists of recordings of 30 sentences spoken by six English speakers (three male and three female) and corrupted by noise due to cars, babble noise, noise in exhibition halls, restaurant noise, suburban train noise, train-station noise, noise on airports, and street noise. Since white noise, car noise, and babble noise (crowd of people) are usually used in speech processing, these three types of noise were used for the performance evaluation of the speech classifiers chosen in this work.

4.1 | Percentage of the voiced/unvoiced speech classification accuracy

The percentage of the voiced/unvoiced speech classification accuracy for speech signals having different levels of SNRs has been calculated and listed in Table 1. Before adding noise samples in each of the utterance, the percentage of voiced speech samples was maintained at 50% by appending a required duration of silence. Manual classification of speech material was performed by two experienced people. The percentage classification accuracy is computed as:

$$P_{\mathrm{C}} = 1 - \left(0.5 \times P_{\mathrm{VU}} + 0.5 \times P_{\mathrm{UV}}\right) \tag{10}$$

where P_{VU} denotes the percentage of voiced speech classified as unvoiced and P_{UV} denotes the percentage of unvoiced speech classified as voiced.

From the obtained results given in Table 1, it can be seen that the classification accuracy of all speech classifiers is good

TABLE 1 Percentage of the voice/unvoiced speech classification accuracy for the six speech classifiers investigated in this work

	SNR		Speech classifiers					
Noise type	**(dB)**	$P_C/P_{VU}/P_{UV}$	**ACF**	**AMDF**	**Cepstrum**	**WACF**	**ZCR-E**	**NN**
Clean speech	Clean speech	P_C	96.55	95.52	95.60	96.78	96.22	96.92
		P_{VU}	3.12	3.80	3.35	3.32	3.92	3.04
		P_{UV}	3.77	5.16	5.45	3.13	3.64	3.12
Speech corrupted with white noise	15	P_C	96.14	95.10	95.09	96.62	95.12	96.88
		P_{VU}	3.68	3.89	3.92	3.85	7.92	3.24
		P_{UV}	4.04	5.91	5.90	2.91	1.84	3.00
	10	P_C	95.48	84.25	84.40	95.65	84.62	96.16
		P_{VU}	3.54	1.25	6.92	5.26	29.35	4.42
		P_{UV}	5.51	30.26	24.28	3.44	1.41	3.26
	5	P_C	93.12	63.33	63.39	94.75	63.40	95.36
		P_{VU}	3.85	1.03	8.02	6.75	72.88	5.64
		P_{UV}	9.91	72.31	65.20	3.75	0.32	3.64
	0	P_C	61.83	51.05	53.35	85.62	55.32	87.92
		P_{VU}	0.65	0.05	1.35	8.65	19.36	7.71
		P_{UV}	75.69	97.85	91.95	20.11	70.00	16.45
	–5	P_C	50.76	50.00	50.03	64.84	52.34	67.54
		P_{VU}	0.06	00.00	0.01	5.68	15.32	6.95
		P_{UV}	98.42	100.00	99.93	64.64	80.00	57.97
Speech corrupted with car noise	15	P_C	96.02	94.90	94.98	96.56	95.00	96.75
		P_{VU}	3.11	2.25	2.89	3.72	7.89	3.66
		P_{UV}	4.85	7.95	7.15	3.15	2.10	2.84
	10	P_C	95.15	81.81	82.66	95.29	82.86	96.62
		P_{VU}	3.42	1.17	2.91	4.58	8.74	3.92
		P_{UV}	6.28	35.21	31.77	4.84	25.54	2.84
	5	P_C	76.72	56.61	58.88	88.32	61.34	90.86
		P_{VU}	1.15	0.83	1.05	5.32	17.18	5.46
		P_{UV}	45.42	85.95	81.19	18.04	59.56	12.82
	0	P_C	57.13	50.00	52.52	69.34	53.12	72.54
		P_{VU}	0.12	00.00	0.06	6.24	3.12	5.98
		P_{UV}	85.63	100.00	94.90	55.08	90.64	48.94
	–5	P_C	50.22	50.00	50.02	54.59	52.94	59.62
		P_{VU}	00.02	00.00	0.02	1.23	0.75	2.14
		P_{UV}	99.54	100.00	99.94	89.59	93.37	78.62
Speech corrupted with babble noise	15	P_C	95.29	92.21	92.34	96.20	94.12	96.72
		P_{VU}	3.45	2.12	3.14	3.45	6.50	3.61
		P_{UV}	5.98	13.46	12.18	4.15	5.26	2.95
	10	P_C	77.07	69.45	70.11	78.95	76.83	82.98
		P_{VU}	2.31	0.98	1.21	3.12	7.32	4.61
		P_{UV}	43.55	60.12	58.57	38.98	39.02	29.43
	5	P_C	64.99	53.50	55.12	65.54	60.13	71.87
		P_{VU}	0.91	0.05	0.90	1.85	12.31	2.29
		P_{UV}	69.12	92.95	88.86	67.07	67.43	53.97

