Performance of PSO and GWO Algorithms Applied in Text-Independent Speaker Identification

Lucas Schulze
Departamento de Engenharia Elétrica
Universidade do Estado de Santa Catarina
Joinville, Brasil
schulze.lucass@gmail.com

Renan Sebem
Departamento de Engenharia Elétrica
Universidade do Estado de Santa Catarina
Joinville, Brasil
renan.sebem@udesc.br

Douglas Wildgrube Bertol
Departamento de Engenharia Elétrica
Universidade do Estado de Santa Catarina
Joinville, Brasil
douglas.bertol@udesc.br

***Abstract*—In this paper, we analyze the performance of two bio-inspired algorithms applied in text-independent speaker recognition through voice signal. The analyzed algorithms are *particle swarm optimization* and *grey wolf optimizer*. The complete methodology described in this paper was specifically developed in the context of this work. First, a widely known model of the speaker is determined based on discrete transfer functions. Then a method of estimation of the input signal is determined. The bio-inspired algorithms are custom-developed and applied to parameterize the transfer functions based on the models. The proposed method is composed by three major parts, first the fitness used in the bio-inspired algorithms is created based on the cross-correlation. Second, a method to create a database with speakers' identities is proposed, and third, a method to compare the characteristics of the speaker is proposed, to identify or distinguish two different speakers. Finally, experiments were made considering 4 speakers with 2 speech each, a representation of the identity of each speaker was created through both algorithms, totalizing 16 entries on the database. Results show that all comparison results were accurate. The algorithms identify the speaker even when two different speeches were compared, and, as expected, distinguish when two different speakers were compared.**

***Index Terms*—Speaker Identification, Grey Wolf Optimizer, Particle Swarm Optimization, Bio-Inspired Algorithms**

I. Introduction

The information obtained through human voice signal processing may be useful in many applications. With the aid of the current technology, people have access to a popular device which, among other functions, is a high-performance voice processor: the smartphone. In the smartphone are present many voice signal processing techniques, such as noise-canceling filtering, language recognition [1], speech recognition, speaker recognition [2]. Voice processing techniques have another applications, such as speaker diarisation [3] and emotion identification through the voice signal [4].

Due to the influence of speech in speaker identification [5], different approaches are proposed in the literature. The identification can be done considering the speech [6], [7], independently of the speech [8], or independently of the speech and the language [9].

There are many techniques, in the literature, for text independent speaker recognition, such as: linear discriminant analysis [10]; artificial neural networks and hidden Markov chains [11]; data mining [12]; Karhunen-Love transform [13]; VQ distortion [14], mel-cepstral coefficients [15]; ant colony optimization [16]; Particle Swarm Optimization (PSO) [17].

Between the techniques listed above, we note that in [16], [17], bio-inspired optimization algorithms were used. However, these algorithms were applied directly to the voice signal, without the aid of a model of signals and systems theory. The human voice is a stochastic signal, and therefore its identification does not have an analytical solution. One possible application of optimization algorithms is to find parameters of a deterministic model of a speaker, based on his voice signal.

System identification may rely on some information of the system, e.g., the synchronous data of the input and output of the system in a period of time. In the case of human vocal system identification, the input signal is generated by the glottis and vocal cords, which, to the best of our knowledge, it is not possible to physically measure.

In this work, it is proposed a method of speaker identification independently of the speech and language, or in other words, *text independent*. The method use optimization algorithms to parameterize transfer functions of well-known models from the literature of voice processing [18]. When using a model, the efficiency of the optimization algorithms may be enhanced. Further, we propose the use of the Grey Wolf Optimizer (GWO) algorithm [19], which, to the best of our knowledge, has not yet been applied to speaker identification. Yet, the performance of GWO is compared to the PSO algorithm [20].

