Exploratory Analysis of Linguistic Data based on Genetic Algorithm for Robust Modeling of the Segmental Duration of Speech

Edmilson Morais        Fábio Violaro
School of Electrical and Computer Engineering
University of Campinas, São Paulo, Brazil
{emorais,fabio}@decom.fee.unicamp.br

Abstract
This work presents a new method for exploratory analysis of linguistic data. This new method is based on Genetic Algorithm and it is used to improve the performance of linear regression models for predicting the segmental duration of speech. The proposed method was compared with Regression Trees and with a baseline Linear Regression model (a Linear Regression with topologies selected using multivariate analysis of variance). The experimental results has shown that the proposed method presents better generalization performance (properties to deal with database imbalance) than the Regression Trees and the baseline Linear Regression model. All the evaluations presented in this article were carried out using an American English database from the Toshiba Speech Technology Laboratory in Cambridge, UK.

1. Introduction
In Text-to-Speech synthesis (TTS) the modeling and estimation of the segmental duration of speech is based on a set of linguistic attributes (linguistic factors) extracted from the sentence to be synthesized [1]. Some of the reasons which make this modeling and estimation a difficult task are [1]:
- Database imbalance (missing data problems)
- Interactions among linguistic factors
- No existence of an elaborated linguistic theory of the rhythm structure of speech

The main goal of this paper is to try to deal with the problem of database imbalance. In order to deal with this problem an exploratory analysis of the linguistic data based on Genetic Algorithm - GA [3] is proposed and it is used to improve the performance of an additive Linear Regression model. It was chosen additive linear regression models due to its simplicity and also because it is a fully tied model [2]. Despite being applied to additive linear regression models the proposed exploratory data analysis could be easily adapted to work with regression models based on Sum-of-products - SoP [4] and Neural Networks [5], for instance.

The proposed exploratory data analysis is used to try to answer to important questions related to the problem of database imbalance. The first one is about the problem of which phones should be modeled together in a single linear regression model? Some authors have proposed the use of category trees, constructed based on a priori linguistic knowledge, to deal with this problem [6]. In this work, GA is used to search, in an efficient way, the whole space of linguistic factors (for all phones) and build up automatically a clusterization tree indicating which phones should be modeled together. The criterium used to select the clusters is the maximization of the overall performance of the linear regression models. The second question is about which linguistic factor should make part of the linear regression models? (And which factor interactions in the case of SoP). GA and a sort of bagging technique [7], that has been called majority rule, are used to answer this question.

The section 2 of this paper presents a description of the database used on the experiments. Details about the linear regression model used in this work are presented on section 3. The proposed exploratory data analysis based on GA is on section 4. The results are on section 5 and finally section 6 presents some final considerations and suggestions for future works.

2. DataBase description
It was used a database from the Toshiba Speech Technology Laboratory in Cambridge, UK. This database is from an American English Male Speaker and it has 1474 sentences, 18172 words, and 60136 phones. All the linguistic factors were automatically labeled and hand-checked. The list with the 14 linguistic factors used is presented in Table 1.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>phiD</td>
</tr>
<tr>
<td>F1</td>
<td>PosInSyll</td>
</tr>
<tr>
<td>F2</td>
<td>PrevPh</td>
</tr>
<tr>
<td>F3</td>
<td>NextPh</td>
</tr>
<tr>
<td>F4</td>
<td>NNextPh</td>
</tr>
<tr>
<td>F5</td>
<td>PoS</td>
</tr>
<tr>
<td>F6</td>
<td>ACC</td>
</tr>
<tr>
<td>F7</td>
<td>NSyll</td>
</tr>
<tr>
<td>F8</td>
<td>DistEnd</td>
</tr>
<tr>
<td>F9</td>
<td>DistStress</td>
</tr>
<tr>
<td>F10</td>
<td>NextStress</td>
</tr>
<tr>
<td>F11</td>
<td>PrevPause</td>
</tr>
<tr>
<td>F12</td>
<td>Chunk</td>
</tr>
<tr>
<td>F13</td>
<td>PosInWord</td>
</tr>
</tbody>
</table>

The set of phones used is described on Table 2.