(Continues)

TABLE 1 (Continued)

Noise type	SNR (dB)	$P_C/P_{VU}/P_{UV}$	Speech classifiers ACF	 AMDF	 Cepstrum	 WACF	 ZCR-E	 NN
	0	P_C	50.80	50.00	50.10	51.95	52.32	61.06
		P_{VU}	0.04	0.00	0.04	1.63	2.98	1.98
		P_{UV}	98.37	100.00	99.76	94.47	92.38	75.9
	–5	P_C	50.00	50.00	50.00	50.80	51.15	58.64
		P_{VU}	0.00	0.00	0.00	1.16	0.45	2.21
		P_{UV}	100.00	100.00	100.00	97.24	97.25	80.51

for clean speech and gets degraded as the SNR of the sample is decreased. The accuracy of the NN-based speech classifier is observed to be higher than the ACF-, AMDF-, WACF- and ZCR-E-based speech classifiers for all types of noise considered in the performance evaluation. The NN-based classifier has the highest percentage accuracy of 96.92% for clean speech. However, in the case of noisy speech, it is 96.88% for white noise with an SNR level of 15dB. This classifier has the lowest classification accuracy of 58.64% for the babble noise with an SNR of –5dB, where P_{UV} is maximum, that is, 80.51%. However, in case of white noise and car noise with –5 dB SNR, the lowest accuracies for the NN-based classifier of 67.54% and 59.62% respectively are obtained. For these cases, P_{VU} and P_{UV} are 6.95, 57.97, and 2.14, 78.62, respectively. The range of accuracy for the NN-based classifier lies between 58.64% and 96.92%. The highest classification accuracies for the ACF-, AMDF-, cepstrum-, WACF- and ZCR-E-based speech classifiers for clean speech are 96.55%, 95.52%, 95.60%, 96.78%, and 96.22% respectively. However, the lowest classification accuracy for these methods are 50.00%, 50.00%, 50.00%, 50.80%, and 51.15% in the case of speech corrupted with babble noise. In these cases, the P_{UV} are 100.00%, 100.00%, 100.00%, 97.24%, and 97.25%. From the results, it can be observed that for babble noise with –5dB SNR, the ACF-, AMDF- and cepstrum-based classifiers classify all the unvoiced speech segments as voiced speech segments, which results in the lowest classification accuracy. The highest classification accuracies of these classifiers in the case of noisy speech are 96.14%, 95.10%, 95.09%, 96.62%, and 95.12%, respectively, for speech corrupted using white noise with an SNR of 15dB. Based on these results, the ranking of the different classifiers based on their classification accuracy in the decreasing order and corresponding to higher levels of SNR (15 dB to 5 dB) is as follows: NN > WACF > ACF > ZCR-E > cepstrum > AMDF.

From the results corresponding to the white noise and car noise having 0 dB SNR, the accuracy of NN-and WACF-based classifier is observed to be higher than the other classifiers. The ranking based on their performance from high to low accuracy is as follows: NN > WACF > ACF > ZCR-E > cepstrum > AMDF. However, the results corresponding to babble noise with 0 dB SNR, show that the accuracy of NN- and ZCR-E-based speech classifiers is higher than the other classifiers and ranking order in this case (from high to low accuracy) is: NN > ZCR-E > WACF > ACF > cepstrum > AMDF.

From the results corresponding to the white noise and car noise with very low level of SNR (–5 dB), the classification accuracy for the NN- and WACF-based classifiers is higher as compared to the other speech classifiers. The order of their performance from high accuracy to low accuracy is: NN > WACF > ZCR-E > ACF > cepstrum > AMDF. However, in the case of babble noise, NN- and ZCR-E-based classifiers exhibit a higher accuracy as compared to the other classifiers. The ranking from high to low accuracy in this case is: NN > ZCR-E > WACF > ACF > cepstrum > AMDF.