The methodology proposed in this work is based on discrete transfer function models for the glottis, vocal tract, and radiation [18]. The parameters to the transfer function are obtained through an optimization algorithm (GWO or PSO) with a custom fitness function for the speaker identification. Also, a structure for the database with the speakers' identification is proposed. And finally, a method of evaluation for the identification (or distinguishing) between speakers is proposed.

This work is organized as follows. In Section II, the methodology used for the speaker identification is presented. Then, we present the vocal tract modeling techniques and the techniques used in the implementation of the custom optimization algorithms. In Section III, the results of the work are presented, considering experiments with 4 professional speakers. A discussion is made in Section IV. And finally, in

Section V we conclude on the performance of the algorithms.

II. Methodology

In this section, we present the proposed methodology for the speaker identification. Basically, we use optimization algorithms to find the parameters of a discrete transfer function model of the human vocal system, and then, we develop a method of identification over these parameters.

A. Model

The block diagram of the complete model is shown in Figure 1.

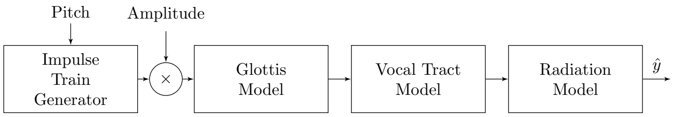

Fig. 1. Basic model of the human vocal system in block diagram [18].

In this model, the vocal tract and vocal cords are considered, which are the parts of the system where the vocalized sounds are generated. Stochastic wave sounds are not considered in this model, e.g., the sound of 's'. The reason for this consideration in the model is because the speaker identity is mainly defined by vocalized sounds. The model is composed of three parts: glottis model, vocal tract model and radiation model (lips model).

1) Glottis Model: The glottis is the source of the sound in the human vocal tract, whereby the pitch is controlled [21]. There are different approaches to model and parameterize the glottis model using the signal from the speaker's voice [22]–[24]. Because of the characteristics of the human glottis, it has a finite impulse response (FIR), and the resulting transfer function of the vocal system would have more zeros than poles, which is difficult to deal in the Z domain. Therefore, the glottis model is implemented in the time domain as an impulse response. A common model for the glottis is presented in Equation 1:

$$g(n) = \begin{cases} \frac{1}{2}\left[1 - \cos(\frac{\pi n}{N_1})\right], & 0 \leq n < N_1; \\ \cos(\frac{\pi(n-N_1)}{2N_2}), & N_1 \leq n \leq N_1 + N_2; \\ 0, & \text{otherwise.} \end{cases} \qquad (1)$$

The parameters N_1 and N_2 are proportional to the opening and closing time of the glottis, respectively. They are important characteristics of the human vocal system. It is possible to observe a typical impulse response of a glottis model in Figure 2.

2) Vocal Tract Model: The vocal tract model is based on the lossless tube model, where the tube is composed of a sequence of modules that represents each vocal tract part, e.g., trachea, larynx, and pharynx. The transfer function of the vocal tract model is given by:

$$V(z) = \Pi_{k=1}^{M} V_k(z), \qquad (2)$$

with:

$$V_k(z) = \frac{1 - 2|z_k|\cos(2\pi F_k T) + |z_k|^2}{1 - 2|z_k|\cos(2\pi F_k T)z^{-1} + |z_k|^2 z^{-2}}, \qquad (3)$$

where M is the number of modules; $|z_k|$ is the absolute value from the poles of $V_k(z)$; F_k is the resonance frequency of $V_k(z)$; and, T is the sampling time. Each module $V_k(z)$ is associated to two complex conjugate poles resulting in an underdamped system. For each part of the vocal tract, the quantity is increased by M modules $V_k(z)$ in the transfer function, where each module represents a different resonance frequency.

