# Perceptual Optimization of Unit-Selection Text-to-Speech Synthesis Systems by means of active interactive Genetic Algorithms

Lluís Formiga* and Francesc Alías

Enginyeria i Arquitectura La Salle,
C/Quatre Camins, 20 - 08022 Barcelona
{llformiga,falias}@salle.URL.edu
http://www.salle.URL.edu

**Abstract.** The tuning process of Unit Selection TTS (US-TTS) system is usually performed by an expert that typically conducts the task of weighting the cost function by hand. However, hand tuning is costly in terms of the required training time and inaccurate and ambiguous in terms of methodology. With the purpose of easing the task of properly tuning the weights of the cost function, this thesis make its contribution from a perceptual-based approach using of active interactive Genetic Algorithms (aiGAs). The thesis pursues four major guidelines: *i*) accuracy when tuning the weights, *ii*) robustness of the obtained weights, *iii*) real world applicability of the methodology to any cost function design, and *iv*) finding consensus of the different users when tuning the weights. The experimentation is carried out through a small and medium sized corpus (1.9h) applied to different configurations (type of features) of the US-TTS cost function. The thesis concludes that aiGAs are highly competitive in comparison to other weight tuning techniques from the state-of-the-art.

**Keywords:** speech synthesis,unit selection,weight tuning,interactive evolutionary computation,human computer interaction,latent models

## 1 Introduction

State-of-the-art speech synthesis strategies base their methodology in the use of large speech corpora with the aim to obtain high-quality synthetic speech. Corpus based TTS are considered third generation of TTS systems. The two main corpus-based strategies are Unit-Selection (US) [1] and Hidden Markov Models (HMM) [2]. US-TTS synthesis is based on the waveform concatenation of acoustic units retrieved from very large speech corpora, while HMM-based TTS (hereafter named HMM-TTS) builds an statistical parametric model of the acoustic units from a smaller amount of speech data.

US-TTS systems are able to achieve very high quality synthetic speech, both in terms of naturalness and expressiveness, when the target message matches the acoustic characteristics of the recorded speech. However, the synthetic speech shows a significant quality decrease when the US-TTS system is asked to generate out-of-domain

[*] The thesis document might be downloaded at http://www.tdx.cat/handle/10803/21796 and the presentation slides at http://www.slideshare.net/llformiga/llus-formiga-phd-thesis

or out-of-coverage speech (i.e., far from the speaking styles contained in the speech corpus).

The classical expert-based tuning techniques generally finish the weight optimization problem by finding a *unique* set of weights (weight vector) for the whole corpus, without considering that the relative importance of the costs may vary depending on the contextual and phonetic specificity of the acoustic units when being selected. This issue has been previously identified by the literature related to objective weight tuning [3], but as far as we know this issue has not still been addressed by perceptual weight tuning strategies. Hence, developing a weight tuning strategy that respects the units contexts besides embedding subjective preferences could be of great interest.

This thesis studies an human-aided evolutionary strategy of perceptual weight-tuning that combines active interactive genetic algorithms (aiGA) with a clustering step in order to respect the contextual and phonetic features that present a similar behavior. In particular, the thesis first studies a proof-of-concept of the strategy [4], where the aiGA-based proposal was successfully applied to obtain reliable subjectively tuned weights under a classical unit selection scheme. In a deeper study, the seminal aiGA-based weight-tuning scheme is improved by adapting it to obtain context-dependent weights under a real unit selection scenario (i.e., with a larger corpus and a complex cost function).

This extended abstract is organized as follows. Section 2 reviews the background of a evolutionary-based strategy for tuning the weights of the cost function. Section 3 explains the motivations and the goals pursued by the thesis. Section 4 explains the main contributions of the thesis. Section 5 details how the overall strategy was validated with respect to other baseline perceptual tuning strategies and the discussion and final conclusions are detailed in Section 6.

