EFFICIENT INTERACTIVE WEIGHT TUNING FOR TTS SYNTHESIS: REDUCING USER FATIGUE BY IMPROVING USER CONSISTENCY

Francesc Alías[†]*, Xavier Llorà*[‡*]*, Lluís Formiga*[†]*, Kumara Sastry*[‡*]*, and David E. Goldberg*[‡*]

[†]Department of Communications and Signal Theory, Enginyeria i Arquitectura La Salle, Ramon Llull University, 08022 Barcelona, Spain.

{falias,llformiga}@salleURL.edu

[‡]Illinois Genetic Algorithms Lab, University of Illinois at Urbana-Champaign, Urbana, 61801 IL.

{xllora,kumara,deg}@illigal.ge.uiuc.edu

ABSTRACT

The quality of corpus-based text-to-speech systems depends on the accuracy of the unit selection process, which in turn relies on the cost function definition. This function should map the user perceptual preference when selecting synthesis units, which is a very difficult task. This paper continues our previous work on fusing the human judgements with the cost function by means of interactive weight tuning. The application of active interactive genetics algorithms mitigates user fatigue by improving user consistency. As a result, the obtained weights generate more natural synthetic speech when compared to previous objective and subjective proposals.

1. INTRODUCTION

The aim of any Text-to-Speech (TTS) system is the generation of synthetic speech from text. The performance of such systems is evaluated by human beings based on the perceived speech quality. Hence, it is essential to somehow embed this subjective criterion into the tuning process of the TTS system for achieving highly natural synthetic speech. The corpus-based or *unit selection* TTS approach is one of the state-of-the-art techniques that try to reach this aim [1]. This method generates the synthetic speech signal by means of the selection and concatenation of recorded speech units. The tuning of the unit selection module is one of the most important processes in getting high quality synthetic speech [2]. The selection process is driven by a cost function [3], which is typically computed as the combination of several weighted sub-costs. A key issue involves the accurate tuning of these weights, that is, mapping the user subjective preferences among candidate units—a complicated task [3, 4]. Several approaches have been proposed for weight training, distinguishing between *i*) hand-tuning [5] and *ii*) machine-driven—purely objective methods [3, 6, 7] or perceptually optimized techniques [4, 8, 9].

*This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF (F49620-03-1-0129).

In a previous work, we introduced genetic algorithms (GA) for tackling the weight tuning problem [10]. This technique overcame the restrictions of classic approaches [3, 6], attaining better results with a feasible computational effort. Nevertheless, this approach, as all the previous techniques, needs to face a key challenge: the reliable estimation of the subjective perception of the speech attributes (i.e. it is very difficult to define a solid perception mapping function). Thus, it is necessary to actually incorporate user preferences for accurately tuning the weights of the cost function. As a first step, we applied a simple interactive genetic algorithm (iGA) for weight tuning, allowing an actual perception-guided adjustment [11]. However, the conducted experiments evidenced two main problems: the tediousness of the process (user fatigue) and the complexity of maintaining a stable comparison criterion throughout the whole process (user consistency), which are weaknesses related to iGAs. Later, *active* iGAs (aiGAs) introduced several advances for combating the user fatigue [12], showing that learning from user interaction and exploiting the learned knowledge to guide the process of collecting user evaluations can greatly reduce the number of evaluations required to achieve high-quality solutions.

This paper focuses on interactive weight tuning processes and how aiGAs can reduce user fatigue by boosting user consistency. Section 2 describes the main features of aiGAs and their application to weight tuning for corpus-based TTS synthesis. Section 3 introduces a new consistency measure related to the *active* iGA paradigm, which allows controlling the user robustness during the tournaments. Section 4 describes the achieved improvements on user consistency and synthetic speech quality, using efficient subjective-based tuning techniques. Finally, the conclusions are presented in section 5.

2. SUBJECTIVE WEIGHT TUNING

The purpose of this section is twofold. First, it reviews some related research topics in subjective weight tuning, and second, it focuses on the surrogate model of the synthetic subjective fitness proposed in [12], as the main pillar of aiGAs.

Table 1. Algorithmic description of the aiGA model [12], where h is the height of the tournaments tree and $\hat{r}(v)$ is the estimated rank for vertex v.