Thus, from the results discussed above, it can be seen that the classification accuracy of the AMDF-based speech classifier is worst in all the cases (that is, for both clean and noisy speech signals) and its accuracy range lies between 50.00%–95.52%. Further, it can be observed that the degradation in the performance of all speech classifiers (in most of the cases) is mainly due to the P_{UV} error (where the unvoiced speech is classified as voiced), which increases with the degradation in the SNR levels. In addition, the accuracy of the speech classifiers is also observed to vary with the different types of noise.

4.2 | Results of the subjective (MOS) and objective (PESQ) speech quality tests

In this work, the MOS listening test, which is a well-known subjective method for measuring the speech quality, has been used for comparing the performance of the six speech classifiers. In this test, 25 listeners were chosen to rate their response on a scale of 1 and 5 (where a score of 1 corresponds to a low speech quality and 5 corresponds to an excellent speech quality). The test results thus obtained are presented in Table 2. The MOS scores for the original unprocessed speech have also been presented in the table for reference.

An objective speech quality test, namely, PESQ, based on the ITU-T P.862 recommendation, has been performed for evaluating the performance of the six speech classifiers.

TABLE 2 Results of the MOS test for the six speech classifiers

Noise type	SNR (dB)	Original unprocessed speech	Speech classifiers					
			ACF	AMDF	Cepstrum	WACF	ZCR-E	NN
Clean speech	Clean speech	4.60	3.45	3.25	3.25	3.52	3.30	3.72
Speech corrupted with white noise	15	4.45	3.39	3.16	3.18	3.47	3.18	3.54
	10	4.18	3.13	2.72	2.75	3.16	2.82	3.29
	5	3.42	2.78	2.31	2.35	2.82	2.36	3.02
	0	2.66	1.98	1.80	1.88	2.16	1.96	2.29
	–5	2.13	1.49	1.38	1.40	1.75	1.52	1.88
Speech corrupted with car noise	15	4.45	3.38	3.10	3.10	3.45	3.12	3.68
	10	4.15	3.11	2.70	2.75	3.15	2.80	3.23
	5	3.40	2.75	2.30	2.32	2.80	2.35	2.98
	0	2.65	1.98	1.80	1.85	2.15	1.95	2.21
	–5	2.12	1.47	1.36	1.40	1.62	1.52	1.79
Speech corrupted with babble noise	15	4.43	3.34	3.02	3.05	3.42	3.09	3.67
	10	4.10	3.05	2.66	2.72	3.12	2.75	3.20
	5	3.36	2.66	2.24	2.30	2.76	2.31	2.94
	0	2.60	1.87	1.75	1.79	1.91	1.92	2.02
	–5	2.03	1.35	1.27	1.32	1.39	1.50	1.68

In the PESQ test, the processed speech signal has been compared with the original speech signal. The resultant PESQ score is marked with a scale ranging between –0.5 and 4.5.

From the results of MOS and PESQ presented in Tables 2 and 3, respectively, it can be seen that the speech quality for all speech classifiers is satisfactory for clean speech as well as speech corrupted with lower SNRs. The speech quality for the NN-based speech classifier is better than that for the other five speech classifiers in the case of both clean speech and noisy speech with SNR levels from –5 dB to 15 dB. The highest MOS and PESQ scores obtained for the NN-based classifier are 3.72 and 3.44, respectively for the clean speech samples. The range of the MOS score for this classifier lies between 1.88–3.54, 1.79–3.68, and 1.68–3.67

