INTELLIGENT AUTOMATION & SOFT COMPUTING, 2017
http://dx.doi.org/10.1080/10798587.2017.1278961

Soft Computing Techniques for Classification of Voiced/Unvoiced Phonemes

Mohammed Algabri[a,c], Mohamed Abdelkader Bencherif[c], Mansour Alsulaiman[b,c], Ghulam Muhammad[b] and Mohamed Amine Mekhtiche[c]

[a]Computer Science Department, King Saud University, Riyadh, Saudi Arabia; [b]Computer Engineering Department, King Saud University, Riyadh, Saudi Arabia; [c]Center of Smart Robotics Research (CS2R), King Saud University, Riyadh, Saudi Arabia

ABSTRACT

A method that uses fuzzy logic to classify two simple speech features for the automatic classification of voiced and unvoiced phonemes is proposed. In addition, two variants, in which soft computing techniques are used to enhance the performance of fuzzy logic by tuning the parameters of the membership functions, are also presented. The three methods, manually constructed fuzzy logic (VUFL), fuzzy logic optimized with genetic algorithm (VUFL-GA), and fuzzy logic with optimized particle swarm optimization (VUFL-PSO), are implemented and then evaluated using the TIMIT speech corpus. Performance is evaluated using the TIMIT database in both clean and noisy environments. Four different noise types from the AURORA database—babble, white, restaurant, and car noise—at six different signal-to-noise ratios (SNRs) are used. In all cases, the optimized fuzzy logic methods (VUFL-GA and VUFL-PSO) outperformed manual fuzzy logic (VUFL). The proposed method and variants are suitable for applications featuring the presence of highly noisy environments. In addition, classification accuracy by gender is also studied.

KEYWORDS

Voiced and Unvoiced Detection; Fuzzy Logic; Genetic Algorithm; PSO; Optimization

1. Introduction

The classification of speech into voiced and unvoiced phonemes is a significant pre-processing step in many speech applications, such as speech segmentation, speech recognition, reconstruction, and de-noising (Beritelli, Casale, Russo, & Serrano, 2009; Fisher, Tabrikian, & Dubnov, 2006). Conventional models of speech are based on the voiced and unvoiced characteristic, as it is easily discernable in speech (Narayanan, Zhao, Wang, & Fosler-Lussier, 2011). Speech is composed of phonemes, which are produced by the vocal cords. Voiced sounds, such as "z" or "g" consist of periodic, oscillatory signals due to the vibration of the vocal folds. Unvoiced sounds, such as "k" or "q" are non-periodic and more noise-like, caused by air passing through some constriction in the vocal folds.

A variety of techniques for voiced/unvoiced classification has been reported in the literature. Wavelet transform has been used for feature extraction in speech-recognition applications, proving to be an effective technique for unvoiced phoneme classification (Sahu, Biswas, Bhowmick, & Chandra, 2014). Kumar, Hussain, and Kanhangad (2015) proposed a novel approach based on empirical wavelet transform and multilevel local patterns (MLP) for classification of voiced/nonvoiced speech signals. In their proposed approach, MLP and a modified version of 1D-local binary patterns (LBP) are used as features, and nearest neighbor (1-NN) is used as a classifier. They used the CMU Arctic database to evaluate the proposed method, with experiments conducted on 60 male speech signals and a set of 60 female speech signals. Zero-crossing rate and short-term energy have also been used to make voiced/unvoiced decisions in broad phoneme classification (Deekshitha & Mary, 2014), in which one broad phoneme symbol (Vowels, Nasals, Plosives, Fricatives, Approximants, and Silence) is assigned for each frame. Faycal and Bensebti (2014) conducted a comparative performance study of several features for voiced/nonvoiced classification of speech. They developed five classification schemes by combining one or two features from among the following: Energy (E), Zero-Crossing Rate (ZCR), Autocorrelation Function (ACF), Average Magnitude Difference Function (AMDF), Weighted ACF (WACF), and Discrete Wavelet Transform (DWT). The performance of their classifiers was evaluated on a subset of speech data extracted from the TIMIT Database. To validate the developed classifiers in noisy environments, they used two different noise types, white and babble, taken from the NOISEX92 database. For 26 speakers (13 females and 13 males), they achieved a performance of approximately 97%, but the classifiers significantly degraded in noise to 50%. Driaunys, Rudžionis, and Žvinys (2015) presented a classification approach comprising features and rules for the detection of phoneme groups using phonetically labeled data. Their classification approach produced an overall 3% improvement for phoneme recognition accuracy using the LTDIGITS corpora. The utterances of 100 speakers (50 males, 50 females) were selected for their experiments. Alam, Jassim, and Zilany (2014) proposed a method for phoneme classification under noisy conditions. In their proposed method, neurograms are constructed from the responses of a model auditory nerve to speech phonemes, which are then used as features to train a recognition system that utilizes a Gaussian Mixture Model (GMM). Using the TIMIT database, performance was evaluated for different types of phonemes, such as fricatives, stops, and vowels, in both clean and noisy environments. The results obtained suggest that their proposed method based on neural response is more robust to noise for phoneme classification.