## 2 Background

The abilities of interactive evolutionary computation are well known when it comes to fuse human and computer efforts with the subjective perception of speech [5]. To reduce user's fatigue in the perceptual tuning, and hence limit the noise at evaluation stage, Llorà et al.[6] defined active interactive Genetic Algorithms (aiGA). This approach allows to discard noisy users or sentences, split the optimization task into a collaborative multi-user task, identify the ambiguity of specific target pairs or obtain a complete rankings of solutions. There are, already, works where aiGA have been used for the perceptual optimization of speech applications (e.g., see [5]).

Figure 1 presents the core execution flow of the proposed aiGA-based method for tuning the cost function weights of US-TTS synthesis systems [4]. Given an initial set of different synthesized versions of the same stimuli (obtained from different weights), these are presented to the user in a tournament hierarchy, then he/she listens to them and selects the winners according his/her subjective criteria. Then, the preferences of the user in front of the proposed synthetic pairs are collected, and a partial-ordering graph is incrementally built by adding this round of user preferences. This graph is used to compute the synthetic fitness function described on the previous section. Once it is available, the continuous-variable based PBIL (cPBIL) optimizes the fitness. The

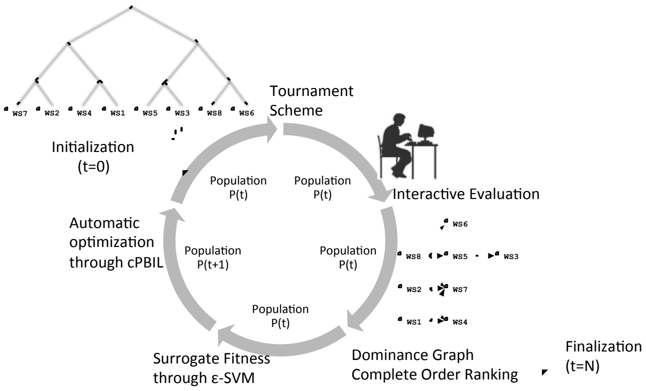

**Fig. 1.** Process diagram of the aiGA-based weight tuning strategy for US-TTS.

important output of this process is a probability distribution over the weight configurations. This probability distribution models the current user preferences towards good solutions. The new round of solutions to be presented to the user is a fifty-fifty combination of previous shown top ranking solutions, and new solutions sampled out of the learned probability distribution—representing promising solutions according to the observed user preferences, also known as *educated guesses*. The process is repeated for three iterations, as our previous work showed it was a good stop criterion [4].

Following this scheme, the pairwise comparisons that would be necessary to establish a complete ranking of the weighting alternatives ($N$) for a particular sentence is dramatically reduced with respect to an exhaustive pairwise comparison scheme ($\frac{N \cdot (N-1)}{2}$), e.g., for $N = 16$, the number of required evaluations is reduced by more than $87\%$ ($15 < 120$).

Several approaches have been introduced in acoustics research for combining human global evaluation based on ranking and automatic optimization schemes [7]. However, these works never addressed the problems of human contradictions and ambiguity. On the other hand, other schemes based in graph theory allow to obtain more complex schemes addressing this human interaction problems and therefore reduce the user fatigue by means of *educated guesses* of the user preferences.

## 3 Motivation and goals: Accuracy, robustness, consensus and real world applicability

The seminal idea of applying an evolutionary strategy for tuning the weights needed a more thorough investigation. Prior to the thesis, the research has bounded to assess the feasibility of the strategy through a small prototype under controlled environment [4]. Concretely, experiments were performed with a extremely small corpus (8 min.) with manual supervision and a single type of unit selection costs (acoustic).

The main objective of the thesis was to design a methodology to perceptual cost function that satisfies the principles of accuracy, robustness, consensus and real world applicability of US-TTS. In this sense, the thesis presents several contributions to meet each of these principles. Once the new methodology is defined the thesis studies its suitability in a real selection scenario making and comparing it with other state-of-the-art tuning strategies, focusing on the competitiveness, efficiency and significance of the final quality of the synthetic speech obtained.