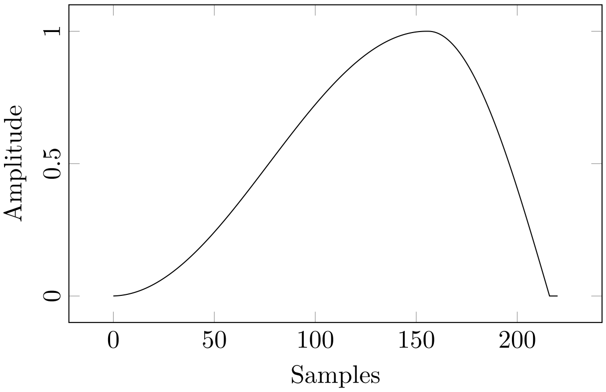

Fig. 2. Impulse response from glottis model.

3) Radiation Model: The lips characterize the last stage of the vocal system. Their function is the radiation of the sound wave, forming a wave front. They radiate the wave of the sound, forming a wavefront that directs the sound in all directions.

The radiation model is defined by a transfer function with only one zero:

$$R(z) = R_o(1 - z^{-1}), \qquad (4)$$

where R_o is the radiation gain.

B. Input Signal Estimation

The input signal is composed of two variables: pitch and amplitude. There are no transducers for this signal, then, it cannot be measured. Therefore, it must be estimated from the voice signal [18], [25].

Since the objective is to identify the speaker and not the speech, a low pass filter is applied to the voice signal, removing the frequencies not related to the speaker identity, e.g., the sound of 's'. The resulting signal is composed only of frequencies from the vocalized sounds, which allows the identification of the speaker [25], [26].

After the filtering, the power spectral density (PSD) is calculated through the correlogram [27]. The PSD is calculated for each window, in time, of 23ms with 11ms of overlap. Then, the pitch is determined as the frequency which has the peak in each window, and the amplitude is determined as the square root of the peak in each window. In Figure 3, it is presented the pitch signal from a speech (dashed line).

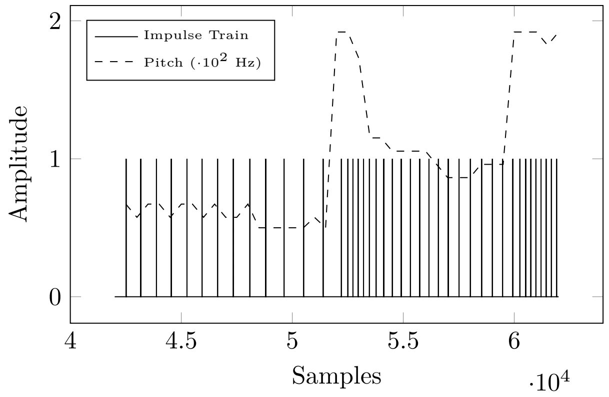

Fig. 3. *Pitch* estimation (dashed), and resulting impulse train (continuous).

Next, the pitch signal is converted to an impulse train with constant amplitude, where the time between each impulse is the period of the pitch. That way, the higher the pitch, the smaller the time between two impulses. A train impulse obtained from the pitch is presented in Figure 3.

From the PSD computed with a window of 23ms, the amplitude is determined as the square root of the peak in each window [25]. The estimated amplitude should be the envelope of the voice signal, demonstrated in Figure 4.

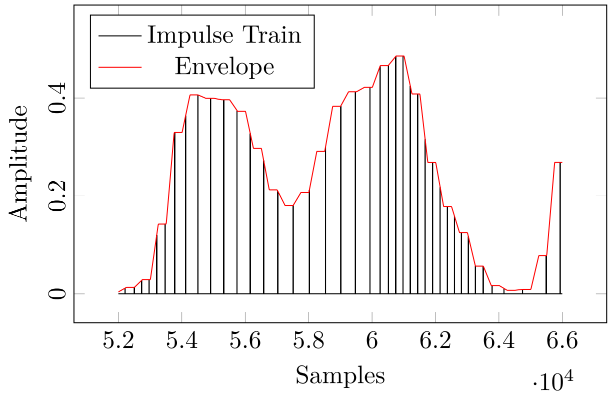

Fig. 4. Estimated amplitude (red) and original voice signal (black).