CONTACT Mohammed Algabri malgabri@ksu.edu.sa

© 2017 TSI® Press

Beritelli et al. (2009) proposed an adaptive system for Voiced/Unvoiced (V/UV) speech detection in background noise, in which a genetic algorithm is used to select features that achieve the best V/UV detection. They implemented the system and performed tests using the TIMIT speech corpus and its phonetic classification. Four different types of noise and five different SNRs were used. Further, they used sentences uttered by two speakers, one male and one female, from each of eight different geographical areas (DR1-DR8), resulting in 16 different speakers in total being used for the training and test phases. Their experimental results showed that the adaptive V/UV classifier outperforms traditional solutions, giving an improvement of 25% in very noisy environments compared with non-adaptive classification and the V/UV detection system in the ETSI ES 202 212 v1.1.2 and with the speech classification in the Selectable Mode Vocoder (SMV) algorithm. Dhananjaya and Yegnanarayana (2010) proposed a new method for voiced/nonvoiced detection based on epoch extraction, in which instants of significant excitation (or epochs) are extracted using a zero-frequency-filtered speech signal. The performance of the proposed algorithm was evaluated on 40 speakers from the TIMIT and CMU ARCTIC databases, using two different noise types from the NOISEX database at five different SNRs.

All of the studies presented above use various techniques to detect voiced/unvoiced speech, but the performance of their classifiers significantly degrades in noise. Further, the classifiers were only applied to small groups of speakers. To the best of our knowledge, this present study is the first to use soft computing techniques for voiced/unvoiced classification with a large set of phonemes in highly noisy environments.

The remainder of this paper is organized as follows; Section 2 presents the proposed method and its variants. Section 3 develops the fuzzy logic for voiced/unvoiced classification. Section 4 outlines the enhancement of fuzzy logic using a genetic algorithm. Section 5 describes the enhancement of fuzzy logic using PSO. Section 6 evaluates the performance of the proposed methods under noisy conditions. Section 7 concludes this paper.

2. Proposed Method

Figure 1 shows a block diagram of the proposed voiced and unvoiced classification system. The input to the system is a speech file. In the feature extraction step, the ZCR and short-term energy are computed for each phoneme. These features are applied for each controller. We investigated three classifiers: A manually constructed fuzzy logic controller and two classifiers that each use a fuzzy logic controller tuned using soft computing techniques (GA and PSO).

2.1. Feature Extraction

One of the simplest methods used to perform voiced/unvoiced phoneme classification is based on zero-crossing rate (ZC) and short-term energy (STE).

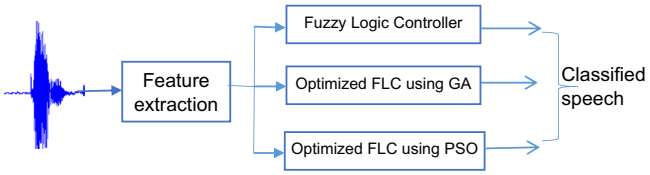

Figure 1. Block Diagram of the Proposed Methods.

2.1.1. Zero-crossing rate (ZC)

ZC (Panagiotakis & Tziritas, 2005) is the rate of sign changes along a signal. The ZCR of a voiced signal is less than that of an unvoiced signal. The ZC for each frame is calculated using Equation (1):

$$ZC = \sum_{n=1}^{N} \left| sign(x_n) - sign(x_{n-1}) \right| \quad (1)$$

Where N represents the number of samples in a phoneme.

2.1.2. Short-term energy

Energy for voiced speech is significantly greater than that for unvoiced speech; hence, the short-term energy can be used to distinguish voiced and unvoiced speech. The long-term definition of signal energy is given by Equation (2):

$$E = \sum_{m=-\infty}^{\infty} x^2(m) w(n-m) \qquad n - N + 1 \le m \le n \quad (2)$$

Where w is a function, and n is the number of samples in a phoneme.