Next, the four important aspects pursued by the thesis are presented.

### 3.1 Accuracy

Prior to the thesis, human-aided tuning was only considered towards obtaining a *unique* weight combination regardless the different phonetic or contextual characteristics where the selection of the units takes place [4]. In contrast, the classical methods for tuning the weights that do not have human intervention, are considered at unit level. That is that a specific, and different, weight configuration is applied when recovering the units with the aim to generate a more naturally sounding voice (e.g. the relative importance of the duration near the stops). Hence, the appropriate weighting should be specified by type of the unit to be selected (e.g. occlusive vs. liquid) [3]. However, under a perceptual approach, it is unfeasible to obtain weights specifically for each unit, and therefore, to obtain fine-grained human-supervised knowledge of the relative importance of each feature. Thus, perceptual approaches only obtain weights with low accuracy but qualitatively good in terms of prediction. This issue arises the first motivation of the thesis: set a methodology that allows to obtain high-quality and fine-grained perceptual weights adapted to the different types of contexts and units that take place within the cost function.

### 3.2 Robustness

*Robustness of automatic optimization:* Among all the state-of-the-art approaches to perform the weight tuning, the most wide-accepted method is to perform a linear regression between the different cost features and a qualitative estimate by means of Cepstral distances [8]. However regression estimates (RMSE and $R^2$) yield poor regression models due to the noise and ambiguity of the task. The thesis makes a thorough study of different data preprocessing schemes in order to overcome the noise and obtain more reliable models. In addition, other optimization techniques such as evolutionary strategies are studied as they are well-known for their robustness with respect to noisy scenarios. Up to date of the thesis, we have not found in the literature exhaustive studies that address the issue of cost function feature normalization.

*Robustness of human-aided optimization* The aiGA prototype [4] highlighted the problem of contradictions of different users yielding to some unreliable models. Therefore, it emerged the need of a methodology for assessing the quality of the solutions provided by the users according to their consistency. Furthermore, the thesis studies other aspects like ambiguity: a user might be consistent, and yet have a passive or heavy hesitant attitude providing excessive evaluation ties compared to other users (evolutionary ambiguity). In addition, the thesis analyzes, the evolution of the user towards a single or multiple criteria and also quantifies the level of consensus among different users.

### 3.3 Real world applicability:

The seminal prototype [4] demonstrated the feasibility of the aiGA only under a controlled environment (8 min. speech) and poorly populated cost function: only F0, energy, duration and spectral coefficients obtained from the recorded signal. In addition, the annotation of the unit selection corpora was fully supervised avoiding classical errors from an automatically annotated corpus. Thus, the prototype scenario was far beyond from a real selection scheme. The thesis fulfills the need of studying the applicability of the aiGA scheme in a real selection scenario with a larger corpus (at least $>$ 1h) and with a cost function that considers different type of features (acoustic, linguistic, contextual...).

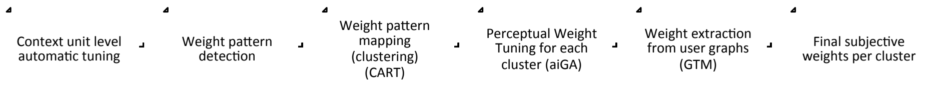
Context unit level automatic tuning
Weight pattern detection
Weight pattern mapping (clustering) (CART)
Perceptual Weight Tuning for each cluster (aiGA)
Weight extraction from user graphs (GTM)
Final subjective weights per cluster

**Fig. 2.** Step diagram of the proposed methodology to tune weights at context-dependent cluster level.

### 3.4 Consensus

When different users evolve the same problem, the integration of the different solutions arises a problem when it exist some contradiction in the criteria (eg. intelligibility vs. similarity to a natural voice). Alías et al. [4] solved this problem by a second pass by the users, where the best solutions obtained by aiGA were competitively selected by other users. In this sense, it was pointed the need of a method able to integrate the different criteria of the users. This issue was also identified by the creators of aiGA [9]. Whereas individual users may disagree about the best solution to the problem, the multiple criteria can be mapped into a single global model, and therefore obtain a consensus model that satisfies the preferences of different users. In this work we study this problem using models with latent variable, which solve the problem of contradictions under a Gaussian non-heuristic adaptive approach.