The obtained impulse train and the estimation of the amplitude are combined to establish an impulse train enveloped by the amplitude value, as illustrated in Figure 5.

Finally, the impulse train δ_{T_p} enveloped with the amplitude $a(n)$ is convoluted with the transfer function from the glottis $g(n)$ in order to obtain the input signal $x(n)$:

$$x(n) = \big(a(n) \cdot \delta_{T_p}(n)\big) * g(n). \tag{5}$$

A graphical example of the input signal is presented in Figure 6.

C. Bio-inspired Algorithms

The algorithms and their implementation details are presented in this section.

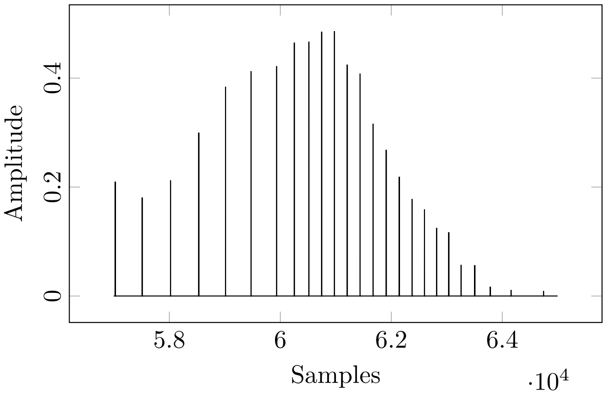

Fig. 5. Estimated impulse train enveloped by the estimated amplitude.

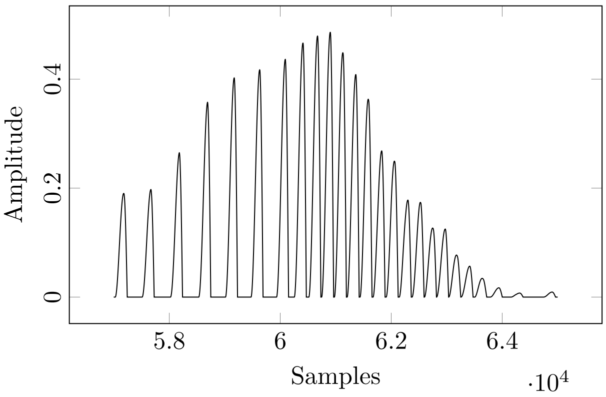

Fig. 6. Example of estimated input signal for the vocal tract system.

1) Grey Wolf Optimizer (GWO): The Grey Wolf Optimizer is a meta-heuristic algorithm inspired by the wolf pack social structure [19]. The wolves hierarchy is composed of 4 levels, alpha, beta, delta, and omega, where the alpha wolf is considered the most capable of making decisions.

In the mathematically model of the social hierarchy, the fittest solution is the alpha, the second and third best solutions are the beta and delta, respectively. And the rest of the solutions are designated as omegas.

After initialization with random values, the algorithm consists in every iteration determine the fittest three solutions, and then update the next position of each particle based on its distance to alpha, beta, and delta. Accordingly, the next position $\vec{X}(n+1)$ from each wolf is updated as follow:

$$\vec{X}(n+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3}, \tag{6}$$

where $\vec{X}_1$ is the next position with regard to alpha wolf, determined by:

$$\vec{X}_1 = \vec{X}_\alpha(n) - \vec{A}_1 \cdot \vec{D}_\alpha, \tag{7}$$

with $\vec{X}_\alpha(n)$ is the current alpha position, and $\vec{D}_\alpha$ computed accordingly to:

$$\vec{D}_\alpha = |\vec{C}_1 \cdot \vec{X}_\alpha(n) - \vec{X}(n)|, \tag{8}$$

$\vec{A}_1$ and $\vec{C}_1$ are coefficient vectors defined as:

$$\vec{A}_1 = 2\vec{a} \cdot \vec{r}_1 - \vec{a}, \tag{9}$$

$$\vec{C}_1 = 2\vec{r}_2, \tag{10}$$

where the elements of $\vec{a}$ decrease linearly from 2 to 0 over the iterations; and, $\vec{r}_1$ and $\vec{r}_2$ are vectors with random values between 0 to 1.