<table>
<thead>
<tr>
<th>Toshiba</th>
<th>IPA</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ii</td>
<td>iː</td>
<td>ease</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
<td>rise</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
<td>pet</td>
</tr>
<tr>
<td>ae</td>
<td>ae</td>
<td>pat</td>
</tr>
<tr>
<td>aa</td>
<td>aː</td>
<td>calm</td>
</tr>
<tr>
<td>uh</td>
<td>uː</td>
<td>cut</td>
</tr>
<tr>
<td>AR</td>
<td>AR</td>
<td>g(r)</td>
</tr>
</tbody>
</table>

Table 2: List of the phones used in this article

Table 1: List of 14 linguistic factors used. (Distances are given in number of syllables)
4. Exploratory data analysis using GA

In order to improve the robustness of the QMTI models to deal with the database imbalance we have investigated two important problems:

- **Optimal clusterization tree:** Should each phone be individually modeled or should some of them be grouped together and modeled by a single QMTI model?
- **Optimal topology for each QMTI model:** Which linguistic factors should be used in each QMTI model?

Sections 4.1 and 4.2 present the proposed solution for these two problems.

### 4.1. Optimal topology estimation

The optimal solution to this problem is to search over all possible subsets of linguistic factors and to examine all regression equations constructed out of a given list of linguistic factors, along with some measure of fit for each [8]. However, a direct implementation of this method can be computationally very expensive. One of the main proposals of this article is to use GA to estimate the optimal subset of linguistic factors to be used in the regression models without having to explore the whole space of possible combination of linguistic factors. A pseudo-code of the proposed algorithm is presented on Figure 1.

#### Figure 1: Algorithm for optimal topology estimation

Algorithm for optimal topology estimation

For each phone class (one or more phones per classes)

```
BEGIN
for i = 1 to Ni (Ni: number of intermediate topologies) BEGIN
    Partition of the database (training X validation)
    Estimation of the ith intermediate topology
END
Estimation of the optimal topology by majority rule
END
```

Routine for estimation of the ith intermediate topology

```
for i = 1 to Ne (Ne: Number of epochs) BEGIN
    Generation of the initial population
    Evaluation of the fitness function
    Reproduction and selection of individuals
    for j = 1 to Ng (Ng: Number of generations ) BEGIN
        Crossover and mutation
        Evaluation of the fitness function
        Reproduction and selection of the individuals
        END
    IndivEpoch (j) = Best individual in generation j
END
Intermediate Topology = Best individual over all IndivEpoch (j)
```

Some important parts of this algorithm are the cromossomic representation of each individual (topology), the partition of the database into training and validation sets, the objective and fitness function, and the majority rule. Some discussions on these topics are presented on the next subsections.

#### 4.1.1 Cromanomic representation of each individual

It was used a very simple binary representation where each cromosome corresponds to a 14 dimensional binary vector. In this representation, if an allele is equal to 1, then it's...
4.1.2 Partition of the database (Training and validation)

The estimation of each intermediate topology was done using a different partition of the database into training and validation sets. It was used 80% of the data for training and 20% of the data for validation. The aim of using different partitions was to explore the possibility of having different intermediate topologies for different splitting of the data. In other words the aim was to obtain a very diverse space of intermediate topologies.

4.1.3 Objective function and the fitness function

The objective function used is based on the Pearson correlation coefficient [8] between the original durations in the validation set and the predicted durations. The fitness function was estimated from the objective function using a linear smoothing function to reduce the selective pressure of the algorithm [3].

4.1.4 Majority rule

This rule performs a sort of average of the space of intermediate topologies in order to generate the optimal (final) topology of the model. What the majority rule really does is to select to the optimal topology only the linguistic factors that appear in more than 50% of the intermediate topologies.

4.2. Algorithm for phone clustering

In order to deal with problems of sparsing data and also to explore the fact that some phones have some similar properties in terms of their durational structure, it was adopted a procedure based on GA to automatically estimate the phone classes that can be modeled together. This clustering procedure was implemented using a Top-Down hierarchical binary clustering. This procedure allows each cluster to be divided into only two son clusters. This clusterization technique starts with all phones in a single cluster and then divides this cluster in two son clusters. After that these two clusters are divided generating 4 son clusters. This procedure of division/duplication is repeated until the clusterization tree achieves a single phone per cluster. A pseudo code of the algorithm is described on Figure 2.

Some important operations on this algorithm are the cromossomic representation and the definition of the objective function.

4.2.1 Cromossomic representation of each individual

Table 4 shows an example describing the representation used. If the allele of the ith position is equal to 0 then, the ith phone will be classified in the class L (left); otherwise the ith phone will be classified in class R (Right).