Rabiner and Schafer (2011) state that "voiced speech should be characterized by relatively high energy and relatively low ZCR, while unvoiced speech will have relatively high ZCR and relatively low energy." They also state that "we have not said what we mean by high and low values of short-term ZCR, and it is really not possible to be precise." We view this as a problem that can be appropriately solved using soft computing techniques as they deal with imprecision, partial truth, and uncertainty.

2.2. TIMIT Speech Corpus

The TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT—Texas Instruments [TI] and Massachusetts Institute of Technology [MIT]) was recorded to provide speech data for acoustic-phonetic studies and automatic speech-recognition systems. TIMIT contains recordings of 630 speakers of eight major dialects of American English. All sentences are manually segmented at the phoneme level. The phonemes are grouped into voiced and unvoiced groups (Beritelli et al., 2009; Huang, Acero, Hon, & Foreword By-Reddy, 2001), as shown in Table 1.

3. Fuzzy Logic for Voiced and Unvoiced Classification

In this section, the design of the fuzzy logic controller called VUFL for voice and unvoiced classification based on our study in Algabri et al. (2015) is presented. VUFL is used to classify speech based on ZCR and short-term energy. The ZCR and energy features are extracted for each phoneme and then used as input for VUFL. VUFL comprises three components; linguistic variables (inputs and outputs), membership functions,

Table 1. TIMIT Phoneme Classification.

Class	Phoneme
Voiced	b, d, g, dx, jh, z, zh, v, dh, m, n, ng, em, en, eng, nx, l, r, w, y, hv, el, iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, ax, ix, axr.
Unvoiced	P, t, k, q, ch, s, sh, f, th, hh, kcl, tcl, gcl, epi, dcl, ax-h.

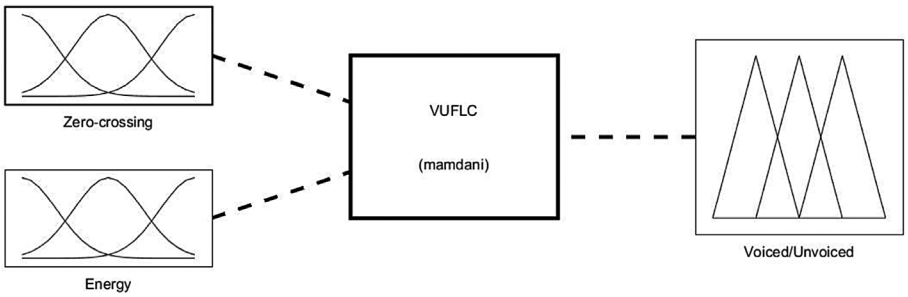

Figure 2. VUFLC Components.

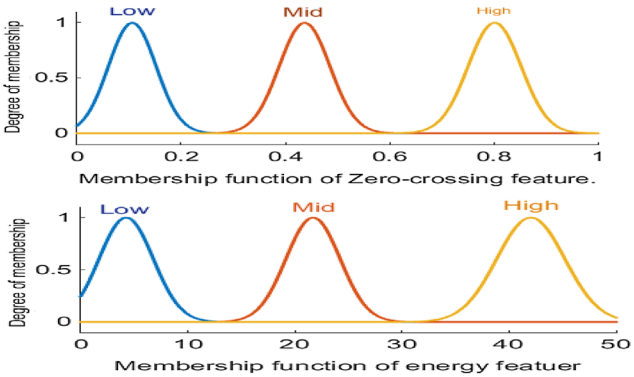

Figure 3. Input Membership Functions of VUFLC.

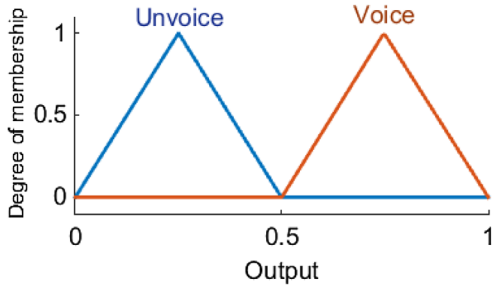

Figure 4. Output Membership Functions of VUFLC.

and fuzzy rules. We used a MATLAB Fuzzy Logic Toolbox to build the classifier, as shown in Figure 2. In our design, the controller has two inputs; zero-crossing rate (ZC) and short-term energy (STE). The output of the controller is classified speech.