## 4 Main contributions: Adaptive Weight Tuning through active interactive Genetic Algorithms

### 4.1 Perceptual weight tuning at context-dependent cluster level

The first contribution of thesis (namely aiGAClustered) allows to obtain perceptual weights and respect the phonological, linguistic and contextual specificity of the selected units.

Generally, weight tuning can be addressed at three levels of precision (see figure 3) [1]: at unit level [10, 8] (one weight pattern for each phoneme), at cluster level [4] (one weight pattern per group of units, e.g., fricatives), or at global level [11] (a unique

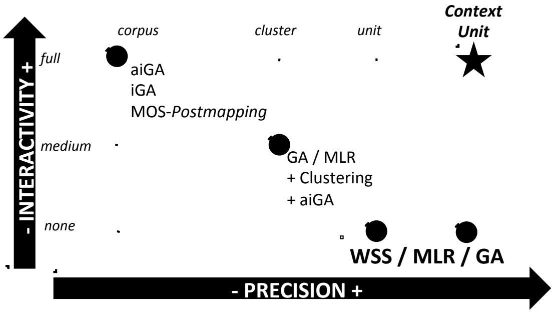

**Fig. 3.** Context-dependent reformulated weight-tuning levels depending on the interactivity and precision offered by the tuning strategy (the star shows the desired setting).

weight pattern for *all* the units contained in the corpus). Generally, the level of adjustment is directly related to the technical difficulty of the tuning process: weights are typically tuned at unit level when the process is automatic and without human intervention (e.g., MLR or GA), the tuning is performed at global level when it is performed perceptually through a reduced set of representative utterances, both for computer-aided [4, 11] and expert-based hand-tuning approaches [12].

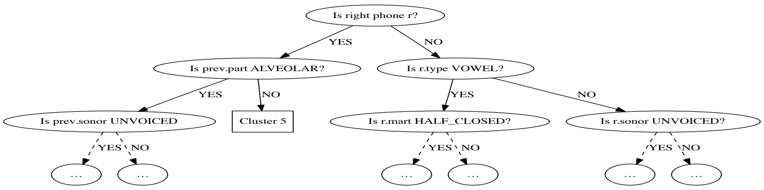

**Fig. 4.** Question set CART-clustering in order to determine weight groups.

Therefore, the thesis considers that similar units can be grouped according to their common behavior (pattern) when tuning weights automatically. Afterwards the weights may be refined through a perceptual tuning stage (see figure 4). It is important to highlight that clusters are only obtained to detect weight behavior patterns, but they do not decide the final weight values of the group as other approaches do [13]. The strategy obtains an intermediate level of accuracy (see figure 3 between global (all units together) and unit level tuning, respecting the inherent specificity of each unit [3]. Hence, weight tuning at cluster level overcomes the drawbacks of fatigue and design complexity that are typically involved in perceptual tuning [4].

The proposed methodology is as follows: *i*) initial weights are obtained by means of an automatic methodology (e.g. GA) – that is a different weight value for each unit in a specific context. *ii*) contextual units with similar weight values are grouped (see figure 4) with a state-of-the-art algorithm (we have chosen Expectation Maximization after a detailed comparative study among other techniques). *iii*) a classification tree (CART) algorithm [10]) is used to map the clusters to the distinct contextual scenarios of the cost function. The CART question set was defined to include phonological (e.g. point of articulation),linguistic (e.g. POS) and contextual information for each half-phone of the corpus. The number of clusters was been constrained by analyzing several impurity measures and selecting the number of clusters that obtained the most favorable validity scores [4].