1.	Create an empty directed graph $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$.
2.	Create 2^h random initial solutions ($\mathcal{V}$ set).
3.	Create the hierarchical tournament set $\mathcal{T}$ using the available solutions in $\mathcal{V}$.
4.	Present the tournaments in $\mathcal{T}$ to the user and update the partial ordering in $\mathcal{E}$.
5.	Estimate $\hat{r}(v)$ for each $v \in \mathcal{V}$.
6.	Train the surrogate ε-SVM surrogate synthetic fitness based on $\mathcal{G}$ and $\hat{r}(v)$.
7.	Optimize the ε-SVM synthetic fitness using the compact GA.
8.	Create a $\mathcal{S}'$ set with 2^{h-1} new different solutions, where $\mathcal{V} \bigcap \mathcal{V}' = \emptyset$, sampling out of the probabilistic model evolved by cGA.
9.	Create hierarchical tournament set $\mathcal{T}'$ with $2^h - 1$ tournaments using 2^{h-1} solutions in $\mathcal{S}$ and 2^{h-1} solutions in $\mathcal{V}'$.
10.	$\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{V}'$
11.	$\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{T}'$
12.	Go to 4 while not converged.

Our previous work proved the usefulness of using GAs (as an objective method) and iGAs (as a subjective method) for tackling the weight tuning problem [10, 11], as iGAs allow the fusion of human and computer efforts for problem solving [13, 14]. However, putting the evaluation process into the hands of a user sets up a different scenario when compared to normal optimization tasks [14]. In this sense, we realized that further research was needed to improve the quality of the achieved synthesis and to combat user fatigue. In the quest to address these issues, aiGAs rely on learning from the interaction with the user and anticipate what hypotheses the user may be interested via *educated guesses*, guiding the breading process of new solutions (see table 1). Further details may be found elsewhere [12].

The key element of an aiGA is its synthetic fitness function. The minimal scenario for collecting meaningful domain-independent information from the user is provided by a binary tournament scheme ($s = 2$) [15]. User evaluations introduce a partial order among the solutions presented so far—in this paper, the synthetic representation of the weights configurations. Durant et al.[16] also attempted to ensemble global rankings based on pair-wise comparisons. However, they never explored the model building over the obtained graph to reduce user fatigue by means of *educated guesses* of the user preferences, as later explained. A partial order can be made explicit by using a partial-ordering graph ensemble $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$—as firstly suggested in [12]. A vertex in $\mathcal{V}$ represents the solutions presented to the user, whereas the edges in $\mathcal{E}$ represent the partial-ordering evaluations provided by the user. Given two solutions $\{s_1, s_2\} \in \mathcal{V}$ the user is able to provide three possible outcomes: *i*) $s_1 > s_2$, *ii*) $s_1 < s_2$, and *iii*) $s_1 = s_2$—or *equal/don't know/don't care*. Such a graph $\mathcal{G}$ can be transformed into a normalized graph $\mathcal{G}'$ containing only *bigger than* relations [12]. A global ordering measure may be computed using a heuristic based on two dominance measures, δ and ϕ, inspired by multiobjective optimization [17, 18]. Let $\delta(v)$ be the number of different nodes present on the paths departing from vertex v, and $\phi(v)$ the number of different nodes present on the paths arriving to v. The estimated fitness of a given solution v may be computed as $\hat{f}(v) = \delta(v) - \phi(v)$. Intuitively, the more solutions a vertex v dominates (is *greater than*), the greater the fitness. Otherwise, the more solutions dominate (are *greater than*) a solution v, the smaller the fitness. The final global estimated ranking $\hat{r}(v)$ is obtained sorting the vertex $v \in \mathcal{V}$ by $\hat{f}(v)$. This global estimate is used to train a ε-SVM for creating the synthetic surrogate fitness [12]. By optimizing such a synthetic fitness we can obtain *educated guesses* about the user preferences. In this paper, the optimization step is conducted by a continuous PBIL [19], instead of using compact GA [12], due to the real-valued representation of the weight tuning problem.

3. MEASURING USER CONSISTENCY

Given a normalized partial-ordering graph $\mathcal{G}'$, if a vertex v appears more than once in a path of $\delta(v)$ or $\phi(v)$, then a cycle exists. If such property exists, it represents an inconsistency in the user evaluations. Thus, due to the *greater than* relations, the consistency of the user evaluations can be identified. This property is the basis of the consistency metric proposed in this paper. In order to compute such a measure we need two components: *i*) cycle detection capabilities for a given graph $\mathcal{G}'$ at time t ($\mathcal{G}'^t$), and *ii*) an heuristic to quantify how much inconsistency the detected cycle is introducing.

