Perception-Guided and Phonetic Clustering
Weight Tuning Based on Diphone Pairs for Unit Selection TTS

Francesc Alías †, Xavier Llorà ‡, Ignasi Iriondo †,
Xavier Sevillano †, Lluís Formiga †, Joan Claudi Socoró †

†Dept. of Communications and Signal Theory
Enginyeria i Arquitectura La Salle
Ramon Lull University
Psg. Bonanova 8, 08022-Barcelona, Spain
{falias, iriondo, xavis}@salleURL.edu
{liformiga, jclaudi}@salleURL.edu

‡Illinois Genetic Algorithms Lab (IlliGAL) &
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
104 S. Mathews Avenue
Urbana, IL 61801, USA
xllora@illigal.ge.uiuc.edu

Abstract
The quality of corpus based text-to-speech systems depends on
the accuracy of the unit selection process, which relies on
the values of the weights of the cost function. This paper is focused
on defining a new framework for the tuning of these weights.
We propose a technique for taking into account the subjective
perception of speech in the selection process by means of Interactive Genetic Algorithms. Moreover, we introduce a CART-
based method for unit clustering. Both techniques are applied
to weight tuning based on diphone pairs. The conducted experiments analyze the feasibility of both proposals separately.

1. Introduction
A key issue in corpus based text-to-speech (TTS) synthesis is
the tuning of the weights involved in the unit selection cost function [1]. Such tuning determines the synthetic speech quality
achieved. Several approaches have been proposed for weight
tuning, distinguishing between (1) hand-tuning [2] and (2) machine-driven tuning (purely objective approaches [1, 3, 4]
or perceptually optimized techniques [5, 6, 7]). Our previous work used a genetic algorithm for simultaneously adjusting the
target and concatenation weights based on diphone pairs [8].

Moreover, the use of diphones as basic units induces a consid-
erable increase in the size of the search space. It also produ-
ces the appearance of scarcely populated units—rare events
[9]—hindering a reliable unit-dependent weight tuning
[10]. Thus, efficient clustering of the available diphones be-
comes essential for the weight tuning process when compared to
the approaches using phones [1, 3]. This paper presents a novel approach for the tuning of the unit selection cost function
weights. Such process relies on (1) the subjective perception of
humans by means of an Interactive Genetic Algorithm [11], and
(2) a phonetic clustering of the units using CART [12]. The feasi-
bility of such approaches are analyzed throughout the paper,
including empirical validations based on several experiments.

Section 2 presents the proposed method for subjective
weight tuning. Section 3 describes the algorithm for diphone
clustering. The experiments are presented in section 4. Finally,
section 5 discusses some conclusions about the presented work.

2. Perception-guided Weight Tuning
The aim of any TTS system is the generation of speech, whose
naturalness is evaluated by a human being in terms of perceptual
criteria. Hence, tuning methods based on subjective—human—
evaluation is essential for achieving natural sounding synthetic
speech. In a corpus based TTS context, the perceptual component
may be modeled by the subcost functions and their rele-
vance adjustment (weights), among others. Therefore, the qual-
ity of the synthesized speech is highly dependent on their values.

Interactive Genetic Algorithms (IGAs) constitute an opti-
mization model capable of combining the adjustment of quanti-
tative parameters and the subjective evaluation of the results.
IGAs replace the traditional computer-based fitness and selec-
tion scheme [13, 14] by a human-driven selection process. This
kind of algorithms have been employed in several disciplines
to fuse human and computer efforts when subjective evaluation
is a key element [11]. The algorithm evolves a vector of indi-

ciduals \( w = (w_1, \ldots, w_n) \)—the weights of the cost function in
our case—through a two-stage process: (1) the selection of the
best solutions contained in the population, and (2) their poste-
rrior recombination in order to generate new solutions (see figure 1). At each iteration, the IGA generates a set of weights
\( w_i \) in order to synthesize the input text. The result of the
TTS process is interactively evaluated by the user, who is prompted
to choose the best realization between two candidates—using a
binary tournament.