Therefore the next position based on beta and delta, $\vec{X}_2$ and $\vec{X}_3$ respectively, are calculated with the same presented procedure.

The linear decrease of $\vec{a}$ tends the wolves to diverge from the fittest solutions in the first half of the iterations, referred to as the *Hunting* phase, and in the second half to converge, the *Attacking* phase. The random values in $\vec{r}_1$ and $\vec{r}_2$ contribute to avoiding local solutions.

2) Particle Swarm Optimization (PSO): Particle Swarm Optimization is an algorithm inspired by the social behavior of bird flocking.

After initialization of all particles position with random values, the position and velocity from each particle is updated in every iteration regarding its best solution and the global best solution, as described by Equations (11) and (12):

$$\begin{aligned}\vec{V}(n+1) = w \cdot \vec{V}(n) + c_1 r_1(\vec{P}_b - \vec{X}(n)) + \\ + c_2 r_2(\vec{P}_g - \vec{X}(n)),\end{aligned} \tag{11}$$

$$\vec{X}(n+1) = \vec{X}(n) + \vec{V}(n+1), \tag{12}$$

where $\vec{V}$ is the particle inertial velocity; w is an inertial coefficient; $\vec{X}$ is the particle position; P_b is the best position from the owns particle; P_g is the global best position; r_1 and r_2 are random values between 0 and 1; and, c_1 and c_2 are learning factors.

3) Fitness Function: to identify the vocal tract system, it is not possible to use conventional methods to determine the fitness [28]. The output signal from the human vocal tract system is a wave signal composed of different harmonics, what increases the complexity of the error computation between two signals. Hence, a fit method is created.

To evaluate the correctness of the calculated parameters, it is used the cross-correlation ($\star$) between the original signal and the signal generated by the estimated parameters from the optimization algorithm. If both signals are similar, the peek from cross-correlation will be high. That way, the peak value of the autocorrelation from the original sound is used as the reference parameter, and it is compared to the peak value from the cross-correlation between the original signal and the signal from the bio-inspired algorithm. The lower the difference between the peaks, the more the signals are similar. The presented procedure is described in Equation (13).

$$fit_{\%} = 100 \cdot \left(1 - \frac{max(y_r \star y_r) - max(y_r \star y)}{max(y_r \star y_r)}\right), \tag{13}$$

where y_r is the reference signal from a database, and y is the signal under evaluation.

D. Speaker Identity

The speaker identity is associated to the maximum value from the autocorrelation of the transfer function defined by the coefficients $[a_1^r \; a_2^r \; \ldots \; a_n^r]$, that is:

$$id_r = \begin{bmatrix} a_1^r & a_2^r & \ldots & a_n^r \end{bmatrix} \begin{bmatrix} a_1^r \\ a_2^r \\ \vdots \\ a_n^r \end{bmatrix}, \tag{14}$$

and it is stored in a database for comparison, where a_i^r represents the reference coefficients from the database.

The verification test for the speaker identification is made through the cross-correlation with zero lag of the database coefficients and the verification coefficients:

$$id_{test} = \begin{bmatrix} a_1^r & a_2^r & \ldots & a_n^r \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \tag{15}$$

where a_i denotes the coefficients computed for the speaker under identification.

The error is calculated by:

$$error_{\%} = 100 \cdot \frac{\mid id_r - id_{test} \mid}{id_r}, \tag{16}$$

where $error$ is the difference between the voice of the speaker used as a reference and the speaker under identification. That way, the smaller the error, the more similar are the voices from the speakers.