Let's $\chi(\mathcal{G}'^t)$ be the set of vertex that are part of at least one cycle in $\mathcal{G}'^t$. Then, the consistency of a user at time t, $\kappa(\mathcal{G}'^t, \omega)$ is defined as follows

$$\kappa\left(\mathcal{G}'^t, \omega\right) = 1 - \left(\frac{1}{|\mathcal{V}'^t|} \cdot \sum_{v \in \chi(\mathcal{G}'^t)} \omega_v \right)^{\alpha} \quad (1)$$

where $|\mathcal{V}'^t|$ is the number of vertex in $\mathcal{G}'$ at time t, ω_v the weight of vertex v (not to be confused with the cost function weights), $\chi(\mathcal{G}'^t)$ the vertexes in the cycles detected in $\mathcal{G}'^t$, and α a global scaling factor bigger or equal than 1. Unless noted otherwise, $\omega_v = 1, \forall v \in \mathcal{V}'^t$ and $\alpha = 1$.

The κ measure allows controlling the user consistency during the evolutionary process, avoiding the explicit inclusion of *control points* (i.e. A-B *vs.* B-A comparisons) along the tournaments. Hence, it's an implicit method for speeding up the weight tuning process, helping to increase user consistency. In this paper, the user consistency is measured at time $t = t_f$—or final time. However, measuring user consistency accurately would require an average integration of the consistency measure along the interactive run. This approach is postponed until further research.

4. EXPERIMENTS

The main goals of the experiments were to explore: *i*) the consistency of user evaluations, *ii*) the implications of using

Table 2. Consistency κ plus the absolute enhancement percentage —denoted as (Δ%)— obtained by aiGA.

simple iGA **Phrase**	**Novice User**	**Knowledgeable User**	**Expert User**
"De la seva selva"	0.944	0.855	0.784
"Fusta de Birmània"	0.857	0.769	0.911
"I els han venut"	0.894	0.867	0.731
"Grans extensions"	0.942	0.800	1
***active* iGA** **Phrase**	**Novice User (Δ%)**	**Knowledgeable User (Δ%)**	**Expert User (Δ%)**
"De la seva selva"	1 (5.89)	0.892 (4.30)	1 (27.50)
"Fusta de Birmània"	1 (16.67)	1 (30.01)	1 (9.81)
"I els han venut"	1 (11.91)	1 (15.00)	0.948 (29.76)
"Grans extensions"	1 (6.12)	1 (25.00)	1 (0.00)
Avg. (Δ%)	1 (10.15)	0.973 (18.58)	0.987 (16.77)

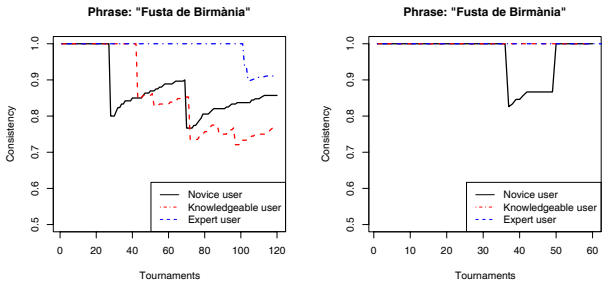
aiGAs in weight tuning for TTS synthesis, and *iii*) the performance of the proposal in terms of perceptual experiments.

We repeated our early experimentation done using the *"Sin-Evo"* platform [11], but replacing the simple iGA with the proposed *active* iGA. The consistency of user evaluations when using both interactive methods was compared by computing $\kappa\left(\mathcal{G}'^{t_f}, \omega\right)$. Then, the synthetic phrases obtained by the tuned weight configurations were presented to naive users. The purpose is to evaluate the synthetic speech quality obtained by the different weight adjustments, validating the impact of the introduced efficient subjective tuning approach, as opposed to previous objective and subjective proposals.

4.1. Objective analysis

First, we measured user consistency via $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ along the evolutionary process presented in our earlier work [11], which used a simple iGA for subjective weight tuning. As shown in upper portion of table 2, only the *expert* user was consistent all the time in a particular experiment. Thus, regardless of the user profile–*novice*, *knowledgeable*, or *expert*—, all the users had troubles maintaining a consistent criterion throughout the tournaments when using a simple iGA (see figure 1(a)). This is basically due to the large number of needful evaluations before converging and the subtle perceptual variations among candidates. Another relevant discovery was that inconsistencies show early on the run of the simple iGA. On average, users, despite of their profile, tend to contradict themselves around tournament 14 across the runs, with a contradiction of 2.83 times per run. Such inconsistencies may be regarded as a noisy subjective fitness function and, hence, be responsible for increasing the number of tournaments required to get a high-quality solution [20]—relieving user fatigue.