The **membership function** is a graphical representation of the values of inputs and outputs. In this work, three "Gaussian" membership functions are used for each input and output. The range distribution of each input is divided into three linguistic variables: Low, Mid (Medium), and High. Figure 3 shows the membership functions of the VUFL input variables.

The range of the output of this classifier is divided into two linguistic variables, namely voiced (V) and Unvoiced (U). Figure 4 illustrates the output membership functions of VUFL.

Fuzzy rules use a simple IF-THEN rule base to control the output of the controller. The fuzzy rules used in this controller for the three membership functions are listed in Table 2.

To build the rule base, we used the IF-THEN rules presented in Algabri et al. (2015), Caruntu, Nica, Toderean, Puşchită, & Buza (2006), and Caruntu, Toderean, & Nica (2005), after testing and verifying each rule manually.

Table 2. Fuzzy Rules of VUFL.

Input		
Zero-crossing	Energy	Output
High	Mid	Unvoiced
Mid	High	Voiced
Low	High	Voiced
High	High	Unvoiced
Mid	Mid	Voiced
Low	Low	Unvoiced
Low	Mid	Voiced
Mid	Low	Unvoiced
High	Low	Unvoiced

4. Optimized Fuzzy Logic using a Genetic Algorithm (VUFL-GA)

The performance of fuzzy logic depends on the membership functions and fuzzy rules (Herrera, Lozano, & Verdegay, 1995). The process of tuning the membership function parameters is difficult and time consuming. In this section, we outline how the membership function parameters are automatically tuned using a genetic algorithm to overcome this difficulty. GAs are evolutionary algorithms that use biological evaluation to solve optimization problems. The general idea underlying GAs is first representation of each candidate solution of the problem as a chromosome. Then, crossover and mutation operators are applied to generate new solutions. Solutions are then selected according to a fitness function. Figure 5 shows a block diagram of the proposed VUFL-GA system. In the initial step, the first generation of membership functions is generated randomly. Because we use Gaussian membership functions, each is represented using two variables: the center (**c**) and width (σ) in a chromosome.

Each chromosome represents one candidate solution and a cost function is computed for each solution using Equation (3):

$$Cost = 100 \times \left[1 - \frac{1}{n \times m \times w \times p} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{w} \sum_{l=1}^{p} Class(i,j,k,l)\right] \quad (3)$$

Where n is population size, m is number of speakers, w is number of wav files for each speaker, and p is the number of phonemes in each wav file. $Class(i, j, k, l) = 1$ if the system is classified correctly and is equal to zero otherwise. This cost function should be minimized.

In the reproduction step, crossover and mutation are applied on the parents, selected using the roulette wheel method (Lipowski & Lipowska, 2012), to generate children. The new generation is selected based on the best cost of parents and children. Termination occurs when the system has reached its maximum number of generations.

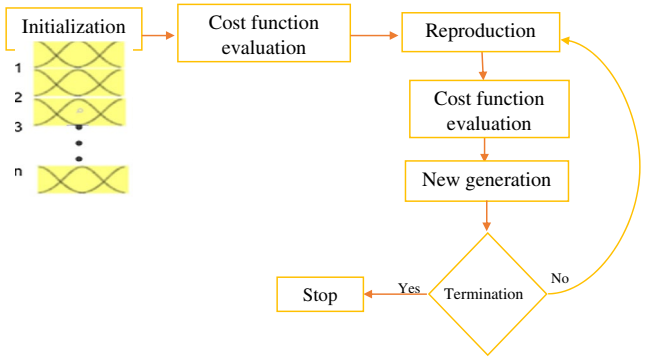

Figure 5. Block Diagram Showing the Typical Operation of a Genetic Algorithm.

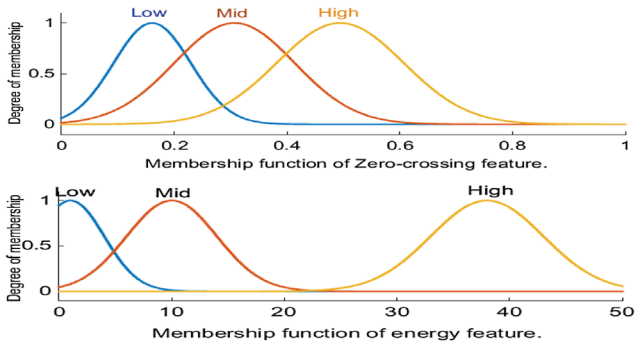

Figure 6. Membership Functions for VUFL-GA after Tuning using GA.