E. Speaker Identification

After determining the error between the identity from the speaker under identification and the speakers from the database, it is necessary to define an identification criterion. In this case, it is defined as an error threshold, that is, if the error is lower than the threshold, both speakers are considered the same. The threshold value can be determined experimentally by statistical analysis.

The error that is lower than the threshold is denominated as identification error. Otherwise, it is defined as rejection error, in which case the speeches have been spoken by two different speakers.

III. Results

To compare the algorithms presented in Section II, it is evaluated two speeches of four professional native english speakers, two men and two women of 30 years old, from a radio commercials database [29]. The speeches are recorded without noisy in a professional studio.

Each of the eight speeches is evaluated by both algorithms, resulting in 16 references for the database, as described by Equation (14). The obtained parameters from each speech are correlated with each other by Equation (15). That way, the relative errors obtained with Equation (16) among all the tested speeches are presented in Table I.

TABLE I
RELATIVE ERROR (%) AMONG ALL EVALUATED SAMPLES.

[]	B1G	B1P	B2G	B2P	G1G	G1P	G2G	G2P	C1G	C1P	C2G	C2P	K1G	K1P	K2G	K2P
B1G	[]	0.072	0.018	0.038	2.638	2.550	2.609	2.501	1.518	1.426	1.434	1.347	0.834	0.692	0.783	0.653
B1P	0.072	[]	0.090	0.110	2.703	2.616	2.675	2.567	1.443	1.350	1.359	1.271	0.760	0.618	0.709	0.579
B2G	0.018	0.091	[]	0.020	2.621	2.534	2.593	2.484	1.538	1.445	1.454	1.366	0.853	0.711	0.802	0.671
B2P	0.038	0.111	0.020	[]	2.603	2.516	2.575	2.466	1.559	1.466	1.475	1.387	0.874	0.731	0.823	0.692
G1G	2.903	2.982	2.882	2.860	[]	0.097	0.033	0.152	4.571	4.470	4.479	4.382	3.820	3.662	3.763	3.620
G1P	2.796	2.875	2.776	2.753	0.096	[]	0.064	0.055	4.459	4.358	4.367	4.270	3.710	3.553	3.654	3.511
G2G	2.865	2.944	2.845	2.822	0.033	0.064	[]	0.119	4.531	4.430	4.438	4.341	3.780	3.623	3.724	3.581
G2P	2.735	2.814	2.715	2.692	0.151	0.055	0.119	[]	4.394	4.293	4.302	4.205	3.647	3.490	3.591	3.448
C1G	1.443	1.374	1.460	1.479	3.947	3.864	3.920	3.818	[]	0.088	0.080	0.163	0.651	0.785	0.699	0.822
C1P	1.359	1.290	1.376	1.395	3.870	3.787	3.844	3.740	0.088	[]	0.008	0.075	0.564	0.699	0.612	0.736
C2G	1.366	1.297	1.384	1.403	3.878	3.795	3.851	3.748	0.080	0.008	[]	0.083	0.572	0.707	0.620	0.744
C2P	1.287	1.218	1.305	1.324	3.807	3.723	3.780	3.676	0.164	0.075	0.083	[]	0.491	0.626	0.539	0.663
K1G	0.810	0.740	0.828	0.847	3.373	3.288	3.345	3.240	0.665	0.575	0.584	0.499	[]	0.138	0.049	0.176
K1P	0.676	0.606	0.694	0.713	3.252	3.167	3.225	3.119	0.807	0.717	0.725	0.640	0.138	[]	0.089	0.038
K2G	0.762	0.692	0.780	0.799	3.330	3.245	3.302	3.197	0.716	0.626	0.634	0.549	0.049	0.088	[]	0.127
K2P	0.638	0.568	0.656	0.676	3.217	3.132	3.190	3.084	0.847	0.756	0.764	0.679	0.177	0.039	0.127	[]

The rule for the denomination presented in Table I is described as follows. The first character denotes the first character from the speaker's name, where the speakers are Bill, Garth, Charlotte e Kelly. The second character detail if it is the first or the second speech. The last character is related to the algorithm used: "1" for GWO and "2" for PSO.