TABLE 3 Results of the PESQ test for the six speech classifiers

Noise type	SNR (dB)	Original unprocessed speech	Speech classifiers					
			ACF	AMDF	Cepstrum	WACF	ZCR-E	NN
For clean speech	Clean speech	4.50	3.28	3.12	3.12	3.29	3.12	3.44
Speech corrupted with white noise	15	4.01	2.94	2.79	2.82	2.97	2.85	3.10
	10	3.77	2.56	2.41	2.46	2.71	2.55	2.94
	5	2.28	2.26	1.88	1.90	2.33	2.21	2.67
	0	2.05	1.45	1.35	1.36	1.59	1.40	1.76
	–5	1.47	0.98	0.93	0.93	1.01	1.00	1.20
Speech corrupted with car noise	15	3.98	2.92	2.77	2.80	2.95	2.85	3.02
	10	3.59	2.55	2.35	2.40	2.70	2.53	2.90
	5	2.25	2.20	1.85	1.89	2.31	2.19	2.53
	0	1.98	1.39	1.31	1.33	1.54	1.36	1.68
	–5	1.42	0.95	0.90	0.92	1.00	0.96	1.17
Speech corrupted with babble noise	15	3.95	2.91	2.77	2.79	2.95	2.83	3.02
	10	3.45	2.55	2.34	2.37	2.69	2.53	2.89
	5	2.19	2.19	1.69	1.73	2.30	2.13	2.48
	0	1.85	1.14	1.06	1.11	1.23	1.35	1.55
	–5	1.31	0.89	0.82	0.85	0.92	0.95	1.12

for speech corrupted with white, car, and babble noise, respectively, whereas, the range of the PESQ score for the same lies between 1.20–3.10, 1.17–3.02, and 1.12–3.02, respectively. The speech quality for the AMDF-based classifier is poor as compared to that for the other five classifiers in all the cases. The highest and lowest MOS scores for the AMDF-based classifier are 3.25 (for clean speech) and 1.27 (for speech corrupted with babble noise having a –5dB SNR), whereas, the highest and lowest PESQ scores for this classifier are 3.12 and 1.27, respectively.

For speech corrupted by white noise and car noise with a 0 dB SNR level, the speech quality for the NN-based classifier is superior to that of the other five classifiers. In this case, the MOS scores for the NN-based classifier are 2.29 and 2.21, respectively, while the PESQ scores are 1.76 and 1.68, respectively for the two types of noise. The speech quality for the WACF-based speech classifier is higher than that for the ACF-, AMDF-, cepstrum-, and ZCR-E-based classifiers. For the WACF-based classifier, the MOS scores corresponding to white noise and car noise with 0 dB SNR are 2.16 and 2.15, respectively, whereas, the PESQ scores for the same are 1.59 and 1.54, respectively. In the case of ACF-, AMDF-, cepstrum-, and ZCR-E-based classifiers applied to speech data corrupted with white noise and car noise having 0 dB SNR, the MOS scores are 1.98, 1.80, 1.88, and 1.96 and 1.98, 1.80, 1.85, and 1.95, respectively. The PESQ scores for the same are 1.45, 1.35, 1.36, and 1.40 and 1.39, 1.31, 1.33, and 1.36, respectively. In the case of speech samples corrupted by babble noise having 0 dB SNR, the speech quality obtained from the ZCR-E-based classifier is higher than that obtained from the ACF-, AMDF-, cepstrum-, and WACF-based speech classifiers. In this case, the MOS and PESQ scores for the ZCR-E-based classifier are 1.92 and 1.35, respectively. Similarly, the MOS and PESQ scores for the ACF-, AMDF-, cepstrum-, and WACF-based classifiers are 1.87, 1.75, 1.79, and 1.91 and 1.14, 1.06, 1.11, and 1.23, respectively.

For very low SNR (–5 dB) where the speech has been corrupted by white noise and car noise, the speech quality for the NN-based classifier is the highest, followed by the WACF-ACF-, ZCR-E-, cepstrum-, and AMDF-based classifiers. In this case, the highest MOS and PESQ scores, 1.75 and 1.01, respectively are obtained for the WACF-based classifier applied to speech corrupted by white noise. The MOS and PESQ scores obtained for this classifier when applied to speech corrupted by car noise are 1.62 and 1.00, respectively. The MOS and PESQ scores for ACF-, ZCR-E-, cepstrum- and AMDF-in the case of white noise (having –5 dB SNR) are 1.49, 1.52, 1.40, and 1.38 and 0.98, 1.00, 0.93, and 0.93, respectively. However, in the case of speech corrupted by car noise, the MOS and PESQ scores for these classifiers are 1.47, 1.52, 1.40, and 1.36 and 0.95, 0.96, 0.92, and 0.90, respectively. In case of speech corrupted by babble noise having a very low SNR of –5 dB, the quality of the synthesized speech for the NN-based classifier is the highest, with MOS and PESQ scores of 1.68 and 1.12, respectively, and the quality for the ZCR-E-based classifier is better than that for the ACF-, WACF-, cepstrum-, and AMDF-based classifiers. In this case, the MOS and PESQ scores for ZCR-E-based classifiers are 1.50 and 0.95, respectively, while the corresponding scores for the ACF-, WACF-, cepstrum-, and AMDF-based classifiers are 1.35, 1.39, 1.32, and 1.27 and 0.89, 0.92, 0.85, and 0.82, respectively.