Table 3. GA Parameter Ranges.

Parameters	Range
Number of generations	20–160
Population size	20–160
Crossover rate	60–90
Mutation rate	1–20

Table 4. Parameters Selected for GA.

Exp. ID	No. of generations	Popula-tion size	Crossover rate	Mutation rate	Cost
1	56	59	0.69	0.12	9.880
2	118	96	0.90	0.15	7.063
3	101	111	0.88	0.10	8.492
4	132	92	0.92	0.11	8.492
5	90	84	0.73	0.15	8.530
6	65	87	0.84	0.20	9.880
7	135	111	0.72	0.50	9.800
8	94	109	0.93	0.60	9.840
9	68	52	0.92	0.10	12.620
10	95	52	0.83	0.90	9.880

4.1. Selecting Parameters

Selection of parameters is a critical step in the application of genetic algorithms (Gates, Merkle, Lamont, & Pachter, 1995). In this section, the experimental results obtained using GA are applied to choose the optimal number of generations, population size, crossover rate, and mutation rate. Several parameter sets were selected randomly from the ranges given in Table 3. The range of each GA parameter was selected based on previous work (Gates et al., 1995).

Table 4 shows that the best result, cost of 7%, was obtained in experiment number 2 using parameter set (118, 96, 0.90, and 0.15). In the final step, we generated a fuzzy logic controller from the best solution obtained using the genetic algorithm. Figure 6 shows the tuning membership functions for VUFL-GA.

5. Fuzzy Logic Optimization using Particle Swarm Optimization (VUFL-PSO)

This section presents the optimization of the fuzzy logic controller for voiced and unvoiced classification using PSO (VUFL-PSO). The proposed method optimizes the controller by tuning the parameters of the membership function. PSO is a population-based optimization technique developed by Kennedy and Eberhart (1995). It is inspired from the social behavior of bird flocks or fish schools. In PSO, each candidate solution is modeled by particles flying around the search space.

Table 5. Parameters Selected for PSO.

Exp. Id	# of iteration	Swarm size	c_1	c_2	Cost
1	77	20	0.8	0.4	9.840
2	42	66	0.2	1.6	8.490
3	78	52	2.0	0.6	7.060
4	109	76	1.2	2.0	4.285
5	**124**	**66**	**1.8**	**1.0**	**2.817**
6	119	46	1.4	1.4	7.100
7	28	95	0.4	1.8	5.640
8	59	91	1.6	0.2	7.100
9	142	80	0.6	1.2	7.060
10	87	76	0.6	2.0	15.510

Note: Bold values represent the parameters of the best cost.

While searching, each particle updates its velocity and position using Equations 4 and 5 (Algabri, Mathkour, Ramdane, & Alsulaiman, 2015):

$$v(t+1) = w \cdot v(t) + c_1 \cdot rand_1 \cdot (pbest_i - x(t)) + c_2 \cdot rand_2 \cdot (gbest - x(t)) \quad (4)$$

$$x(t+1) = x(t) + v(t+1) \quad (5)$$

Where $v(t)$ is the current velocity, $v(t+1)$ is the updated velocity, c_1 and c_2 are acceleration coefficients, w is the inertial weight, $pbest$is the personal best fitness of particle i, $gbest$ is the global best fitness among all the particles, $x(t+1)$ is the updated position of the particle, $x(t)$ is the current position of the particle, and $v(t+1)$ is the updated velocity, from Equation (3), of the particle.

As with the GA, we conducted several experiments using PSO to choose the optimal number of iterations, swarm size,c_1, andc_2. Parameter sets were selected randomly, as shown in Table 5.

Table 5 shows that the result with the best cost, 2.8%, was obtained in experiment number 5 using parameter set (124, 66, 1.8, and 1) for number of iterations, swarm size,c_1, andc_2, respectively. Finally, we generated a fuzzy logic controller from the best solution obtained by PSO. Figure 7 shows the tuning membership functions for VUFL-PSO.

6. Performance Evaluation

We evaluated the performance of the proposed methods (VUFL, VUFL-GA, and VUFL-PSO) experimentally. All 630 speakers (192 females and 438 males) in the TIMIT Corpus were used in the evaluation. The overall number of phonemes in this sample is more than 43,000. We divided the entire database into train and test for each male and female set, as shown in Table 6. The training data were used to optimize the fuzzy logic via genetic algorithm and PSO.