The developed codes are available on GitHub[1].

For GWO, it is necessary to set the number of iterations and elements, selected as 30 and 2000, respectively. From Equation (11), it is necessary to tune extra parameters for PSO, namely: an initial value for w, w, c_1 and c_2, resulting in six parameters. In Table II are presented the parameters used for PSO.

TABLE II
PARAMETERS SELECTED FOR PSO.

w_{ini}	w	c_1	c_2
0.06	0.04	0.06	0.028

A. Estimation Parameters

Based on the models presented, the Vocal Tract Model is considered to be a fourth-order system ($M = 2$), then $V(z) = V_1(z) \cdot V_2(z)$.

Therefore, the parameters to be estimated are:

- Glottis Model: N_1 and N_2;
- Vocal Tract Model: z_1, z_2, F_1, and F_2;
- Radiation Model: R_o.

From Equations (14) and (15), four parameters (a_1 to a_4) are necessary to be determined for each speech from each speaker.

B. GWO and PSO Comparison

To summarize the results and compare the algorithms, the average and the standard deviation from the identification error and the rejection error are computed.

Ideally, the identification error is considered 0%. The identification error is computed between two samples from the database related to different speeches from the same speaker and by the same algorithm.

The rejection error does not have an ideal value, because it is related to the differences from each speaker's voice. Although, the higher the rejection error, the more different are the voices. The rejection error is determined between two samples obtained by the same algorithm from different speeches and different speakers.

The average and standard deviation for the identification and rejection errors are presented in Table III.

TABLE III
AVERAGE ERROR AND STANDARD DEVIATION FROM IDENTIFICATION AND REJECTION FOR EACH ALGORITHM.

		GWO	PSO
Identification	Average Error	0.045 %	0.070 %
	Standard Deviation	0.027 %	0.031 %
Rejection	Average Error	2.16 %	2.17 %
	Standard Deviation	1.29 %	1.22 %

IV. DISCUSSION

It is necessary to tune four extra parameters for PSO in comparison with GWO. Hence, as the parameters were tuned empirically, the results from GWO are more consistent due to the smaller number of parameters, resulting in a small sensitivity to the tuning procedure.

The increase of the transfer function order results in a more reliable result. However, due to the increase of the computational cost, it was not possible to increase the order from the fourth-order defined previously in Subsection III-A. That way, the transfer functions used are of fourth-order with two pairs of complex conjugate poles.

The application of the GWO presented more efficient results, with average error and standard deviation smaller than the PSO. Also, the rejection efficiency, when two different speakers are compared, demonstrated similar results for both

[1] https://github.com/Schulze18/Who-is-this

algorithms. The obtained values for the average error and standard deviation are presented in Table III.

Based on the analysis of all the errors in Table I, it is important to remark that no false positives or false negatives occurred in both identification and rejection. This is verified by analyzing the average error presented in Table III, the rejection is 29 bigger than the identification average for PSO and 47 for GWO, which also demonstrates the higher efficiency of GWO when compared to PSO.

Despite the limited number of speakers and speeches evaluated, the results for both algorithms are consistent. We have not considered more speakers due to the high computational cost: 30 minutes per run on average.

V. Conclusion

Both algorithms, GWO and PSO, demonstrated to be reliable for the speaker identification, where no false positive nor false negative occurred.

Also, it is possible to conclude that the use of the GWO is more efficient for the speaker identification by transfer functions model. The main reasons are two: the first and more important is the performance, where the average identification error is smaller than the PSO, and the rejection error are similar for both algorithms. The second is that the tuning process of the GWO is simpler, reducing possible errors from the user in the setup phase.

Acknowledgement

The first author acknowledges UNIEDU/FUMDES for the Master degree scholarship program of the Santa Catarina State.

The second author acknowledges UNIEDU/FUMDES for the Ph.D. scholarship program of the Santa Catarina State.