4.3 | Computation time

4.3.1 | Simulation time

A total of 20 speech files were used to calculate the average simulation time for the system using the six speech classifiers. Each file was tested 10 times and the simulation time taken during each test iteration was noted using the *profiler* tool [23]. This process was repeated for each classifier and an average of these measurements for each speech file was calculated for the six classifiers. The resultant values are presented in Table 4.

From the table, it can be seen that the simulation time for the AMDF-based speech classifier is the smallest as compared to that corresponding to the other speech classifiers. The minimum and maximum simulation times for the AMDF-based classifier are 19.98 ms and 20.88 ms, respectively. The NN-based method takes a longer amount of simulation time compared to the other five speech classifiers. The minimum simulation time for the NN-based method is 20.84 ms. The cepstrum-based method takes a longer amount of simulation time as compared to the other methods except the NN-based speech classifier. The minimum and maximum simulation times for the cepstrum-based classifier are 20.39 ms and 21.32 ms, respectively. Further, the simulation time for the ACF-based speech classifier is less than that for the WACF-, NN-, cepstrum- and ZCR-E-based speech classifiers. The minimum simulation time for the ACF-based classifier is 20.10 ms.

4.3.2 | Execution time (real-time implementation)

The execution times for the analysis-synthesis system employing the six speech classifiers have been calculated using the breakpoint method [23]. The results thus obtained are presented in Table 5. The execution time was calculated as follows:

$$\text{Execution Time} = \\ (\text{Average number of cycles for each frame} \times \text{Execution time per cycle}) \tag{11}$$

where the average number of cycles was calculated for each frame. The length of each frame was 20 ms. The execution time per cycle was 4.44 ns for the TMS320C6713 processor (having a clock frequency of 225 MHz).

TABLE 4 Values of the average simulation time for the six speech classifiers

Speech file (.wav)	Length of speech (sec)	Total number of frames processed	Speech classifiers (Time/frame in ms) ACF	AMDF	Cepstrum	WACF	ZCR-E	NN
s1	6.88	344	20.26	20.21	20.45	20.40	20.38	21.04
s2	7.41	371	20.12	20.05	20.39	20.32	20.31	20.95
s3	7.99	400	20.65	20.62	20.90	20.81	20.74	20.94
s4	8.06	403	20.55	20.35	20.85	20.72	20.65	20.89
s5	9.43	472	20.41	20.39	20.85	20.55	20.52	21.02
s6	9.16	458	20.95	20.88	21.32	21.19	21.07	21.85
s7	7.14	357	20.44	20.37	20.92	20.75	20.65	21.10
s8	6.87	344	20.62	20.52	20.90	20.84	20.81	21.12
s9	6.75	338	20.26	20.12	20.75	20.51	20.44	20.94
s10	6.87	344	20.10	19.98	20.45	20.39	20.33	20.84
s11	6.16	308	20.46	20.35	20.92	20.78	20.62	20.98
s12	7.47	374	20.22	20.16	20.69	20.55	20.49	20.97
s13	7.92	396	20.47	20.45	21.02	20.91	20.74	21.21
s14	5.61	281	20.17	20.15	20.68	20.52	20.31	20.95
s15	6.82	341	20.39	20.31	20.90	20.73	20.52	21.14
s16	5.60	280	20.14	20.09	20.72	20.51	20.38	20.85
s17	6.74	337	20.51	20.36	20.98	20.89	20.76	21.08
s18	7.12	356	20.40	20.34	20.76	20.67	20.66	20.88
s19	6.12	306	20.29	20.22	20.85	20.77	20.72	20.96
s20	6.62	331	20.20	20.12	20.69	20.59	20.49	20.98

TABLE 5 Execution time for the six speech classifiers

Speech classifier used	Avg no. of cycles	Execution time (in ms)
ACF	4 428 878	19.664
AMDF	4 156 321	18.454
Cepstrum	4 459 988	19.802
WACF	4 458 984	19.797
ZCR-E	4 451 572	19.764
NN	4 491 198	19.941