The recombination of genetic material (in our case, the set
of weights) exchanges fragments of the genetic material of two
parents of the selected population. One point crossover operator
[13] has been employed for such purpose. After the recombi-
nation stage, the sets of weights are probabilistically perturbed
[13], in order to simulate errors in the recombination process
(mutation).

3. Diphone clustering
Dividing the unit space into clusters offers an intermediate
level of precision between global (all units together) and unit-
dependent (one weight set per unit) adjustment techniques
[1, 3]. Such approach allows obtaining different weights for
different kinds of units. It also avoids the drawbacks of sparsely
populated units, by means of distributing them among the clus-
ters.
Diphones are clustered according to their phonetic features. For each unit, we take into account its type (vowel, consonant, semivowel, and silence), the sonority (voiced or unvoiced), the manner of articulation (plosive, fricative, etc.) and the place of articulation (bilabial, dental, etc.). The clustering process also aims the creation of well-balanced clusters. The goal is both avoiding predominant and isolated clusters. For achieving such purpose, the clustering method and the optimal number of clusters need to be carefully selected, distributing the units among clusters as uniformly as possible.

The clustering method implemented is based on the Classification and Regression Tree algorithm (CART) [12], which was adapted to solve the categorical diphone clustering problem. CART implicitly deals with sparseness of units [15], obtaining the set combination of phonetic features that best minimizes the entropy of each cluster. After building the clustering tree, a greedy algorithm [15] prunes the nodes until the desired number of clusters (N) is found. The goal is to diversify the weight tuning by having a sufficient number of clusters, and to avoid scarcely or massively populated clusters.

The results of the clustering process were evaluated using a multicriteria approach. Such criteria was based on: (1) the number of units in the least populated cluster (MIN); (2) the number of units in the most populated cluster (MAX); (3) the standard deviation of the number of the units per cluster (STD); (4) the difference MAX-MIN; and (5) the slope of the ordered distribution of units per cluster (SLOPE).

4. Experiments

The experiments have been conducted on a speech corpus in Catalan composed of 1520 sentences (containing around 10000 units). It is to note that the referred corpus has not been intendedly designed for its use in a unit selection TTS system. Hence, not all of the units of the corpus (in this case, diphones) present a number of instances that provide sufficient diversity for test purposes [8].

The conducted experiments intend to (1) adjust the tunable parameters of the designed clustering algorithm according to the described speech database, (2) evaluate its performance in terms of statistical indicators, (3) validate the feasibility (convergence) of the subjective IGA-based weight tuning and (4) compare its results with respect to objective-based approaches.

4.1. Clustering Experiments

The following experiments evaluate three aspects of the clustering process. Firstly, the best phonetic question set is chosen according to the distribution of the units in the corpus. Secondly, the designed CART-based algorithm is compared to two classic clustering methods in order to evaluate the correctness of our approach. And finally, the optimal number of clusters is defined regarding to the statistical multicriteria. The clustering tests have been carried out from 3 to 100 clusters (N).

4.1.1. Choosing the question set

In order to determine the best question set for the CART-based clustering algorithm, all possible combinations of the four kinds of phonetic questions (unit type, sonority, manner and place of articulation) have been tested.

Figure 2 compares the clustering results obtained by CART with 4 questions (CART-4q) against two samples of the four possible combinations of 3 questions (CART-3q), in terms of the number of units in the least populated cluster. The higher the MIN, the better the clustering, given a particular value of N.

After analyzing the results using the previously introduced statistical indicators, we conclude that CART-4q is more stable across the different N cluster values. Thus, 4q configuration is selected for the following experiments. However, CART-3q also offers good performance for small values of N, when the manner of articulation is not deemed, and slightly better results for large values of N, when the sonority is excluded. Moreover, we noticed that the unit type is crucial to obtain a good partition of the search space. CART-3q is not able to find any clustering when N < 8 as the result of excluding the unit type from the clustering process, as figure 2 shows.

4.1.2. Comparison with other clustering methods

The performance of the implemented CART-based clustering algorithm is evaluated by comparison with categorical K-means and Expectation-Maximization (EM) clustering methods provided by the WEKA package [16].