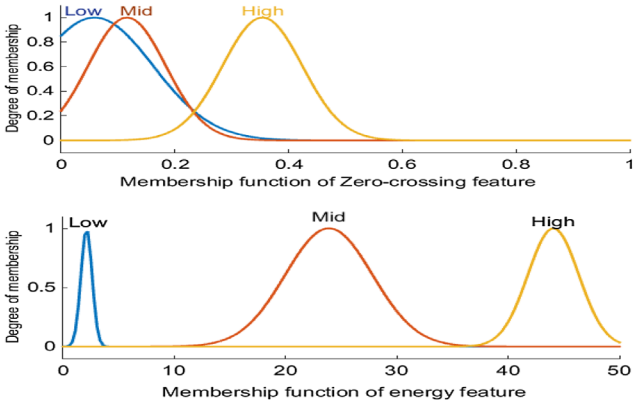

Figure 7. Membership Functions for VUFLC-PSO after Tuning using PSO.

Table 6. Training and Testing Datasets.

Datasets	Number of speakers
Train (Male)	326
Test (Male)	112
Train (Female)	136
Test (Female)	56
Total	630

Table 7. Accuracy (%) Comparison between Proposed Methods.

	SNR	FLC	VUFL-GA	VUFL-PSO
babble noise	Clean	80.10	89.04	88.48
	30 dB	80.10	87.75	87.99
	20 dB	79.84	80.89	84.02
	10 dB	73.47	75.42	78.30
	5 dB	71.12	74.59	75.46
	0 dB	70.68	73.73	72.73
	−5 dB	70.64	72.40	71.04
white noise	30 dB	80.10	87.78	87.72
	20 dB	79.99	80.73	83.98
	10 dB	73.25	75.57	78.64
	5 dB	70.97	74.77	75.49
	0 dB	70.65	74.03	72.60
	−5 dB	70.65	72.40	70.77
restaurant noise	30 dB	80.10	87.84	87.73
	20 dB	79.85	81.29	84.45
	10 dB	74.05	75.69	78.56
	5 dB	71.28	74.73	75.96
	0 dB	70.69	74.00	73.03
	−5 dB	70.65	72.78	71.16
car noise	30 dB	80.10	87.79	87.81
	20 dB	79.87	80.94	84.00
	10 dB	73.46	75.66	79.11
	5 dB	71.04	74.88	76.09
	0 dB	70.64	74.16	72.83
	−5 dB	70.65	72.33	70.87

The performances were compared under four noise types (babble, white, restaurant, and car) in addition to clean speech. Six levels of SNR, shown in Table 7, were used in the experiments. These noise types were taken from the AURORA noisy speech evaluation (Zhu, Iseli, Cui, & Alwan, 2001). Analysis of these results showed that improved performance was obtained from the fuzzy logic classifier enhanced using GA and PSO (VUFL-GA, and VUFL-PSO).

Table 7 gives results in terms of the average accuracy of voiced and unvoiced classification using the TIMIT Corpus. In all the experiments, the tuned fuzzy logic controllers using soft computing (VUFL-GA and VUFL-PSO) obtained better results than the FLC without tuning. All classifiers performed

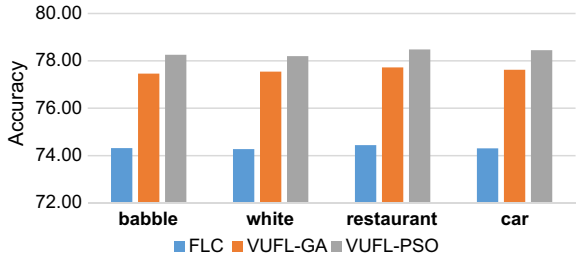

Figure 8. Average Classification Accuracy in Noisy Environments.

well in the clean environment, with decreasing performance when noise was added to the speech signal. The VUFL-GA had the best performance in white and restaurant noise. Conversely, VUFL-PSO had the best performance in babble and car noise with high SNRs (30 dB, 20 dB, and 10 dB). Our proposed methods proved more robust than other V/UV classification systems presented in (Faycal & Bensebti, 2014) in very noisy environments (0 dB and -5 dB). The performance of the method in (Faycal & Bensebti, 2014) was 51.83 and 69.95 at low SNRs under babble and white noise, respectively.