### 4.2 Generative Topographical Maps to find consensus models

The second contribution of this thesis focuses on the study of integration (consensus) of the different solutions evolved by the users. Previously [4], due to the lack of a methodology to automatically obtain consensus, it was necessary a second pass to choose the best weights among the different solutions. The thesis studies the combination of user preferences at genotype level, i.e. in terms of weight vectors. Although some proposals of aiGA consensus have been proposed at phenotype (graph combination) level [9], it did not solve the problem of breaking the cycles (inconsistencies) formed by the different users, at least without heuristics. The consensus by genotype allows to search for the best solution with a Gaussian adaptive approach that fits to the data (weight values) distribution.

In this sense, this thesis proposes latent models in order to model and group the best preferences of evolving users and therefore extract the best set of weights from the model. The rationale after the latent models is motivated by their capacity to find models beyond linear dependencies of the data and their EM strategy makes them robust to noise. Furthermore, Gaussian latent methods implicitly deal with the contradictions between different users without having to resolve them heuristically. After some experimentation, the thesis concludes that generative topographic maps (GTM) are able to integrate different preferences robustly (thanks to the EM adaptation). That is coherent with previous work carried out in speech processing, as it confirms that Gaussian-EM based algorithms are good to deal with speech problems either for *i*) detecting patterns of weights (with EM) and  em ii) finding consensus models, as GTM is a constrained version of GMM.

### 4.3 Evolutionary Process Indicators

Prior to the thesis [4] there was only a single indicator of consistency defined ($\kappa$) which measured the degree of consistency of user preferences graph model. During the experimentation of the thesis, we observed this measure as insufficient to obtain all the necessary information about the quality of the solutions. Thus, we expanded the aiGA indicators to: three new indicators: *i*) A measure that about the confusion within the graph (Certainty Ratio $\lambda$), *ii*) A measure about the convergence of the perceptual test to a single or multiple solutions (*Intra-user* Convergence Ratio $\rho$) and *iii*) A measure about the consensus of the test conducted by different users (*Inter-user* Correlation Ratio $\tau$). The new indicators allowed to discard runs that had not converged or didn't correlate with other users' runs and detect whether the graph building process had to be stopped prematurely because it was not evolving.

### 4.4 Other contributions

Other remarkable contributions of the thesis that were a byproduct of the performed experimentation are: *i*) a detailed study of different preprocessing, normalization and transformation strategies to obtain more reliable models. *ii*) the demonstration of the capabilities of *aiGAClustered* methodology under a real-world US-TTS scenario with a large training corpus with automatic annotations prone to errors and *iii*) the suitability of the methodology to complex cost functions (with a large amount and different type of cost features). Rigorous experimentation and analysis has been carried out to all the contributions explained.

## 5 Methodology Validation

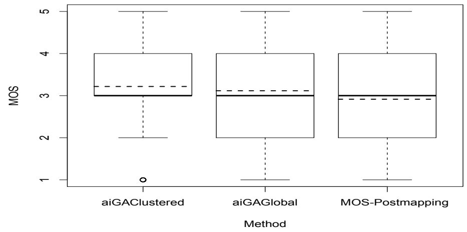

**Fig. 5.** MOS Comparison of different methods of perceptual weight tuning studied.

The weights obtained by the methodology proposed in the thesis (aiGA at context-dependent level) were compared to a perceptual baseline approach (MOS-Postmapping – [11, 11]). MOS-Postmapping tunes a single weight combination for all the different units and contexts in the corpus (global granularity).

Furthermore, we also extracted a single global weight combination from aiGA graphs,i.e. for all contexts and units. This evaluation scheme allowed completing the study in two aspects: *i*) the impact of the clustering strategy to the perceptual tuning methodology (Global aiGA vs. Clustered aiGA), and *ii*) a fair comparison between aiGA and MOS-Postmapping working both at the same global level of tuning.