After averaging the results of the K-means and EM clustering methods for 10 different seed initializations, CART attains the best performance throughout the experiment (according to the multicriteria statistics), maximizing the uniformity of the cluster distribution (see figure 3).
4.1.3. Optimal number of clusters

The optimal number of clusters for weight tuning \((N^*)\) is defined as the \(N\) attaining the maximum of MIN and the minimum of MAX, STD, MAX-MIN and SLOPE, i.e. simultaneously avoiding predominant and isolated clusters. Unfortunately, these statistical indicators are insufficient for determining \(N^*\) unequivocally. Hence, the value of \(N^*\) has been selected by means of a heuristic criterion. In the case of our corpus, the optimal number of clusters is \(N^* = 10\), which presents the best statistical multicriteria behavior. The optimal value was, hence, determined as the best trade off of the different statistics used to analyze the results.

Figure 4 presents the resulting splitting tree for the optimal number of clusters. Although the tree has been built by means of CART-4q, notice that only three kinds of phonetic questions are finally used: unit type, sonority and place of articulation.

Thus, as discussed in section 4.1.1, a CART-3q would be sufficient for parting the diphone space into 10 clusters in this case.

4.2. IGA-based weight tuning experiments

This section describes the IGA-based weight tuning process, which was conducted by means of a web-based platform. The developed experiments intend to (1) evaluate the appropriateness of the proposal in terms of its convergence, and (2) compare the obtained weights against two objective methods: multilinear regression (MLR) and genetic algorithms (GAs) [8]. The considered cost function [8] takes into account six different weights: target unit mean pitch (PIT T), target unit mean energy (ENE T), target unit duration (DUR T), concatenation unit local pitch (PIT C), concatenation unit local energy (ENE C) and Mel Frequency Cepstrum (MFC C) at the point of concatenation.

4.2.1. Evaluating the subjective tuning process

As a first step, before facing a larger-scale experimental process, only five phonetically balanced sentences extracted from a television documentary have been selected for IGA-based global weight tuning. The inputs to the synthesis system are the phonetic transcription and the prosody extracted from the target sentences. At each test step, the user must choose the best individual between two candidate sentences (binary tournament), using the documentary sentence as a comparison benchmark.

Several conclusions concerning the test process and the tuning of the developed platform were reported by three expert users after the developed experiments:

- It is complicated to maintain a stable comparison criterion throughout the whole test process. Moreover, the criteria applied by the users seldom coincide.
- The user automatically discards the sentences that have been affected by any error (e.g. a small noise, a wrong phone, ...), although this error might be due to segmentation or labeling failures, and not to the weight set itself.
- Differences between synthesized sentences become extremely subtle after several iterations and the test process turns out to be tedious. This situation can be motivated by (1) a rapid convergence of the IGA or (2) the presence of some sparsely populated units in the corpus.
- Two different speech corpora are used during the process: the television documentary and the corpus for speech synthesis. These corpora were recorded by two different speakers, thus, the prosody information extracted from the first corpus differs from the speech contained in the second corpus. As a future step, both corpora should be recoded by the same speaker, in order to enable more precise rythmic and tonal comparisons.

4.2.2. Comparing IGA with objective methods

The tests were performed using the following parameters: \(pop\_size = 15\), \(p_c = 0.6\), and \(p_m = 0.1\) [13, 14]. The mutation and the crossover probabilities \((p_m\) and \(p_c\), respectively) are increased with respect to the GA approach presented in [8], in order to compensate the notable decrease of individuals in the population \((pop\_size)\) due to the computational constraints of the synthesis process. After conducting the test, it was stated that 7 was the average number of iterations required before perceptual saturation of the users.

![Figure 3: Log-MIN comparison between CART - K-means and CART - EM, where N = 3 corresponds to the left-top point and N = 100 to the bottom-right point of each pair (N = 3 : 100).](image)

![Figure 4: Clustering tree obtained by CART-4q for \(N^* = 10\), indicating the number of units per cluster.](image)
This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF (F49620-03-1-0129), and by the Technology Research, Education, and Commercialization Center (TRECC), at University of Illinois at Urbana-Champaign, administered by the National Center for Supercomputing Applications (NCSA) and funded by the Office of Naval Research (N00014-01-1-0175).

7. References