The average values for classification accuracy at different SNRs (30, 20, 10, 5, 0, and -5 dB) are shown in Figure 8 for different types of noise. It can be seen that VUFL-GA and VUFL-PSO are more robust than VUFL in very noisy environments. Finally, we compared the performance of our classifiers with that of the classifier proposed by Dhananjaya and Yegnanarayana (2010). In white noise, their classifier achieved voiced/unvoiced detection accuracy in the range [94.2–85.7] with different SNRs. In contrast, our method achieved performance in the range [87.78–72.40]. The higher performance by Dhananjaya and Yegnanarayana's classifier may be attributed to the fact that they used only a small subset of TIMIT, comprising 38 speakers (24 males and 14 females), whereas we used all 630 speakers from TIMIT (438 males and 192 females).

6.1. Study of Classification Accuracy by Gender

In this section, we look at the accuracy of classification by gender (male and female) using SAS 9.2. First, we calculated the mean classification accuracy for males and females separately for the VUFL-GA method. Then, these data were analyzed with SAS using a Two Sample t-test, as shown below.

Two Sample t-tests for accuracy by gender

```
Sample Statistics
Group          N      Mean       Std. Dev.   Std. Error
-----------------------------------------------------------
F            192    88.92189      4.0239      0.2904
M            438    88.84331      4.3235      0.2066
Hypothesis Test
Null hypothesis:    Mean 1 - Mean 2  = 0
Alternative:        Mean 1 - Mean 2 ^= 0

If Variances Are    t statistic      Df       Pr > t
-----------------------------------------------------------
Equal                   0.214        628       0.8303
Not Equal               0.221      389.62      0.8256
95% Confidence Interval for the Difference between Two Means
Lower Limit    Upper Limit
-----------    -----------
-0.64           0.80
```

The mean of classification accuracy for male speakers was 88.9%, whereas that for female speakers was 88.8%. We are 95% confident that (μ male—μ female$\in$ (-0.64, 0.80)). Because $p = 0.8303 > 0.05$, we cannot reject the null hypothesis

H_0: μ female = μmale. No significant differences were found in the means for classification accuracy by gender.

7. Conclusion

In this paper, an automatic voiced and unvoiced classification system based on zero-crossing rate and short-term energy was proposed. In the proposed system, a fuzzy logic-controller called VUFL is used for classification, with its performance depending on the membership function parameters. Two soft computing techniques were also proposed to enhance the performance of fuzzy logic and automatic tuning of membership function parameters conducted using a genetic algorithm (GA) and particle swarm optimization (PSO). The findings of the speech classification study on voiced and unvoiced speech can be summarized as follows:

- Fuzzy logic is a simple and good classification method.
- Soft computing, either GA or PSO enhances the performance of fuzzy logic.
- The system is robust and suitable for application in noisy environments; that is, in environments with SNRs lower than 5 dB.
- There are no significant differences in the mean classification accuracy for male and female speakers.

TIMIT Corpus was used in these experiments. The highest accuracy, 89.3%, was obtained using VUFL-GA.

Acknowledgment

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-MED2474-02).

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Mohammed Algabri is a PhD student in the Computer Science Department, College of Computer and Information Science, King Saud University. He received his Master's degree from the Department of Computer Science in 2015 at King Saud University and a BSc degree in Computer Science from Umm Al-Qura University. His research areas include soft computing techniques, speech recognition, robotics, and pattern recognition.

Dr. Mohamed Abdelkader Bencherif obtained an Engineer Degree in Control from INELEC, Boumerdes, Algeria in 1992. He worked in diverse industrial projects, mainly on project management. He completed his master's in Signals and Systems in 2005, and obtained his Ph.D. in the Classification of Remote Sensing Images in 2015 from Saad Dahleb University, Blida, Algeria. He is actually working at the Center of Smart Robotics Research, King Saud University. His areas of interest are robotic design, speech classification, and pattern recognition.

Professor Mansour Alsulaiman obtained his PhD degree from Iowa State University, USA in 1987. Since 1988, he is with the Department of Computer Engineering, College of Computer and Information Sciences. His research areas include automatic speech/speaker recognition, automatic voice pathology assessment systems, computer-aided pronunciation training system, and robotics. He has authored and co- authored more than 80 publications in journals and conferences. He was the editor of Journal of King Saud University- Computer Science from 2008 till 2015. He is the director of Center of Smart Robotics Research (CS2R).

Dr. Ghulam Muhammad is an associate professor in Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He received his Ph.D. degree in 2006 from the Department of Electronic and Information Engineering at Toyohashi University of Technology, Japan. He has authored and co-authored more than 140 publications in journals and conferences. His research interest includes speech and image processing. He owns a US patent.