The validation of the three strategies was conducted with 33 users (including the 8 expert tuning evaluators), considering different levels of speech-related expertise between users throughout 10 randomly sentences and 3 weight configurations (Global aiGA, Clustered aiGA and MOS-Postmapping). That made a total of 30 pairwise comparisons for each user, collecting a total of 990 pairwise annotations and 1980 absolute MOS scores.

Figure 5 depicts the results obtained on the MOS scale considering the preferences of the users. The absolute scores obtained were as follows: $MOS_{aiGAClustered} =$ 3.21, $MOS_{aiGAGlobal} =$ 3.11 and $MOS_{MOS-Postmapping} =$ 2.91. The direct comparison of the different configurations (via a CMOS-like sign analysis of the pair-wised absolute scores) revealed that all differences were significant

In absolute terms, results showed a clear preference for the clustered aiGA methodology as is preferred on 96/241 pairs (39.83%). That number of votes was significantly better than the 74/241 votes (30.70%) obtained by the global aiGA and the 71/241 votes (29.46%) by the MOS-Postmapping method. ($p < 5 \cdot 10^-3$).

## 6 Discussion and Final conclusions

The thesis describes context-dependent methodology based on active interactive genetic algorithms for the perceptual tuning of the cost function weights of TTS systems including unit selection (both US-TTS and hybrid TTS systems). The seminal aiGA-based methodology was improved from the original proof-of-principle scheme (small corpus and simple cost function) to a real unit selection scenario (large corpus and complex cost function). Therefore, the aiGA-based strategy is a robust strategy when combining different kind of weights (linguistic vs.ãcoustic) for different types of costs (discrete vs.c̃ontinuous). In addition the thesis makes two contributions that are crucial to find weights that improve significantly the quality of the synthesized speech: *i*)Evolutionary process indicators of aiGA and *ii*) the application of a consensus modeling method (GTM). Moreover, the experiments demonstrate the importance of considering the context of units when it comes to tune the weights. *aiGAClustered* was preferred in front of other strategies in the perceptual preference tests involving different ways of user-interaction and tuning level grain. Therefore, the context-dependent aiGA-based is an excellent tool for tuning different weights perceptually respecting the contextual and phonetic specificity of the involved units.

The research developed under the framework of this thesis has had a considerable impact in the community involving: 1 peer-reviewed journal [4], 5 conferences related to Speech Processing [14–18], 2 conferences related to Artificial Intelligence [19, 20] and one book chapter [21].

## References

1. Hunt, A., Black, A.W.: Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Volume 1., Atlanta (USA) (1996) 373–376
2. Tokuda, K., Zen, H., Black, A.: An HMM-Based Speech Synthesis System Applied to English. In: Proc. of IEEE Workshop on Speech Synthesis, IEEE (2003) 227–230
3. Campillo, F., Alba, J.L., Rodríguez-Banga, E.: A Neural Network Approach for the Design of the Target Cost Function in Unit-Selection Speech Synthesis. In: Proc. of the 9th International Conference on Speech Communication and Technology (InterSpeech), Lisbon (Portugal), ISCA (2005) 2533–2536
4. Alías, F., Formiga, L., Llorá, X.: Efficient and Reliable Perceptual Weight Tuning for Unit-Selection Text-to-Speech Synthesis based on active interactive Genetic Algorithms: A Proof-of-Concept. Speech Communication **53**(5) (2011) 786–800
5. Alm, C., Llorà, X.: Evolving Emotional Prosody. In: Proc. of 9th International Conference on Spoken Language Processing (ICSLP), Pittsburgh, PA (USA), ISCA (2006) 1826–1829
6. Llorà, X., Sastry, K., Goldberg, D.E., Gupta, A., Lakshmi, L.: Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness. Proc. of the Genetic and Evolutionary Computation Conference (GECCO) (July 2005) 1363–1371