Table 3: Cromossomic representation of the topologies

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Selection factors</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>...</th>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>F1</td>
<td>F2</td>
<td>...</td>
<td>F12</td>
<td>F13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2: Algorithm for optimal topology estimation**

**Algorithm for phone clustering**

Initialization: Start with all phones in a single cluster

\[
\text{for } i = 0 \text{ to } N_t \text{ (Number of clusters in level } l) \\
\text{BEGIN} \\
\text{Split the } i\text{th cluster in level } l \\
\text{END} \\
\text{l = l + 1 (Go to next level of the tree) } \\
\text{END}
\]

**Routine for splitting the ith cluster in level l**

\[
\text{for } i = 1 \text{ to } Ne \text{ (Ne: Number of epochs) } \\
\text{BEGIN} \\
\text{Generation of the initial population} \\
\text{Evaluation of the fitness function} \\
\text{Reproduction and selection of individuals} \\
\text{for } j = 1 \text{ to } Ng \text{ (Ng: Number of generations) } \\
\text{BEGIN} \\
\text{Crossover and mutation} \\
\text{Evaluation of the fitness function} \\
\text{Reproduction and selection of the individuals} \\
\text{END} \\
\text{IndivEpoch } (j) = \text{Best individual in generation } j \\
\text{END} \\
\text{BestSpliting } = \text{Best individual over all IndivEpoch } (j)
\]

**Table 4: Cromossomic representation for cluster selection**

<table>
<thead>
<tr>
<th>Phone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>43</th>
<th>44</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cromosomo</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

4.2.2 Objective function

At each division of Cluster into two son clusters, Cluster\textsuperscript{L} and Cluster\textsuperscript{R}, the following objective function was used:

\[
\text{Obj}(\text{Cluster}\textsuperscript{L}, \text{Cluster}\textsuperscript{R}) = 1 - \frac{\text{Corr(Cluster}\textsuperscript{L}) + \text{Corr(Cluster}\textsuperscript{R})}{2}
\]

where \text{Corr(Cluster}\textsuperscript{L}) and \text{Corr(Cluster}\textsuperscript{R}) are the Pearson correlation coefficients for the clusters Cluster\textsuperscript{L} and Cluster\textsuperscript{R} respectively.

5. Results

5.1. Optimal topologies: One phone per model

Figure 3 shows the optimal topologies obtained when considering one phone per regression model. In this Figure the phones are represented over the horizontal axis and the linguistic factors through the vertical axis. If the intersection between phone and linguistic factor is black, then the correspondent factor is part of the model, otherwise the correspondent factor is not part of the model. For example, the model of phone /zh/ is:

\[
dur(zh) = dur(zh) + a_1 \cdot \text{PosInSyl} + a_2 \cdot \text{NextPh}
\]
5.2. Optimal topologies: Phone classes

In order to identify which phones should be grouped and modeled together in a single regression model a bottom up analysis of the whole clusterization tree was performed. The selected clusters (classes of phones) were the ones which presented higher or equal Pearson correlation coefficient than the sum of its correspondent two son clusters.

Figure 4 shows the topologies that were obtained for each phone class. It was obtained 33 phone classes; 24 classes with single phones and 9 classes with 2 or more phones: C01 (aa,sh); C02 (Q,b,ng); C03 (H,z); C04 (AR,OR); C05 (dh,dx); C06 (ccc,u,v,w,y); C07 (g,i,zh); C08 (ae,ii); C09 (ai,oi)

5.3. Overall performance of the method

Figure 5 shows the overall performance of the proposed method (QMTI+GA) when compared to Regression Trees (RT) and a baseline QMTI model with topologies selected using multivariate ANOVA (QMTI+ANOVA).

The experiment showed on Figure 5 was specifically designed to evaluate the performance of the proposed method to deal with database imbalance problems. In order to obtain the results showed on Figure 5 50 experiments were performed. In each of these 50 experiments the database was divided in different training (75%) and validation sets (25%). QMTI+GA, QMTI+ANOVA and RT models were trained using only the training data and evaluated using the validation data. The results showed on Figure 5 are an average over the 50 experiments.

6. Final Considerations

This work presented a new method for exploratory analysis of linguistic data based on GA. This new method was used to improve the performance of a linear method for predicting the segmental duration of speech.

The proposed method uses hierarchical clustering techniques plus GA to identify the phones that should be grouped into a single regression model and it also uses GA to estimate the most important linguistic factors to be used in the each prediction model.

Experiments were performed showing the efficiency of the proposed method when compared to Regression Trees and a baseline linear regression model with topologies selected using multivariate ANOVA.

In future works the authors intend to use the proposed technique to model segmental duration of speech based on Sum-of-Product models.

Acknowledgements

The authors would like to thank the Toshiba Speech Technology Laboratory in Cambridge, UK, for the database used on this work, and in special to Dr. Kate Knill.

References