From the result presented in Table 5, it can be seen that the execution time for the AMDF-based speech classifier is 18.454 ms, which is the least as compared to the other five speech classifiers. The execution time for the NN-based speech classifier (19.941 ms) is higher than that for the other speech classifiers while that for the cepstrum-based classifier is 19.802 ms, which is more than that for the ACF-, AMDF-, WACF- and ZCR-E-based classifiers, but less than that for the NN-based classifier. The execution times for the ACF-, WACF- and ZCR-E-based classifiers are 19.664 ms, 19.797 ms, and 19.764 ms, respectively. The NN-, cepstrum-, WACF- and ZCR-E-based speech classifiers have higher execution times as compared to the ACF- and AMDF-based classifiers. This can be explained as follows: the AMDF function involves only a modulus and addition operation and hence is computationally simpler than the other speech classifiers. The ACF-based classifier is has a higher computational complexity as compared to the AMDF-based classifier since it involves a summation of products [23,29,30]. The cepstrum method is computationally complex because it involves the computation of the Fourier transform, inverse Fourier transform, and logarithmic operation of the power spectrum. However, due to the multiple features used in the hybrid classifiers (WACF, ZCR-E and NN), they are computationally more complex than the ACF- and AMDF-based classifiers.

From the results of the performance comparison presented above, it can be seen that the overall ranking of the speech classifiers investigated in this work is not a simple task, since all the classifiers do not perform well (in terms of the percentage classification accuracy and speech quality) in all the situations and their performance differs for different SNR. In addition, their performance also depends on the type of background noise. From the results of the computation time, it is seen that the classifiers that exhibit a higher classification accuracy and speech quality also exhibit higher computational complexity, while some exhibit lower computational complexity as well as lower classification accuracy and speech quality. Thus, as a solution of this

TABLE 6 Performance ranking[†] of the six speech classifiers on the basis of their synthesized speech quality and percentage classification accuracy

		Speech classifiers					
SNR level	Noise type	ACF	AMDF	Cepstrum	WACF	ZCR-E	NN
15 dB, 10 dB, 5 dB	white, car, babble	3rd	6th	5th	2nd	4th	1st
0 dB	white, car	3rd	6th	5th	2nd	4th	1st
	babble	4th	6th	5th	3rd	2nd	1st
–5 dB	white, car	4th	6th	5th	2nd	3rd	1st
	babble	4th	6th	5th	3rd	2nd	1st

[†] The 1st rank corresponds to the highest classification accuracy and speech quality, while 6th rank represents the lowest accuracy and synthesized speech quality.

TABLE 7 Performance ranking[‡] of the speech classifiers based on their computation time

ACF	AMDF	Cepstrum	WACF	ZCR-E	NN
2nd	1st	5th	4th	3rd	6th

[‡] The 1st and 6th ranks represent the lowest and the highest computational complexity, respectively.

problem, the ranking was divided into two categories. In the first category, two parameters, namely, the percentage classification accuracy and speech quality, were used to rank the performance of the six speech classifiers, as presented in Table 6. In the second category, the computational complexity, measured on the basis of the simulation and execution time, was used to rank the different speech classifiers, as presented in Table 7.

Based on the results of performance ranking, we can see that the NN-based speech classifier is better than the other five classifiers for clean as well as noisy speech with higher as well as lower values of SNR. For all the cases, the NN-based classifier turns out to be a superior choice. However, the WACF-based classifier may be used as a second choice as it shows a better performance as compared to the ACF-, AMDF- and ZCR-E-based classifiers in the case of noisy speech with an SNR of 15 dB, 10 dB, 5 dB, and 0 dB (except babble noise having 0 dB SNR). However, both these hybrid speech classifiers are computationally complex than the ACF-, AMDF- and ZCR-E-based speech classifiers. Based on the performance evaluation results in case of higher SNRs(15dB, 10dB and 5dB), the performance ranking of the classifiers, in terms of their classification accuracy and synthesized speech quality, from first to last are: NN, WACF, ACF, ZCR-E, cepstrum and AMDF. In case of lower SNRs (0 dB and –5 dB), NN-, ACF-, cepstrum- and AMDF-based classifiers are ranked 1st, 4th, 5th, and 6th respectively in terms of their classification accuracy and speech quality, while the performance of WACF- and ZCR-E-based classifiers in this case is comparable to each other. From Table 7, it can be seen that the AMDF-, ACF-, ZCR-E-, WACF-, cepstrum and NN-based classifiers are ranked 1st, 2nd, 3rd, 4th, 5th, and 6th, respectively, on the basis of their computational complexity. Further, the AMDF-based classifier exhibits less computational complexity than the other classifiers and may turn out to be used as the first choice in the cases where the computational complexity is an issue. However, the NN-based classifiers can be used as the first choice where a higher classification accuracy and synthesized speech quality are desirable.