7. Lee, M., Lopresti, D., Olive, J.: A Text-to-Speech Platform for Variable Length Optimal Unit Searching using Perception Based Cost Functions. International Journal of Speech Technology **6**(4) (2003) 347–356
8. Meron, Y., Hirose, K.: Efficient Weight Training for Selection based Synthesis. In: Proc. of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Volume 5., Budapest (Hungary) (1999) 2319–2322
9. Llorà, X., Yasui, N.I., Goldberg, D.: Graph-theoretic measure for active igas: Interaction sizing and parallel evaluation ensemble. In: Proc. of the 10th Conference on Genetic and Evolutionary Computation (GECCO), Nova York (USA), ACM (2008) 985–992
10. Black, A.W., Taylor, P.: Automatically Clustering Similar Units for Unit Selection in Speech Synthesis. In: Proc. of the 5th European Conference on Speech Communication and Technology (EuroSpeech), Rhodes (Greece) (1997) 601–604
11. Chu, M., Peng, H.: An Objective Measure for Estimating MOS of Synthesized Speech. In: Proc. of the 7th European Conference on Speech Communication and Technology (EuroSpeech), Aalborg (Denmark), ISCA (2001) 2087–2090
12. Clark, R., Richmond, K., King, S.: Multisyn: Open-Domain Unit Selection for the Festival Speech Synthesis System. Speech Communication **49**(4) (2007) 317–330
13. Colotte, V., Beaufort, R.: Linguistic Features Weighting for a Text-to-Speech System Without Prosody Model. In: Proc. of the 9th European Conference on Speech Communication and Technology (EuroSpeech), Lisbon (Portugal), ISCA (2005) 2549–2552.
14. Alías, F., Llorà, X., Iriondo, I., Formiga, L.: Ajuste subjetivo de pesos para selección de unidades a través de algoritmos genéticos interactivos. Procesamiento del Lenguaje Natural **31** (September 2003) 75–82
15. Alías, F., Llorà, X., Iriondo, I., Sevillano, X., Formiga, L., Socoró, J.: Perception-Guided and Phonetic Clustering Weight Tuning Based on Diphone Pairs for Unit Selection TTS. In: Proc. of the 8th International Conference on Spoken Language Processing (ICSLP), Jeju Island (South Korea), ISCA (2004) 1221–1224
16. Alías, F., Llorà, X., Formiga, L., Sastry, K., Goldberg, D.E.: Efficient Interactive Weight Tuning for TTS Synthesis: Reducing User Fatigue by Improving User Consistency. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Volume I., Toulouse (France), IEEE (2006) 865–868
17. Monzo, C., Formiga, L., Adell, J., Iriondo, I., Alías, F., Socoró, J.: Adaptación del cth-url para la competición albayzin 2008. In: Actas de las V Jornadas en Tecnologias del Habla. (2008) 87–90
18. Formiga, L., Trilla, A., Alías, F., Iriondo, I., Socoró, J.C.: Adaptation of the URL-TTS system to the 2010 Albayzin Evaluation Campaign. In: Proc. of Fala 2010, VI Jornadas en Tecnología del Habla, Vigo (Spain) (2010) 363–370
19. Formiga, L., Alías, F.: Extracting User Preferences by GTM for aiGA Weight Tuning in Unit Selection Text-to-Speech Synthesis. Lecture Notes in Computer Science - Computational and Ambiental Intelligence **4507** (June 2007) 654–661 Proc. of the 9th International Work-Conference on Artificial Neural Networks (IWANN).
20. Formiga, L., Alías, F., Llorà, X.: Evolutionary Process Indicators for active iGAs Applied to Weight Tuning in Unit Selection TTS Synthesis. In: Proc. of the IEEE Conference on Evolutionary Computation (CEC), Barcelona (Espanya), IEEE (July 2010) 2322–2329
21. Formiga, L., Alías, F.: GTM User Modelling for aiGA Weight Tuning in TTS Synthesis. In: Encyclopedia of Artificial Intelligence. Information Science Reference (2008) 788–795