Second, this analysis was repeated replacing the simple iGA in *"Sin-Evo"* platform by the proposed aiGA. The results obtained using aiGAs boosted the consistency of the criteria employed by the user, as only two out of the twelve experiments ended in an inconsistent status —$\kappa\left(\mathcal{G}'^{t_f}, \omega\right) < 1$ (see the lower portion of table 2). The *active* selection of tournaments based on the partial-ordering graph $\mathcal{G}'$ helped the user to get back on the track of consistency—see figure 1(b). Another effect of using aiGAs was to drastically cut the number of evaluations required from the user, in a successful effort to combat user fatigue [12]. No special efforts were done to adapt the aiGA model to this particular problem. However, as figure 1 shows, an automatic reduction on the number of required user evaluations was collected. On average, the usage of an off-the-shelf aiGA slashed in half the number of evaluations required to tune weights for the considered phrases, providing a minimum speedup of 2.



(a) simple iGA (b) *active* iGA

Fig. 1. Evolution of the user consistency measured using $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ for the Catalan phrase *"Fusta de Birmània"*. Figures compare the consistency of users when using simple iGA or *active* iGA.

4.2. Subjective evaluation

Finally, we evaluated the acceptance of the synthetic phrases generated from four different weight tuning schemes: aiGA and simple iGA [11]—based on subjective criteria—or Multilinear Regression (MLR) [3] and GA [10]—based on objective measures. The considered cost function [10] takes into account six different sub-costs at diphone level: mean pitch, mean energy and duration *target* costs, plus local pitch, local energy and MFCC *concatenation* costs. After conducting the interactive weight tuning stage described in the previous experiment, each user—*novice*, *knowledgeable* and *expert*—, although being consistent, converged to different weight configurations. For this reason, these configurations needed to undergo a perceptual validation stage. Ten different users—not involved whatsoever in the tuning process—were asked to select the best aiGA weight configuration among the candidates. The configurations proposed by the *expert* user were clearly preferred among the other profiles but no one was discarded completely (44% for the *expert vs.* 27% for the *knowledgeable* and 29% for the *novice* users).

Subsequently, the winning aiGA (waiGA) weight configuration was then compared to the ones obtained by iGA, MLR and GA methods, following a perceptual test. Each subjective test lasted around 15-20 min per user. Figure 2 shows that in all phrases more than a 50% of the users always pre-

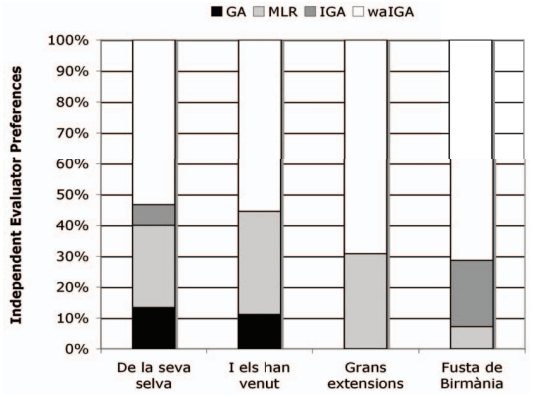


Fig. 2. User preferences among differents methods for cost function weight tuning. Regardless of the phrase analyzed, more than a 50% of the users preferred the synthesis produced by the winning weights obtained by aiGA.

ferred the synthesis produced by the weights obtained using the waiGA. These results represent first empirical evidence on the importance of pursuing such efficient subjective tuning methods. However, a deeper analysis of these results reveals that the greater the difficulty of choosing among the aiGA candidates (number of turns until decision), the lower the degree of acceptance of the waiGA solutions among participants. The smaller the difference among candidates, the harder the problem—a well-known result of the GA literature [21]. Thus, we believe that clustering units may lead to better agreement by helping the evaluators to focus the comparison on particular signal differences [11], instead of conducting global weight tuning.

5. CONCLUSIONS

This paper has continued our work of fusing the human judgement and the weight tuning for corpus-based TTS synthesis. The paper combined a state-of-the-art TTS technique and interactive optimization via *active* interactive genetic algorithms. The use of aiGAs allowed us to evaluate the consistency of user evaluations thanks to partial-ordering graphs and a newly proposed metric. Results shows that aiGAs slashed in half the number of evaluations required to achieve efficient subjectively tuned weights, reducing user fatigue during the tuning process. Moreover, the aiGAs provided better user guidance, drastically boosting the user consistency along the tuning process. The experiments also allowed us to provide sound evidences that efficient subjective weights tuning provide —when compared to previous approaches—a better synthesis acceptance when presented to users.