Mohamed Amine Mekhtiche is a researcher at Center of Smart Robotic Research in Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He received his Master's degree in 2012 from the Department of Electronic at Saad Dahleb University, Algeria. He has eight publications in conferences and journals.

References

Alam, M.S., Jassim, W., & Zilany, M.S. (2014). *Neural response based phoneme classification under noisy condition*. Paper presented at the 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS).

Algabri, M., Alsulaiman, M., Muhammad, G., Zakariah, M., Bencherif, M., & Ali, Z. (2015). *Voice and Unvoiced Classification Using Fuzzy Logic*. Paper presented at the Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), Las Vegas, Nevada, USA.

Algabri, M., Mathkour, H., Ramdane, H., & Alsulaiman, M. (2015). Comparative study of soft computing techniques for mobile robot navigation in an unknown environment. *Computers in Human Behavior, 50*, 42–56.

Beritelli, F., Casale, S., Russo, A., & Serrano, S. (2009). Adaptive V/UV speech detection based on characterization of background noise. *EURASIP Journal on Audio, Speech, and Music Processing, 2009*, 11.

Caruntu, A., Nica, A., Toderean, G., Puşchită, E., & Buza, O. (2006). *An improved method for automatic classification of speech*. Paper presented at the 2006 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca, Romania.

Caruntu, A., Toderean, G., & Nica, A. (2005). *Automatic silence/unvoiced/voiced classification of speech using a modified Teager energy feature*. Paper presented at the Proceedings of the 2005 WSEAS international conference on Dynamical systems and control, Venice, Italy.

Deekshitha, G., & Mary, L. (2014). Broad phoneme classification using signal based features. *International Journal on Soft Computing (IJSC), 5*, 1–11.

Dhananjaya, N., & Yegnanarayana, B. (2010). Voiced/nonvoiced detection based on robustness of voiced epochs. *Signal Processing Letters, IEEE, 17*, 273–276.

Driaunys, K., Rudžionis, V., & Žvinys, P. (2015). Implementation of hierarchical phoneme classification approach on LTDIGITS corpora. *Information Technology and Control, 38*(4), 1.

Eberhart, R.C., & Kennedy, J. (1995). *A new optimizer using particle swarm theory*. Paper presented at the Proceedings of the sixth international symposium on Micro Machine and Human Science, Nagoya, Japan.

Faycal, Y., & Bensebti, M. (2014). Comparative performance study of several features for voiced/non-voiced classification. *International Arab Journal of Information Technology, 11*, 293–299.

Fisher, E., Tabrikian, J., & Dubnov, S. (2006). Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model. *Audio, Speech, and Language Processing, IEEE Transactions on, 14*, 502–510.

Gates Jr, G.H., Merkle, L.D., Lamont, G.B., & Pachter, R. (1995). *Simple genetic algorithm parameter selection for protein structure prediction*. Paper presented at the IEEE International Conference on Evolutionary Computation, 1995, Perth, Australia.

Herrera, F., Lozano, M., & Verdegay, J.L. (1995). Tuning fuzzy logic controllers by genetic algorithms. *International Journal of Approximate Reasoning, 12*, 299–315.

Huang, X., Acero, A., Hon, H.-W., & Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Upper Saddle River, NJ: Prentice Hall PTR.

Kumar, T.S., Hussain, M., & Kanhangad, V. (2015). *Classification of voiced and non-voiced speech signals using empirical wavelet transform and multi-level local patterns*. Paper presented at the 2015 IEEE International Conference on Digital Signal Processing (DSP).

Lipowski, A., & Lipowska, D. (2012). Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications, 391*, 2193–2196.

Narayanan, A., Zhao, X., Wang, D., & Fosler-Lussier, E. (2011). *Robust speech recognition using multiple prior models for speech reconstruction*. Paper presented at the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic.

Panagiotakis, C., & Tziritas, G. (2005). A speech/music discriminator based on RMS and zero-crossings. *Multimedia, IEEE Transactions on, 7*, 155–166.

Rabiner, L., & Schafer, R. (2011). *Theory and applications of digital speech processing*. Upper Saddle River, NJ: Pearson Education.

Sahu, P., Biswas, A., Bhowmick, A., & Chandra, M. (2014). Auditory ERB like admissible wavelet packet features for TIMIT phoneme recognition. *Engineering Science and Technology, an International Journal, 17*, 145–151.

Zhu, Q., Iseli, M., Cui, X., Alwan, A. (2001). *Noise robust feature extraction for ASR using the Aurora 2 database*. Paper presented at the 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark.