IJICTR International Journal of Information & Communication Technology Research

Volume 8- Number 2- Spring 2016 (33-44)

Hybrid of Evolutionary and Swarm Intelligence Algorithms for Prosody Modeling in Natural Speech Synthesis

Mansour Sheikhan
Electrical Engineering Department
Islamic Azad University-South Tehran Branch
Tehran, Iran
msheikhn@azad.ac.ir

Received: June 4, 2014- Accepted: February 24, 2016

Abstract— **To reduce the number of input features to a prosody generator in natural speech synthesis application, a hybrid of an evolutionary algorithm and a swarm intelligence-based algorithm is used for feature selection (FS) in this study. The input features to FS unit are word-level and syllable-level linguistic features. The word-level features include punctuation information, part-of-speech tags, semantic indicators, and length of the words. The syllable-level features include the phonemic structure and position indicator of the current syllable in a word. A modified Elman-type dynamic neural network (DNN) is used for prosody generation in this study. The output layer of this DNN provides prosody information at the syllable-level including pitch contour, log-energy level, duration information, and pause data. Simulation results show that the prosody information is predicted with an acceptable error by this hybrid soft-computing method as compared to Elman-type neural network prosody generator and binary gravitational search algorithm-based FS unit.**

Keywords- **speech synthesis; genetic algorithm; ant colony optimization; neural network; prosody.**

I. INTRODUCTION

Generating accurate prosody information is critical in improving the naturalness and intelligibility of synthetic speech [1, 2]. Generally, the prosody information synthesizers provide pitch frequency (F_0) contours, energy levels, word durations, and inter-word pause durations. In this way, both low-level lexical features (such as the phonetic structures) [3] and high-level features (such as the syntactic information) [4] have been used.

The prediction of prosody aids syntactic/semantic parsing [5], speech recognition [6], diagnosis of speech and language disorders [7], understanding of situational context [8], emotional processing [9], speech to speech translation [10, 11], and automatic dialogue systems [12].

The methods for prosody generation can be classified as follows: a) rule-based [13]; b) statistical models [6]; c) neural networks [14]; and d) hybrid models [15]. The rule-inference and manually exploring the effect of mutual interactions among linguistic features are the hard and complex processes in a rule-based method. On the other hand, the phonological rules are automatically determined using the training data without the help of a linguistic expert in statistical models or neural network-based models.

In this paper, a modified Elman-type dynamic neural network (DNN) is used for prosody generation in a text to speech (TTS) system. The context units are also fed from the output layer in this modification like Jordan networks. This network can maintain a sort of states and is expected to perform better in prediction tasks as compared to Elman and Jordan simple dynamic networks and standard multi-layer perceptron. It is noted that the context layer in DNNs is able to cope with historical data. So, DNNs give better solution than static feed-forward networks. The performance of this modified Elman-type DNN is compared with standard Elman-type network in Section 6.

The inputs of this DNN are both word-level and syllable-level linguistic features. To reduce the number of inputs to the DNN, a hybrid of an evolutionary algorithm (i.e., genetic algorithm (GA)) and a swarm intelligence (SI)-based algorithm (i.e., ant colony optimization (ACO)) [16] is used for feature selection (FS). The F_0 contour and log-energy level of a syllable, the duration of a syllable and its constituent vowel, the vowel onset time in a syllable, and the inter-syllable pause duration are the prosody information generated by the proposed model. It is noted that inspection on the pronunciation process of humans and consistency analysis on acoustic modules in TTS systems show that the consistency can be interpreted as a high correlation of a warping curve between the spectrum and the prosody intra a syllable [17]. As compared to the previous models proposed by the author for Farsi in [18, 19], using an intelligent hybrid feature selection method as compared to [18] and employing semantic-related features and also more contextual information (to enrich the word-level and syllable-level input features) as compared to [19] are the contributions of this study to provide a light neural model for prosody generation.

The rest of paper is organized as follows: Related work is reviewed in Section 2. The schematic of proposed system is introduced in Section 3. The GA-ACO hybrid FS method is detailed in Section 4. The word-level and syllable-level input features are introduced in Section 5. The simulation and experimental results are given in Section 6. Finally, Section 7 concludes the paper and points to possible future works.

II. Related Work

In recent decades, several prosody generation models were proposed for different languages such as English [20], French [21], Chinese [22], Japanese [23], German [24], Arabic [25], and Farsi [26]. In addition, many studies were conducted on deriving prosody generation model of a spoken language by realization of F_0 contours [27-29], energy levels [30], duration of segments [31, 32], and pause data [33, 34]. However, an integrated approach is employed in this study to generate simultaneously the prosody information including intonation, energy, duration, and pause at the syllable level.

As sample researches in this field, Xu and Prom-on [28] developed a trainable (yet deterministic) prosody synthesizer based on an articulatory-functional view of speech. Van Niekerk and Barnard [29] proposed a two-stage feed-forward neural network-based method for modeling F_0 values of a sequence of syllables. They considered linguistic constraints (represented by positional, contextual, and phonological features), production constraints (represented by articulatory features), and linguistic relevance tilt parameters for predicting intonation patterns. Winters and O'Brien [32] determined the relative contributions of suprasegmental and segmental features to the perception of foreign accent and intelligibility in German and English speech. The suprasegmental and segmental features were manipulated independently by transferring the following prosody information: a) native intonation contours and/or syllable durations onto non-native segments; and b) non-native intonation contours and/or syllable durations onto native segments in both languages.

As a syllable-based TTS system, Narendra and Sreenivasa Rao [35] tuned the weights of unit selection cost functions in a syllable-based system by using GA. Huang and Joo Er [36] presented an adaptive neuro-fuzzy controller to reproduce smooth vocal tract trajectories for high quality speech synthesis. Tomaschek *et al.* [37] showed that the perception of vowel length by listeners exhibits the characteristics of categorical perception.

On the other hand, feature reduction techniques have been widely used in data mining, pattern recognition, and forecasting problems to lower the number of dimensions describing the data. The aim of feature reduction is to reduce the computational cost of a system without deteriorating its discriminative capability. Avoiding over-fitting, resisting noise, and strengthening the prediction performance are the most advantages of feature reduction when using learning algorithms.

The feature reduction algorithms are broadly classified into two categories: a) feature transform (or feature extraction); and b) feature selection. Some new features are constructed by projecting an original feature space to a lower-dimension one in the feature transform techniques. On the other hand, a subset of an original feature space is chosen in the FS method according to its discrimination capability to improve the quality of data.

Principal component analysis and independent component analysis are two known feature transform methods [38]. There are three known FS methods: a) filter; b) wrapper; and c) embedded methods. Filter approaches are independent of learning algorithms and are based on inter-class separability criterion for FS. In a wrapper method, the evaluation procedure is tied to the task of learning algorithm (e.g., classification or prediction). In this method, the feature subset space is searched using the estimated accuracy from an induction algorithm as a measure of subset suitability. In the embedded method, the FS and learning algorithm are interleaved [39]. In addition, evolutionary and SI-based algorithms have been employed for FS (Table 1).

Table 1. Some evolutionary and SI-based algorithms for FS

FS algorithm	Research group
GA	De Stefano *et al*. [40]
Particle swarm optimization (PSO)	Xue *et al*. [41]
Differential evolution (DE)	Al-Ani *et al*. [42]
ACO	Chen *et al*. [43]
Bee colony optimization (BCO)	Ghareh Mohammadi and Saniee Abadeh [44]
Gravitational search algorithm (GSA)	Mohseni Bababdani and Mousavi [45]
Binary bat algorithm (BBA)	Nakamura *et al*. [46]
Gene ontology	Gillies *et al*. [47]
Wrapper algorithm based on GA	Lin *et al*. [48]
Hybrid of GA and ACO	Sheikhan and Mohammadi [16]
Simulated annealing (SA) and GA	Gheyas and Smith [49]
Modified micro GA (MmGA)	Tan *et al*. [50]
Binary PSO (BPSO)	Sheikhan *et al*. [51]
Modified BPSO	Vieira *et al*. [52]
Dynamic PSO	Bae *et al*. [53]
Hybrid of PSO and rough sets	Inbarani *et al*. [54]
Self-adaptive DE	Ghosh *et al*. [55]

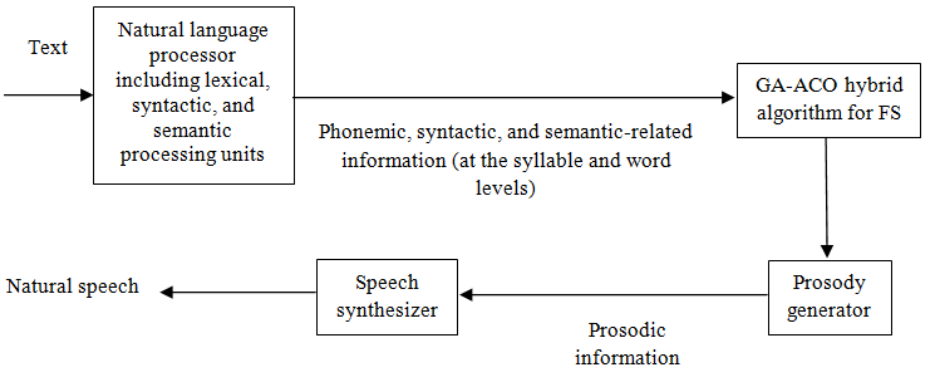

Fig. 1. Block diagram of a text-to-speech system including FS unit based on GA-ACO hybrid algorithm

III. SCHEMATIC OF PROPOSED SYSTEM

The block diagram of proposed method is shown in Fig. 1. As seen in Fig. 1, a natural language processor (NLP) unit is employed in this system consists of lexical processor, syntactic processor, and semantic processor [18, 56-60]. The NLP unit performs the tasks such as follows: a) converting the input text to a list of words; b) morphological analysis; c) POS tagging; and d) an abstract semantic processing to extract semantic-related information for the nouns, verbs, and adverbs. The semantic processor is implemented using a self-organizing map neural network as a semantic map [59]. This unit is also responsible for grapheme to phoneme transcription which is a difficult task in some languages such as Farsi and Arabic [61].

In this way, various syntactic analyzer core functions have been developed at the NLP unit to avoid the need for listing all of the possible derivatives that a word may have in Farsi (including verb segments, plural versus singular forms, definite versus indefinite nouns and adjectives, and comparative or superlative adjectives) [18]. Various types of sentences (such as predicative, interrogative, imperative, and exclamative) and phrases (like verbal,

substantial, and adverbial) have been analyzed and classified in this system. Top-down (context free) and bottom-up parsing methods have been used for parsing.

A set of 32 POS tags are considered in the proposed model including the following items: a) different types of verb (such as intransitive, active/passive transitive, auxiliary, imperative, and prohibitive); b) different types of adjective (such as indefinite, comparative, and superlative); c) different types of noun (such as plural/singular and indefinite plural/singular); d) different types of infinitive (such as simple, compound, and plural); e) different types of pronoun (such as discrete, connected, and demonstrative); f) adjective; g) interjection; h) indefinite; i) particle of interrogation; j) number; k) descriptive mood; l) adverb; m) preposition; n) conjunction; o) sign of direct object, p) object connected to a preposition; q) substantive; and r) noun governing the genitive [62]. It is noted that semantic prosody is the conotational coloring of the semantics of a word, largely uncaptured by dictionary definitions [63, 64]. As semantic-related information, three general

roles are considered for the nouns (i.e., humankind, animal, and object). Similarly, three roles for the adverbs (i.e., positive/negative-stressed such as "always/never", conditional such as "may be", and neutral), and three roles for the verbs (i.e., emotional, motional, and essential) are considered (as simple indications of semantic that can be considered as input features at the word-level) [59].

The hybrid of GA and ACO algorithm is used for FS in this study. Since GA is a domain independent search technique, it is used in this study where domain knowledge was difficult to provide [65]. In other words, GA is based on a global perspective by operating on the complete population from the very beginning [66]. On the other hand, ACO has the important advantage of local searching that does not exist in GA. So, GA and ACO are hybridized in this study to nullify their drawbacks. However, other evolutionary and SI-based algorithms can also be combined for this purpose (such as the algorithms listed in Table 1).

A modified Elman-type DNN with recurrent connections in hidden and output layers is used for prosody generation. This DNN provides a total of 9 output prosodic parameters for the current syllable as follows: a) four parameters to represent the pitch contour (using the first four coefficients of its discrete Legendre polynomial expansion [67]); b) one parameter for the log-energy level; c) one parameter for the syllable duration; d) one parameter for the vowel duration in syllable; e) one parameter for the vowel onset time in syllable; and f) one parameter for the inter-syllable pause duration. All the duration times are in milliseconds. It is noted that all 9 output prosody parameters are normalized to reduce the system complexity due to variations in these parameters caused by lexical phonetic features.

A harmonic plus noise model-version 1 (HNM1) is also used as a concatenative speech synthesizer in this system [18, 68]. To implement HNM speech synthesizer, the following tasks were performed [62]: a) estimation of pitch and maximum voiced frequency [69]; b) calculation of amplitude and phase harmonics [68]; c) using center of gravity algorithm to remove phase mismatches [70]; d) smoothing HNM parameters in concatenation points [71]; e) modification of prosody parameters [72]; f) determination of synthesis time instants [73]; g) estimation of amplitude and phase of harmonics in new pitch frequency [74]; and h) synthesis of speech segments [75].

It is noted that there are six simple vowels (V) and 23 consonants (C) in Farsi [76]. So, the database of speech synthesis in this study consisted of 299 units as follows: a) 23 consonants; b) 138 diphones of CV-type; and c) 138 diphones of VC-type. Equalization of discontinuities was performed at the concatenation point of two vowels to remove the phase, spectral envelope and pitch mismatches.

A piecewise linear curve was used to approximate pitch frequency variations like to MBROLA synthesizer [62]. The synthesizer also took the duration of phonemes in a syllable in form of a vector whose values were given in milliseconds. The intensity variations were also considered by allowing loudness modification through a set of input vectors to the synthesizer.

IV. Hybrid Feature Selection Method

In this section, the foundation of ACO-based feature selection is introduced and then the hybrid GA-ACO algorithm is presented.

A. Feature selection by ACO algorithm

Given the original set, F, of n features, it is desirable to find a subset S (which consists of m features; $m < n$), such that the prediction or classification accuracy is maximized. Assume the following items for feature selection by the ACO algorithm:

- n features that constitute the original set, $F = \{f_1, \ldots, f_n\}$;
- A number of artificial ants, n_a, to search through the feature space;
- τ_i, the intensity of pheromone trail associated with feature f_i, which reflects the previous knowledge about the importance of f_i;
- $\Delta\tau_i$, the amount of change of pheromone trial quantity for feature f_i; and
- For each ant j, a list that contains the selected feature subset, $S_j = \{s_1, \ldots, s_m\}$.

Each ant randomly chooses a feature subset of m features in the first iteration. Only the best k subsets, $k < n_a$, are used to update the pheromone trial and influence the feature subsets of the next iteration. In the following iterations, each ant will start with $m-p$ features that are randomly chosen from the previously selected k-best subsets, where $p \in [1, m-1]$. So, the features that constitute the best k subsets will have more chance to be present in the subsets of next iteration. For an ant j, the features that achieve the best compromise between pheromone trials and local importance with respect to S_j can be considered. The updated selection measure (USM) is used for this purpose which is defined as follows:

$$USM_i^{S_j}(t) = \begin{cases} \dfrac{(\tau_i)^{\alpha}\left(\eta_i^{S_j}\right)^{\beta}}{\sum_{g \notin S_j}(\tau_i)^{\alpha}\left(\eta_i^{S_j}\right)^{\beta}} & ; i \notin S_j \\ 0 & ;otherwise \end{cases} \quad (1)$$

where $\eta_i^{S_j}$ is the local importance of feature f_i given the subset S_j. The parameters α and β control the effect of pheromone trail intensity and local feature importance, respectively. The steps of this algorithm are given in Table 2 [77].

Table 2. Steps of an ACO-based algorithm for FS

Step	Title	Description
1	Initialization	Set $n_a, \tau_i, \Delta\tau_i, k, p$, and maximum number of iterations.
2	Random assignment of a subset of *m* features to S_j	Applicable only in the first iteration, then go to Step 4.
3	Selection of the remaining *p* features for each ant	Given subset S_j, choose feature f_i that maximizes $USM_i^{S_j}$ and set $S_j = S_j \cup \{f_i\}$.
4	Evaluation of the selected subset of each ant	Estimate root mean square error (RMSE) when using the features of S_j ($RMSE_j$) and sort the subsets according to their RMSE. Update the $RMSE_{min}$ and store the corresponding subset of features.
5	Pheromone trail intensity update and initialization of the subsets for next iteration using the feature subsets of the best *k* ants	$\Delta\tau_i = \begin{cases} \frac{\max_{g=1:k}(RMSE_g) - RMSE_j}{\max_{h=1:k}(\max_{g=1:k}(RMSE_g) - RMSE_h)}; & f_i \in S_j \\ 0; & otherwise \end{cases}$ $\tau_i = \rho\tau_i + \Delta\tau_i$ where ρ is a constant and $(1-\rho)$ represents the evaporation of pheromone trails.
6	Termination condition	If the maximum number of iterations or the desired RMSE is not achieved, go to Step 3.

The evaluation of the selected subset of features is performed by estimating the RMSE in predicting prosody parameters when using the features of S_j ($RMSE_j$) and sorting the subsets according to their RMSE (as seen in Table 2). In this way, the $RMSE_{min}$ is updated and the corresponding subset of features is stored.

B. Feature selection by GA-ACO algorithm

This algorithm begins by generating a population in GA and a number of ants. GA generates feature subsets and the resulting subsets are gathered and then evaluated at the end of iterations. The best subset is selected according to evaluation measures. If an optimal subset has been found or the algorithm has been executed a certain number of runs, then the process is stopped and the best feature subset is reported. The flowchart of GA-ACO algorithm is shown in Fig. 2 [16]. It is noted that another hybrid format of GA and ACO is proposed in [66] in which ACO and GA generate feature subsets in parallel.

V. WORD-LEVEL AND SYLLABLE-LEVEL INPUT FEATURES

The input word-level features to the system are as follows:

- POS tag of the current word ($POS(W_i)$), the POS tag of two previous words ($POS(W_{i-1})$ and $POS(W_{i-2})$), and the POS tag of two next words ($POS(W_{i+1})$ and $POS(W_{i+2})$);
- Semantic indicator of the current word ($SMNI(W_i)$), the semantic indicator of two previous words ($SMNI(W_{i-1})$ and $SMNI(W_{i-2})$), and the semantic indicator of two next words ($SMNI(W_{i+1})$ and $SMNI(W_{i+2})$);
- Length of the current word (in terms of syllable counts) ($L(W_i)$), the length of two previous words ($L(W_{i-1})$ and $L(W_{i-2})$), and the length of two next words ($L(W_{i+1})$ and $L(W_{i+2})$); and
- Type of the punctuation mark (PM) after the current word (including comma, point, question mark, sign of exclamation, and sign of quotation).

It is noted that for some words such as prepositions, semantic indicator is not considered. The word-level features are summarized in Table 3. The value "0" in the range of SMNI and PM features show that the semantic indicator is not applicable for that word and there is not punctuation mark after that word, respectively.

Before introducing syllable-level features, it is noted that there are 23 consonants and six simple vowels in Farsi [76]. Based on the articulation manner of consonants, the consonants are classified into six groups in this study (e.g., "p", "t", and "k" phonemes are considered as one group). In addition, the possible syllable structures in Farsi (based on consonant and vowel phonemes) are CV, CVC, and CVCC types.

The input syllable-level features to the system are considered as follows:

- First consonant of the current syllable ($FC(S_i)$), the first consonant of two previous syllables ($FC(S_{i-1})$ and $FC(S_{i-2})$), and the first consonant of two next syllables ($FC(S_{i+1})$ and $FC(S_{i+2})$);

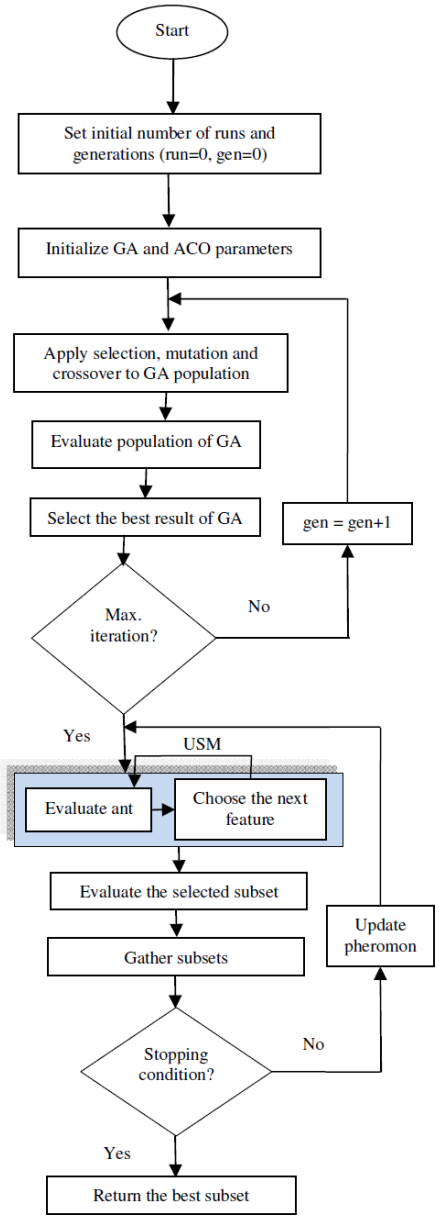

Fig. 2. Flowchart of GA-ACO hybrid feature selection algorithm [16]

Table 3. Word-level input features to FS unit

Feature	Number of features	Range of feature value
$POS(W_i)$, $POS(W_{i-1})$, $POS(W_{i-2})$, $POS(W_{i+1})$, $POS(W_{i+2})$	5	[1,32]
$SMNI(W_i)$, $SMNI(W_{i-1})$, $SMNI(W_{i-2})$, $SMNI(W_{i+1})$, $SMNI(W_{i+2})$	5	[0,3]
$L(W_i)$, $L(W_{i-1})$, $L(W_{i-2})$, $L(W_{i+1})$, $L(W_{i+2})$	5	[1,5]
PM	1	[0,5]
Total	16	

Table 4. Syllable-level input features to FS unit

Feature	Number of features	Range of feature value
$FC(S_i)$, $FC(S_{i-1})$, $FC(S_{i-2})$, $FC(S_{i+1})$, $FC(S_{i+2})$	5	[1,6]
$V(S_i)$, $V(S_{i-1})$, $V(S_{i-2})$, $V(S_{i+1})$, $V(S_{i+2})$	5	[1,6]
$SC(S_i)$, $SC(S_{i-1})$, $SC(S_{i-2})$, $SC(S_{i+1})$, $SC(S_{i+2})$	5	[0,6]
$TC(S_i)$, $TC(S_{i-1})$, $TC(S_{i-2})$, $TC(S_{i+1})$, $TC(S_{i+2})$	5	[0,6]
$P(S_j/W_i)$	1	[1,4]
Total	21	

- Third consonant of the current syllable ($TC(S_i)$), the third consonant of two previous syllables ($TC(S_{i-1})$ and $TC(S_{i-2})$), and the third consonant of two next syllables ($TC(S_{i+1})$ and $TC(S_{i+2})$) (if they are existed); and
- Position indicator of the current syllable ($P(S_j/W_i)$ showing whether the current syllable forms a monosyllabic word, or is the first/intermediate/last syllable of a polysyllabic word).

The syllable-level features are summarized in Table 4. The value "0" in the range of SC and TC features shows that there is not second or third consonant in that syllable, respectively.

VI. Simulation and Experimental Results

The block diagram of the simulated system in this study is shown in Fig. 3. To select the word-level and syllable-level linguistic features, the parameters setting of GA and ACO parts of the hybrid FS method is reported in Tables 5 and 6, respectively.

- Vowel type of the current syllable (V(Si)), the vowel type of two previous syllables (V(Si-1) and V(Si-2)), and the vowel type of two next syllables (V(Si+1) and V(Si+2));
- Second consonant of the current syllable ($SC(S_i)$), the second consonant of two previous syllables ($SC(S_{i-1})$ and $SC(S_{i-2})$), and the second consonant of two next syllables ($SC(S_{i+1})$ and $SC(S_{i+2})$) (if they are existed);

The binary gravitational search algorithm [78] is also used for feature selection in this study as a modern competitive algorithm. The parameters setting of BGSA is reported in Table 7. In addition, a standard Elman-type recurrent neural network [79] is used as a competitive prosody model technique in this study. In this way, the recurrent connections in the output layer of the structure, shown in Fig. 3, are omitted.

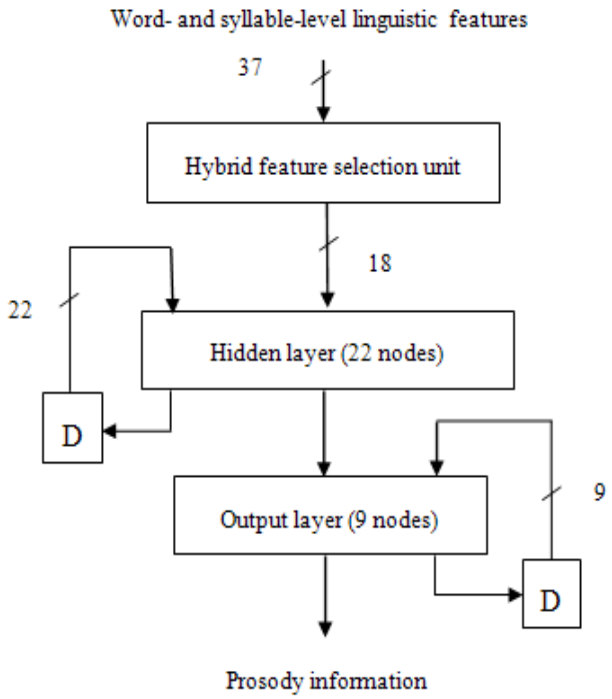

Figure 3. Structure of the modified Elman-type DNN-based prosody generator including hybrid FS unit

Table 5. GA parameters setting in the hybrid algorithm

Parameter	Value
Population size	50
Number of generations	10
Probability of crossover	0.85
Probability of mutation	0.03

Table 6. ACO parameters setting in the hybrid algorithm

Parameter	Value
n_a	50
Number of selected features	18
$\alpha=\beta$	1
k	20

Table 7. BGSA parameters setting for feature selection

Parameter	Value
Number of masses (m)	20
Small constant in the denominator of force equation (ζ)	0.01
Initial value of gravitational constant (G_0)	100
Rate of exponential decrease of gravitation constant (β)	20
Maximum velocity (v_{max})	6
Total number of iterations (t_{max})	50

The training data of proposed DNN-based prosody generator consists of 400 sentences covering possibilities such as positive, negative, short, long, predicative (statement), interrogative, imperative (command), and exclamative sentences. The length of 250 sentences was less than 15 syllables and the length of remaining sentences was less than 50 syllables. The input sentences were spoken at the same rate as an average native speaker (i.e., about four syllables per second). The DNN-based prosody generator was trained using error back-propagation algorithm.

The number of hidden nodes in the hidden layer of the DNN was empirically decided in the experiments of this study. The activation function of all hidden and output nodes was considered as sigmoid and linear, respectively. The number of output nodes was set to 9 in which the pitch contour of a syllable was represented by a smooth curve formed through orthonormal polynomial expansion using four coefficients (by four output nodes of the DNN).

The mean value of the pitch contour is represented by the zeroth-order coefficient. The shape of pitch contour is represented by the other three coefficients. The basis functions of this expansion (i.e., discrete Legendre polynomials) are as follows:

$$\Phi_0(\frac{i}{N})=1 \tag{2}$$

$$\Phi_1(\frac{i}{N})=\left[\frac{12N}{N+2}\right]^{0.5}\left[\frac{i}{N}-0.5\right] \tag{3}$$

$$\Phi_2(\frac{i}{N})=\left[\frac{180N^3}{(N-1)(N+2)(N+3)}\right]^{0.5}\times \left[\left(\frac{i}{N}\right)^2-\left(\frac{i}{N}\right)+\frac{N-1}{6N}\right] \tag{4}$$

$$\Phi_3(\frac{i}{N})=\left[\frac{2800N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right]^{0.5} \times\left[\left(\frac{i}{N}\right)^3-1.5\left(\frac{i}{N}\right)^2+\frac{6N^2-3N+2}{10N^2}\left(\frac{i}{N}\right)-\frac{(N-1)(N-2)}{20N^2}\right] \quad (5)$$

for $0\le i\le N$, and N+1 is the length of the pitch contour. The pitch contour of the kth syllable is approximated by:

$$Pitch_k(i)=\sum_{j=0}^{3}p_j(k)\Phi_j(\frac{i}{N});\qquad 0\le i\le N \quad (6)$$

where $p_j(k)$ is the jth order coefficient in the pitch contour expansion and is calculated by:

$$p_j(k)=\frac{1}{N+1}\sum_{i=0}^{N}\Phi_j(\frac{i}{N})\times Pitch_k(i) \quad (7)$$

The three types of prosodic information have different dynamic ranges, so a distortion measure is used as an objective function for minimization as (8):

$$J(k)=\sum_{j=0}^{3}\{T[p_j(k)]-O[p_j(k)]\}^2+\{T[le(k)]-O[le(k)]\}^2+\sum_{j=0}^{3}\{T[d_j(k)]-O[d_j(k)]\}^2 \quad (8)$$

It is noted that $p_j(k)$ is the jth order coefficient in the pitch contour expansion. The log-energy level of kth syllable is shown by $le(k)$. $d_0(k)$, $d_1(k)$, $d_2(k)$, and $d_3(k)$ represent the syllable duration, the vowel duration in a syllable, the vowel onset time in a syllable, and the inter-syllable pause duration of kth syllable, respectively. $T[p_j(k)]$, $T[le(k)]$, and $T[d_j(k)]$ are the corresponding target values. $O[p_j(k)]$, $O[le(k)]$, and $O[d_j(k)]$ are the actual outputs of the prosody generator.

The number of nodes in the hidden layer of DNN was determined empirically and set to 22. The learning rate for training was initially set to 0.01 and linearly decayed to zero at 200 training epochs. In the first evaluation of the proposed system, the performance is assessed in two cases that can be compared with the original synthesized word: a) using GA-ACO FS unit in the system (shown in Fig. 1); and b) removing FS unit.

It is noted that in the case of no-FS unit in Fig. 3, the number of nodes in the hidden layer was set to 22. The RMSE of the synthesized prosody parameters (averaged over 100 test sentences) for the two mentioned experimental case (i.e., using or removing hybrid FS unit) is reported in Table 8 (in the first and second columns). In addition, the RMSE of the synthesized prosody parameters in the case of removing semantic-related features is included in Table 8 (in the third and fourth columns).

As seen in Table 8, using semantic-related features results in decrease of RMSE of predicted prosodic information (by comparing the RMSEs reported in the first and third columns of Table 8). In addition, using the GA-ACO FS unit results in considerable reduction of connection weights of DNN (as seen in Fig. 3) and this achieves by only a slight increase of RMSE when compared to the case of no-FS unit (by comparing the RMSEs reported in the first and second columns of Table 8 or the RMSEs reported in the third and fourth columns of Table 8).

In the second evaluation of the proposed system, the performance of proposed system is compared with similar systems that used BGSA for feature selection or conventional Elman-type DNN for prosody generation (Table 9).

The number of weight connections in the proposed DNN model when performing/not-performing feature selection is reported in Table 10. As seen in Table 10, the total number of weight connections is reduced by 26.5 percent when feature selection is performed (when employing the same number of hidden nodes in the two experimental cases). This reduction in the size of connections resulted in 32.8 percent reduction in the run time of the prosody generator. It is noted that the activation function of hidden nodes was sigmoid with more computational complexity as compared to the linear activation function of output nodes. So, by performing feature selection, the number of input links to the hidden neurons is reduced significantly which resulted in considerable reduction in the run time of prosody generator.

Using semantic-related features and employing a hybrid GA-ACO feature selection algorithm are the main contributions in this system. By using semantic-related features, the number of input features to the prosody generator is increased by 5 (Table 3). On the other hand, the hybrid FS algorithm reduces the number of features by 19. So, the total number of input features is decreased by 14 as compared to the system which has 32 input features and does not employ semantic-related features and FS unit. In other words, the total 37 input features are reduced to 18 input features (i.e., 51 percent reduction in the number of input features) which is an important achievement in simplification of neural-based prosody generator. However, the performance of this prosody generator is not degraded significantly as compared to the main system (by comparing the performance results reported in the first column and second or fourth columns of Table 8).

The prediction accuracy of the mentioned prosodic information is also analyzed by conducting listening tests to evaluate the quality of synthesized speech. For this purpose, six rating scales of listening effort, comprehension problems, articulation, pronunciation, speaking rate, and voice pleasantness are used to evaluate the quality of synthesized speech as International Telecommunication Union-Telecom sector (ITU-T) recommends in ITU-T P.85 [80]. However, other nonstandard computational models have also been proposed for robust voice quality classification such as fuzzy-input fuzzy-output SVM (F^2SVM) algorithm proposed by Scherer *et al*. [81].

Table 8. RMSE of generated prosody information using modified Elman-type DNN in four experimental cases

Prosodic information	RMSE (using hybrid GA-ACO FS unit and semantic features)	RMSE (using semantic features and no-FS unit)	RMSE (using hybrid GA-ACO FS unit and no-semantic features) [19]	RMSE (using no-FS unit and no-semantic features) [19]
Pitch contour (ms/frame)	0.98	0.89	1.07	0.95
Log-energy level (dB)	4.11	3.85	4.29	4.07
Vowel duration (ms)	36.8	32.1	40.1	35.3
Syllable duration (ms)	51.1	44.3	53.2	48.4
Vowel onset time (ms)	6.1	5.6	6.8	5.9
Inter-syllable pause duration (ms)	35.4	26.8	37.5	29.1

Table 9. Performance comparison of the proposed system with two similar systems using BGSA-based feature selection or conventional Elman-type prosody generator

Prosodic information	RMSE of the proposed system	RMSE of similar system using BGSA for feature selection	RMSE of similar system using Elman-type DNN for prosody generation
Pitch contour (ms/frame)	0.98	1.05	1.13
Log-energy level (dB)	4.11	4.02	4.23
Vowel duration (ms)	36.8	39.1	38.7
Syllable duration (ms)	51.1	52.9	54.3
Vowel onset time (ms)	6.1	6.4	6.6
Inter-syllable pause duration (ms)	35.4	37.7	38.6

Table 10. Number of weight connections in the proposed DNN when selecting/not-selecting input linguistic features

Type of connection	Number of connections when not-performing feature selection	Number of connections when performing feature selection
Input layer to the hidden layer (feed-forward + feedback)	814+484=1298	396+484=880
Hidden layer to the output layer (feed-forward + feedback)	198+81=279	198+81=279
Total number of connections	1577	1159

A total of 20 listeners, from 15 to 48 years old, were participated to subjectively evaluate 24 audio files with lengths ranging from 10-15 seconds. The subjects, 10 males and 10 females, heard each utterance twice; once to rate according to listening effort, comprehension problems, and articulation, and the second time to rate the quality of pronunciation, speaking rate, and voice pleasantness. The listeners filled out a questionnaire based on these factors and scored them on a 5-point scale. Based on these subjective evaluations, the mean opinion score (MOS) for the proposed system was found. It is noted that the responses were translated to a numeric range from 1 to 5 in which 5 is the most positive. The mean and standard deviation of the 5-point scores are given in Table 11.

Table 11. Quality assessment of the proposed method based on ITU-T P.85 recommendation

Quality factor	Mean of 5-point scale	Standard deviation of 5-point scale
Listening effort	3.74	0.78
Comprehension problems	4.17	0.61
Articulation	3.92	0.83
Pronunciation	3.48	0.69
Speaking rate	4.61	0.30
Voice pleasantness	3.24	0.72

VII. Conclision and Future Work

A modified Elman-type DNN was used for prosody generation in this study whose inputs were both word-level and syllable-level linguistic features. A hybrid of GA and ACO algorithms was used for feature selection to result in a light neural model for prosody generation. The pitch contour and log-energy level of a syllable, the duration of a syllable and its constituent vowel, the vowel onset time in a syllable, and inter-syllable pause duration were the prosodic information that were generated by the proposed model at the output layer of DNN. Experimental results showed that the proposed prosody generator can offer an acceptable RMSE of mentioned prosody information when using the hybrid FS unit as compared with the case of no-FS unit. The performance of proposed system was also compared with similar systems that used BGSA for feature selection or conventional Elman-type DNN for prosody generation. The quality of synthesized speech was assessed based on ITU-T P.85 Recommendation that considers intelligibility and naturalness of speech, as well.

As future work, other FS methods such as ones reported in Table 1 can be employed instead of the proposed hybrid method. In addition, purifying the input features to the proposed system by considering more new and informative ones based on an extensive set of behavioral findings related to human speech modification depending on a listener-oriented approach [82] is another proposal for future research.

References

[1] N. G. Ward, A. Vega, and T. Baumann, "Prosodic and temporal features for language modeling for dialog," Speech Communication, vol. 54, pp. 161-174, February 2012.

[2] B. Picart, T. Drugman, and T. Dutoit, "Analysis and HMM-based synthesis of hypo and hyperarticulated speech," Computer Speech & Language, vol. 28, pp. 687-707, March 2014.

[3] M. Ekpenyong, E.-A. Urua, O. Watts, S. King, and J. Yamagish, "Statistical parametric speech synthesis for Ibibio," Speech Communication, vol. 56, pp. 243-251, January 2014.

[4] W. Sandler, I. Meir, S. Dachkovsky, C. Padden, and M. Aronoff, "The emergence of complexity in prosody and syntax", Lingua, vol. 121, pp. 2014-2033, October 2011.

[5] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The use of prosody in the linguistic components of a speech understanding system," IEEE Transactions on Speech and Audio Processing, vol. 8, pp. 519-532, September 2000.

[6] K. Vicsi and G. Szaszák, "Using prosody to improve automatic speech recognition," Speech Communication, vol. 52, pp. 413-426, May 2010.

[7] J. P. H. van Santen, E. T. Prud'hommeaux, and L. M. Black, "Automated assessment of prosody production," Speech Communication, vol. 51, pp. 1082-1097, November 2009.

[8] M. Aguert, V. Laval, L. Le Bigot, and J. Bernicot, "Understanding expressive speech acts: The role of prosody and context in French-speaking 5- to 9-year-olds," Speech, Language and Hearing Research, vol. 53, pp. 1629-1641, December 2010.

[9] L. Chen, X. Mao, P. Wei, Y. Xue, and M. Ishizuka, "Mandarin emotion recognition combining acoustic and emotional point information," Applied Intelligence, vol. 37, pp. 602-612, December 2012.

[10] V. K. R. Sridhar, S. Bangalore, and S. Narayanan, "Enriching machine-mediated speech-to-speech translation using contextual information," Computer Speech & Language, vol. 27, pp. 492-508, February 2013.

[11] B. Zhou, X. Cui, S. Huang, M. Cmejrek, W. Zhang, J. Xue, J. Cui, B. Xiang, G. Daggett, U. Chaudhari, S. Maskey, and E. Marcheret, "The IBM speech-to-speech translation system for smartphone: Improvements for resource-constrained tasks," Computer Speech & Language, vol. 27, pp. 592-618, February 2013.

[12] D. Griol, J. Carbo, and J. M. Molina, "Bringing context-aware access to the web through spoken interaction," Applied Intelligence, vol. 38, pp. 620-640, June 2013.

[13] O. A. Odejubji, A. J. Beaumont, and S. H. S. Wong, "Intonation contour realisation for standard Yorùbá text-to-speech synthesis: A fuzzy computational approach," Computer Speech & Language, vol. 20, pp. 563-588, October 2006.

[14] K. Sreenivasa Rao and B. Yegnanarayana, "Intonation modeling for Indian languages," Computer Speech & Language, vol. 23, pp. 240-256, April 2009.

[15] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," Proc. Int. Conf. Acoustic, Speech, and Signal Processing, Montreal, Canada, pp. 509-512, 17-21 May 2004.

[16] M. Sheikhan and N. Mohammadi, "Neural-based electricity load forecasting using hybrid of GA and ACO for feature selection," Neural Computing and Applications, vol. 21, pp. 1961-1970, November, 2012.

[17] C.-Y. Yeh, S.-C. Chang, and S.-H. Hwang, "A consistency analysis on an acoustic module for Mandarin text-to-speech," Speech Communication, vol. 55, pp. 266-277, February 2013.

[18] M. Sheikhan, M. Nasirzadeh, and A. Daftarian, "Design and implementation of a text to speech system for Farsi language," Journal of School Engineering-Ferdowsi University of Mashhad, vol. 17, no. 2, pp. 31-48, 2005. (In Farsi)

[19] M. Sheikhan, "Synthesizing suprasegmental speech information using hybrid of GA-ACO and dynamic neural network," Proc. 5th Conference on Information and Knowledge Technology, Shiraz, Iran, pp. 175-180, 28-30 May 2013.

[20] D. H. Klatt, "Review of TTS conversion in English," Journal of the Acoustical Society of America, vol. 82, pp. 737-793, September 1987.

[21] G. Bailly, "Integration and rhythmic and syntactic constraints in a model of generation of French prosody," Speech Communication, vol. 8, pp. 137-146, June 1989.

[22] L. S. Lee, C. Y. Tseng, and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, pp. 1309-1320, September 1989.

[23] N. Kaiki, K. Mimura, and Y. Sagisaka, "Statistical modeling of segmental duration and power control for Japanese," Proc. Eurospeech Conf., Genova, Italy, pp. 625-628, 24-26 September 1991.

[24] H. Mixdorff and H. Fujisaki, "A scheme for a model-based synthesis by rule of F_0 contours of German utterances," Proc. Eurospeech Conf., Madrid, Spain, pp. 1823-1826, 18-21 September 1995.

[25] Y. A. El-Imam, "Synthesis of the intonation of neutrally spoken modern standard Arabic speech," Signal Processing, vol. 88, pp. 2206-2221, September 2008.

[26] M. Sheikhan, "Prosody generation in Farsi language," Proc. Int. Symp. Telecommun., Isfahan, Iran, pp. 250-253, 16-18 August 2003.

[27] Q. Sun, K. Hirose, and N. Minematsu, "A method for generation of Mandarin F_0 contours based on tone nucleus model and superpositional model," Speech Communication, vol. 54, pp. 932-945, October 2012.

[28] Y. Xu and S. Prom-on, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," Speech Communication, vol. 57, pp. 181-208, February 2014.

[29] D. R. Van Niekerk and E. Barnard, "Predicting utterance pitch targets in Yorùbá for tone realisation in speech synthesis," Speech Communication, vol. 56, pp. 229-242, January 2014.

[30] P. H. Low and S. Vaseghi, "Application of microprosody models in TTS synthesis," Proc. Int. Conf. Spoken Language Processing, Denver, USA, pp. 2413-2416, 16-20 September 2002.

[31] A. Lazaridis, T. Ganchev, I. Mporas, E. Dermatas, and N. Fakotakis, "Two-stage phone duration modelling with feature construction and feature vector extension for the needs of speech synthesis," Computer Speech & Language, vol. 26, pp. 274-292, August 2012.

[32] S. Winters and M. G. O'Brien, "Perceived accentedness and intelligibility: The relative contributions of F_0 and duration," Speech Communication, vol. 55, pp. 486-507, March 2013.

[33] Y. Hifny and M. Rashwan, "Duration modeling for Arabic TTS synthesis," Proc. Int. Conf. Spoken Language Processing, Denver, USA, pp. 1773-1776, 16-20 September 2002.

[34] E. Tisljár-Szabó and C. Pléh, "Ascribing emotions depending on pause length in native and foreign language speech," Speech Communication, vol. 56, pp. 35-48, January 2014.

[35] N. P. Narendra and K. Sreenivasa Rao, "Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis," Applied Soft Computing, vol. 13, pp. 773-781, February 2013.

[36] G. Huang and M. Joo Er, "An adaptive neural control scheme for articulatory synthesis of CV sequences," Computer Speech & Language, vol. 28, pp. 163-176, January 2014.

[37] F. Tomaschek, H. Truckenbrodt, and I. Hertrich, "Neural processing of acoustic duration and phonological German vowel length: Time courses of evoked fields in response to speech and nonspeech signals," Brain and Language, vol. 124, pp. 117-131, January 2013.

[38] R. Kohavi and J. H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, pp. 273-324, December 1997.

[39] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, pp. 16-28, January 2014.

[40] C. De Stefano, F. Fontanella, C. Marrocco, and A. Scotto di Freca, "A GA-based feature selection approach with an application to handwritten character recognition," Pattern Recognition Letters, vol. 35, pp. 130-141, January 2014.

[41] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," Applied Soft Computing, vol. 18, pp. 261-276, May 2014.

[42] A. Al-Ani, A. Alsukker, and R. N. Khushaba, "Feature subset selection using differential evolution and a wheel based search strategy," Swarm and Evolutionary Computation, vol. 9, pp. 15-26, April 2013.

[43] B. Chen, L. Chen, and Y. Chen, "Efficient ant colony optimization for image feature selection," Signal Processing, vol. 93, pp. 1566-1576, June 2013.

[44] F. Ghareh Mohammadi and M. Saniee Abadeh, "Image steganalysis using a bee colony based feature selection algorithm," Engineering Applications of Artificial Intelligence, vol. 31, pp. 35-43, May 2014.

[45] B. Mohseni Bababdani and M. Mousavi, "Gravitational search algorithm: A new feature selection method for QSAR study of anticancer potency of imidazo[4,5-b]pyridine derivatives," Chemometrics and Intelligent Laboratory Systems, vol. 122, pp. 1-11, March 2013.

[46] R. Y. M. Nakamura, L. A. M. Pereira, D. Rodrigues, K. A. P. Costa, J. P. Papa, and X. S. Yang, "Binary bat algorithm for feature selection," In: Swarm Intelligence and Bio-inspired Computation, Theory and Applications, 1st ed., Elsevier, pp. 225-237, 2013.

[47] C. E. Gillies, M. R. Siadat, N. V. Patel, and G. D. Wilson, "A simulation to analyze feature selection methods utilizing gene ontology for gene expression classification," Journal of Biomedical Informatics, vol. 46, pp. 1044-1059, December 2013.

[48] F. Lin, D. Liang, C. C. Yeh, and J. C. Huang, "Novel feature selection methods to financial distress prediction," Expert Systems with Applications, vol. 41, pp. 2472-2483, April 2014.

[49] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," Pattern Recognition, vol. 43, pp. 5-13, January 2010.

[50] C. J. Tan, C. P. Lim, and Y. N. Cheah, "A multi-objective evolutionary algorithm-based ensemble optimizer for feature selection and classification with neural network models," Neurocomputing, vol. 125, pp. 217-228, February 2014.

[51] M. Sheikhan, M. Pezhmanpour, and M. S. Moin, "Improved contourlet-based steganalysis using binary particle swarm optimization and radial basis neural networks," Neural Computing and Applications, vol. 21, pp. 1717-1728, October 2012.

[52] S. M. Vieira, L. F. Mendonça, G. J. Farinha, and J. M. C. Sousa, "Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients," Applied Soft Computing, vol. 13, pp. 3494-3504, August 2013.

[53] C. Bae, W. C. Yeh, Y. Y. Chung, and S. L. Liu, "Feature selection with intelligent dynamic swarm and rough set," Expert Systems with Applications, vol. 37, pp. 7026-7032, October 2010.

[54] H. H. Inbarani, A. T. Azar, and G. Jothi, "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis," Computer Methods and Programs in Biomedicine, vol. 113, pp. 175-185, January 2014.

[55] A. Ghosh, A. Datta, and S. Ghosh, "Self-adaptive differential evolution for feature selection in hyperspectral image data," Applied Soft Computing, vol. 13, pp. 1969-1977, April 2013.

[56] S. Pan, K. McKeown, and J. Hirschberg, "Exploring features from natural language generation for prosody modeling," Computer Speech & Language, vol. 16, pp. 457-490, July 2002.

[57] M. Sheikhan, M. Tebyani, and M. Lotfizad, "Using symbolic and connectionist approaches to automate editing Persian sentences syntactically," Proc. Int. Conf. Intelligent and Cognitive Systems, Tehran, Iran, pp. 250-253, 23-26 September 1996.

[58] M. Sheikhan, M. Tebyani, and M. Lotfizad, "Continuous speech recognition and syntactic processing in Iranian Farsi language," International Journal of Speech Technology, vol. 1, pp. 135-141, March 1997.

[59] M. Sheikhan, M. Tebyani, and M. Lotfizad, "Conceptual classification and semantic disambiguation of Farsi words using neural networks," Proc. Int. Conf. Intelligent and Cognitive Systems, Tehran, Iran, pp. 35-39, 23-26 September 1996. (In Farsi)

[60] F. Hinskens, B. Hermans, and M. van Oostendorp, "Grammar or lexicon. Or: Grammar and lexicon? Rule-based and usage-based approaches to phonological variation," Lingua, vol. 142, pp. 1-26, April 2014.

[61] A. Ramsay, I. Alsharhan, and H. Ahmed, "Generation of a phonetic transcription for modern standard Arabic: A knowledge-based model," Computer Speech & Language, vol. 28, pp. 959-978, July 2014.

[62] M. Sheikhan, M. Nasirzadeh, and A. Daftarian, "Text to speech for Iranian dialect of Farsi language," Proc. 2nd Workshop on Farsi Computer Speech, Tehran, Iran, pp. 39-53, 27-28 June 2006.

[63] H. Quené and R. Kager, "The derivation of prosody for text-to-speech from prosodic sentence structure," Computer Speech & Language, vol. 6, pp. 77-98, January 1992.

[64] X. Guo, L. Zheng, L. Zhu, Z. Yang, C. Chen, L. Zhang, W. Ma, and Z. Dienes, "Acquisition of conscious and unconscious knowledge of semantic prosody," Consciousness and Cognition, vol. 20, pp. 417-425, June 2011.

[65] H. Vafaie and I. F. Imam, "Feature selection methods: Genetic algorithms vs. Greedy-like search," Proc. 3rd Int. Fuzzy Systems and Intelligent Control Conf., Louisville, USA, March 1994.

[66] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. Hosseinzadeh Aghdam, "A novel ACO-GA hybrid algorithm for feature selection in protein function prediction," Expert

Systems with Applications, vol. 36, pp. 12086-12094, December 2009.

[67] S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," IEEE Transactions on Communications, vol. 38, pp. 1317-1320, September 1990.

[68] D. O'Brien and A. Monaghan, "Concatenative synthesis based on a harmonic model," IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 11-20, January 2001.

[69] Y. Stylianou, "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech," Proc. IEEE Nordic Signal Processing Symp., Espoo, Finland, 24-27 September 1996.

[70] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 232-239, March 2001.

[71] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis," Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Seattle, USA, pp. 273-276, 15 May 1998.

[72] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," IEEE Transactions on Signal Processing, vol. 40, pp. 497-510, March 1992.

[73] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for TTS synthesis using diphones," Speech Communication, vol. 9, pp. 453-467, December 1990.

[74] T. Dutoit, An Introduction to Text-To-Speech Synthesis, Kluwer Academic Publishers, Norwell, USA, 1997.

[75] D. O'Brien and A. Monaghan, "Shape invariant time-scale modification of speech using a harmonic model," Proc. Int. Conf. Acoustic, Speech, and Signal Processing, Phoenix, USA, pp. 381-384, 15-19 March 1999.

[76] Y. Samareh, Phonetics of Farsi Language, Tehran: University Press Center, 1995. (In Farsi)

[77] A. Al-Ani, "Feature subset selection using ant colony optimization," International Journal of Computational Intelligence, vol. 2, pp. 53-58, 2005.

[78] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi, "BGSA: Binary gravitational search algorithm," Natural Computing, vol. 9, pp. 727-745, September 2010.

[79] J. L. Elman, "Finding structure in time," Cognitive Science, vol. 14, pp. 179-211, April 1990.

[80] ITU-T Recommendation P.85, Telephone transmission quality subjective opinion tests. A method for subjective performance assessment of the quality of speech voice output devices, 1994.

[81] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," Computer Speech & Language, vol. 27, pp. 263-287, January 2013.

[82] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," Computer Speech & Language, vol. 28, pp. 543-571, March 2014.

Mansour Sheikhan is currently an Associate Professor in Electrical Engineering Department of Islamic Azad University-South Tehran Branch. His research interests include speech signal processing, neural networks, and intelligent systems. He has published more than 90 journal papers, 70 conference papers, four books in Farsi, and seven book chapters for IET, Springer and Taylor & Francis.