5 | CONCLUSIONS

Six voiced/unvoiced speech classifiers based on ACF, AMDF, cepstrum WACF, ZCR-E and NN have been simulated and implemented in real time using a TMS320C6713 DSP starter kit. The performance of these classifiers has been measured by integrating them into an LPC-based speech analysis-synthesis system. The percentage voiced/unvoiced classification accuracy, speech quality, and computation time were chosen as the parameters for carrying out their performance comparison. On the basis of these parameters, the overall performance ranking of the classifiers was established in two categories. It has been found that all six speech classifiers perform well for clean speech, but their performance degrades with a degradation in the SNR. The performance of all six classifiers also varies with the variation in the type of background noise. Results of the percentage voiced/unvoiced classification accuracy and speech quality show that the NN-based speech classifier performs better than the other five classifiers for all SNR levels. However, the percentage classification accuracy and speech quality for the AMDF-based speech classifier is the poorest in all the cases. Further, the performance of the WACF- and ZCR-E-based classifiers is comparable for very low SNR (–5 dB) level. The computational complexity of the AMDF-based speech classifier is the least as compared to the other five classifiers while that corresponding to the NN-based classifier is the highest.

ORCID

Sandeep Kumar https://orcid.org/0000-0001-9922-2663

REFERENCES

1. S. Ahmadi and A. Spanisa, *Cepstrum-based pitch detection using a new statistical V/UV classification algorithm*, IEEE Trans. Speech, Audio Process. **7** (1999), no. 3, 333–338.
2. A. Mousa, *Speech segmentation in synthesized speech morphing using pitch shifting*, Int. Arab J. Inf. Technol. **8** (2011), no. 2, 221–226.
3. S. Kumar, S. K. Singh, and S. Bhattacharya, *Performance evaluation of a ACF-AMDF based pitch detection scheme in real time*, Int. J. Speech Technol. **18** (2015), no. 4, 521–527.
4. Y. Faycal and M. Bensebti, *Comparative performance study of several features for voiced/ Non-voiced classification*, Int. Arab J. Inf. Technol. **11** (2014), no. 3, 293–299.
5. R. G. Bachu et al., *Voiced/Unvoiced decision for speech signals based on zero-crossing rate and energy*, Advanced Techniques in Computing Sciences and Software Engineering, K. Elleithy (eds), Springer, Dordrecht, Netherlands, 2010, pp. 279–282.
6. L. Janer, J. J. Bonet, and E. L. Solano, *Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms*, in Proc. Int. Conf. Spoken Language Process. (Philadelphia, PA, USA), Oct. 1996, pp. 1209–1212.
7. S. Kumar et al., *Performance evaluation of a wavelet-based pitch detection scheme*, Int. J. Speech Technol. **16** (2013), no. 4, 431–417.
8. K. M. Hassan, E. Hamid, and K. I. Molla, *A method for voiced/unvoiced classification of noisy speech by analyzing time-domain features of spectrogram image*, Sci. J. Circuits, Syst. Signal Process. **6** (2017), no. 2, 11–17.
9. B. S. Atal and L. R. Rabiner, *A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition*, IEEE Trans. Acoust., Speech, Signal Process. **24** (1976), no. 3, 201–212.
10. J. K. Shah et al., *Robust voiced/unvoiced classification using novel features and Gaussian mixture model*, 2004, available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.618.2362&rep=rep1&type=pdf
11. Y. Qi and B. R. Hunt, *Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier,* IEEE Trans, Speech, Audio Process. **1** (1993), no. 2, 250–255.
12. T. Drugman et al., *Traditional machine learning for pitch detection*, IEEE Signal Process. Lett. **25** (2018), no. 11, 1745–1749.