This work was supported by the Fundação de Amparo à Pesquisa e Inovação de Santa Catarina (Fapesc 2021 TR 930).

References

[1] C. R. Salamea Palacios, L. F. D'Haro, and R. Cordoba, "Language recognition using neural phone embeddings and RNNLMs," *IEEE Latin America Transactions*, vol. 16, no. 7, pp. 2033–2039, July 2018.

[2] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, no. Supplement C, pp. 250 – 271, 2017.

[3] M. Moattar and M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 54, no. 10, pp. 1065 – 1103, 2012.

[4] L. Yukio Mano, E. Vasconcelos, and J. Ueyama, "Identifying emotions in speech patterns: Adopted approach and obtained results," *IEEE Latin America Transactions*, vol. 14, no. 12, pp. 4775–4780, Dec 2016.

[5] J. C. Wu, A. F. Martin, C. S. Greenberg, and R. N. Kacker, "The impact of data dependence on speaker recognition evaluation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 5–18, Jan 2017.

[6] G. Chau and G. Kemper, "One channel subvocal speech phrases recognition using cumulative residual entropy and support vector machines," *IEEE Latin America Transactions*, vol. 13, no. 7, pp. 2135–2143, July 2015.

[7] T. Stafylakis, M. J. Alam, and P. Kenny, "Text-dependent speaker recognition with random digit strings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1194–1203, July 2016.

[8] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[9] J. Markel and S. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 1, pp. 74–82, Feb 1979.

[10] A. Khosravani and M. M. Homayounpour, "A PLDA approach for language and text independent speaker recognition," *Computer Speech & Language*, vol. 45, no. Supplement C, pp. 457 – 474, 2017.

[11] F. L. Alegre, "Application of ANN and HMM to automatic speaker verification," *IEEE Latin America Transactions*, vol. 5, no. 5, pp. 329–337, Sept 2007.

[12] P. Univaso, J. M. Ale, and J. A. Gurlekian, "Data mining applied to forensic speaker identification," *IEEE Latin America Transactions*, vol. 13, no. 4, pp. 1098–1111, April 2015.

[13] C. C. T. Chen, C.-T. Chen, and C.-M. Tsai, "Hard-limited karhunen-loeve transform for text independent speaker recognition," *Electronics Letters*, vol. 33, no. 24, pp. 2014–2016, Nov 1997.

[14] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 456–459, Jul 1994.

[15] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, Jan 2006.

[16] S. Nemati and M. E. Basiri, "Text-independent speaker verification using ant colony optimization-based selected features," *Expert Systems with Applications*, vol. 38, no. 1, pp. 620 – 630, 2011.

[17] R. Luo, W. Cai, M. Chen, and D. Zhu, "An improved particle swarm optimization algorithm for speaker recognition," in *2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI)*, Oct 2012, pp. 641–644.

[18] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, ser. Prentice-Hall signal processing series. Prentice-Hall, 1978.

[19] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46 – 61, 2014.

[20] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, Nov 1995, pp. 1942–1948 vol.4.

[21] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117 – 1138, 2014.

[22] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469–472, June 2009.

[23] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20 – 34, 2012.

[24] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, Sep 1999.

[25] L. Sukhostat and Y. Imamverdiyev, "A comparative analysis of pitch detection methods under the influence of different noise conditions," *Journal of Voice*, vol. 29, no. 4, pp. 410 – 417, 2015.

[26] K. S. S. Srinivas and K. Prahallad, "An FIR implementation of zero frequency filtering of speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2613–2617, Nov 2012.

[27] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005.

[28] J. Lu, W. Xie, and H. Zhou, "Combined fitness function based particle swarm optimization algorithm for system identification," *Computers & Industrial Engineering*, vol. 95, pp. 122 – 134, 2016.

[29] VocaliD, "Vocalid's human voicebank," 2021, URL: https://www.vocalid.co/voicebank/.