6. REFERENCES

[1] A.W. Black and K. Tokuda, "Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proceedings of InterSpeech*, Lisbon, Portugal, 2005, pp. 77–80.

[2] A.W. Black, "Perfect Synthesis for all of the people all of the time," in *IEEE TTS Workshop 2002 (Keynote)*, Santa Monica, USA, 2002.

[3] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of ICASSP*, Atlanta, USA, 1996, vol. 1, pp. 373–376.

[4] M. Lee, D. P. Lopresti, and J. P. Olive, "A Text-to-Speech Platform for Variable Length Optimal Unit Searching Using Perceptual Cost Functions," in *4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland, 2001, pp. 75–80.

[5] G. Coorman, J. Fackrell, P. Rutten, and B. Van Coile, "Segment selection in the L&H RealSpeak laboratory TTS system," in *Proceedings of ICSLP*, Beijing, China, 2000, vol. 2, pp. 395–398.

[6] Y. Meron and K. Hirose, "Efficient weight training for selection based synthesis," in *Proceedings of EuroSpeech*, Budapest, Hungary, 1999, vol. 5, pp. 2319–2322.

[7] S. S. Park, C. K. Kim, and N. S. Kim, "Discriminative weight training for unit-selection based speech synthesis," in *Proceedings of EuroSpeech*, Geneva, Switzerland, 2003, vol. 1, pp. 281–284.

[8] H. Peng, Y. Zhao, and M. Chu, "Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation," in *Proceedings of ICSLP*, Denver, USA, 2002.

[9] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing Sub-Cost Functions for Segment Selection Based on Perceptual Evaluations in Concatenative Speech Synthesis," in *Proceedings of ICASSP*, Montreal, Canada, 2004, pp. 657–660.

[10] F. Alías and X. Llorà, "Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis," in *Proceedings of EuroSpeech*, Geneve, Switzerland, 2003, vol. 2, pp. 1333–1336.

[11] F. Alías, X. Llorà, I. Iriondo, X. Sevillano, L. Formiga, and J. C. Socoró, "Perception-Guided and Phonetic Clustering Weight Tuning Based on Diphone Pairs for Unit Selection TTS," in *Proceedings of ICSLP*, Jeju Island, Korea, 2004.

[12] X. Llorà, K. Sastry, D. E. Goldberg, A. Gupta, and L. Lakshmi, "Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness," *Proceedings of Genetic and Evolutionary Computation Conference 2005 (GECCO-2005)*, pp. 1363–1371, 2005, (Also IlliGAL Report No. 2005009).

[13] C. Caldwell and V. S. Johnston, "Tracking a criminal suspect through face-space with a genetic algorithm," in *Proceedings of the Fourth International Conference on Genetic Algorithms*. 1991, pp. 416–421, Morgan Kaufmann.

[14] H. Takagi, "Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation," *Proceedings of the IEEE*, vol. 89, no. 9, pp. 1275–1296, 2001.

[15] D. E. Goldberg, B. Korb, and K. Deb, "Messy genetic algorithms: Motivation, analysis, and first results," *Complex Systems*, vol. 3, no. 5, pp. 493–530, 1989.

[16] E. Durant, G. Wakefield, D. Van Tasell, and M. Rickert, "Efficient Perceptual Tuning of Hearing Aids With Genetic Algorithms," *Trans. IEEE on Speech & Audio Processing*, vol. 12, no. 2, pp. 144–155, 2004.

[17] Carlos A. Coello-Coello, "An updated survey of GA-Based Multiobjective Optimization Techniques," Technical report lania-rd-09-08, Laboratorio Nacional de Informática Avanzada (LANIA), Xalapa, Veracruz, México, December, 1998.

[18] Kalyanmoy Deb, Samir Agrawal, Amrit Pratab, and T. Meyarivan, "A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II," KanGAL report 200001, Indian Institute of Technology, 2000.

[19] Michèle Sebag and Antoine Ducoulombier, "Extending population-based incremental learning to continuous search spaces," *Lecture Notes in Computer Science*, vol. 1498, pp. 418–427, 1998.

[20] B. L. Miller and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," *Complex Systems*, vol. 9, no. 3, pp. 193–212, 1995, (Also IlliGAL Report No. 95006).

[21] D. E. Goldberg, *The design of innovation: Lessons from and for competent genetic algorithms*, Kluwer Academic Publisher, 2002.