13. S. Bagavathi and S. I. Padma, *Neural network based voiced and unvoiced classification using EGG and MFCC feature*, Int. Research J. Eng. Technol. **4** (2017), no. 4, 1934–1937.
14. A. Bendiksen and K. Steiglitz, *Neural networks for voiced/unvoiced speech classification*, in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (Albuquerque, NM, USA), Apr. 1990, pp. 521–524.
15. B. H. Juang and L. R. Rabiner, *Spectral representations for speech recognition by neural networks-A tutorial*, in Proc. Neural Netw. Signal Process. II Proc. IEEE Workshop (Helsingoer, Denmark), 1992, pp. 214–222.
16. M. Sharma and R. Mammone, *Automatic speech segmentation using neural tree networks*, in Proc. IEEE Workshop Neural Netw. Signal Process. (Cambridge, MA, USA), 1995, pp. 282–290.
17. K. Khaldi, A. Boudraa, and M. Turki, *Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement*, IET Signal Process. **10** (2016), no. 1, 69–80.
18. Z. Ali and M. Talha, *Innovative method for unsupervised voice activity detection and classification of audio segments*, IEEE Access **6** (2018), 15494–15504.
19. G. Sun et al., *The complexity analysis of voiced and unvoiced speech signal based on sample entropy*, in Proc. Int. Conf. Math. Comput. Sci. Industry (Corfu, Greece), Aug. 2017, pp. 26–29.
20. K. Struwe, *Voiced-unvoiced classification of speech using a neural network trained with LPC coefficients*, in Proc. Int. Conf. Contr., Artif. Intell., Robot. Opt. (Prague, Czech Republic), May 2017, pp. 56–59.
21. S. S. Park, J. W. Shin, and N. S. Kim, *Automatic speech segmentation with multiple statistical models*, in Proc. INTERSPEECH 2006 - ICSLP (Pittsburgh, PA, USA), 2017, pp. 2066–2069.
22. S. Bhattacharya, S. K. Singh, and T. Abhinav, *Performance evaluation of lpc and cepstral speech coder in simulation and in real time*, in Proc. Int. Conf. Recent Adv. Inf. Technol. (Dhanbad, India), Mar. 2012, pp. 826–831.
23. S. Kumar, *Performance evaluation of a novel AMDF-based pitch detection scheme*, ETRI J. **38** (2016), no. 3, 425–434.
24. C. Yeh and C. Zhuo, *An efficient complexity reduction algorithm for G.729 speech codec*, Comput. Math. Applicat. **64** (2012), no. 5, 887–896.
25. G. Pirker et al., *A pitch tracking corpus with evaluation on multi-pitch tracking scenario*, in Proc. Interspeech – Int. Conf. Spoken Language Process. (Florence, Italy), 2011, pp. 1509–1512.
26. Y. Hu and P. Loizou, *Subjective evaluation and comparision of speech enhancement algorithms*, Speech Commun. **49** (2007), no. 7–8, 588–601.
27. J. R. Deller, J. H. L. Hansen, and J. G. Proakis, Discrete-time processing of speech signal, Wiley, Piscataway, NJ, USA, 2000, pp. 570–579.
28. ITU-T P.862, *Perceptual evaluation of speech quality (PESQ)*, 2004.
29. S. Kumar, S. Bhattacharya, and P. Patel, *A new pitch detection scheme based on ACF and AMDF*, in Proc. IEEE Int. Conf. Adv. Commun., Contr. Comput. Technol. (Ramanathapuram, India), 2014, pp. 1235–1240.
30. S. Kadambe, G. F. Boudreaux-Bartels, *Application of the wavelet transform for pitch detection of speech signals*, IEEE Trans. Inf. Theory **38** (1992), no. 2, 917–924.

AUTHOR BIOGRAPHY

Sandeep Kumar received his BTech degree in Electronics and Instrumentation Engineering from the Institute of Engineering and Technology, MJP Rohilkhand University, Bareilly, India, in 2006 and his MTech and PhD degrees in Electronics and Communication Engineering from the Indian Institute of Technology (ISM), Dhanbad, India, in 2008 and 2015, respectively. He is currently working as an assistant professor in the Department of Electronics and Communication Engineering, National Institute of Technology, Delhi and having more than nine years of teaching and research experience. His research interests include digital signal processing and its application in speech, audio, image, optical, and bio-signal processing.