Report I

Collaboration: Unicamp ↔ Toshiba

Exploratory Analysis of Linguistic Data to Optimal Modelling of the Segmental Duration of American English.

by

Edmilson S. Morais

Sumário

1 PROBLEM DEFINITION ............................................................................................................................................. 1
   1.1 GOALS .............................................................................................................................................................. 1
   1.2 FUNDAMENTAL CONCEPTS: FACTORS, LEVELS AND INTERACTIONS ...................................................... 1

2 METHODOLOGY .......................................................................................................................................................... 3
   2.1 DESCRIPTION, REPRESENTATION AND ANALYSIS OF THE DATA .............................................................. 3
      2.1.1 Description .................................................................................................................................................. 3
      2.1.2 Transformation of the Symbolic Space into a Binary Space ..................................................................... 9
      2.1.3 Data Analysis ......................................................................................................................................... 10
   2.2 ANALYSIS AND ESTIMATION OF THE LINEAR REGRESSION MODELS .................................................... 11
      2.2.1 Type of the Regression Model ................................................................................................................. 11
      2.2.2 Determination of the Optimal Topology ................................................................................................... 12
         2.2.2.a Representation of the Optimal Topologies: One regression model per phone ........................................ 12
         2.2.2.b Objective Function ............................................................................................................................ 12
         2.2.2.c Fitness Function .................................................................................................................................. 13
         2.2.2.d Crossover, mutation, selection e reproduction ..................................................................................... 13
         2.2.2.e Majoritary Rules .................................................................................................................................. 13
         2.2.2.f Computacional Cost ........................................................................................................................... 14
      2.2.3 Clusterization ............................................................................................................................................. 15
         2.2.3.a Representations ..................................................................................................................................... 15
         2.2.3.b Objective Function ................................................................................................................................ 16
         2.2.3.c Fitness Function ................................................................................................................................... 16
         2.2.3.d Crossover, Mutation, Selection and Reproduction ................................................................................ 16
         2.2.3.e Selection of the Best Cluster ................................................................................................................ 16
         2.2.3.f Optimal Topology Determination per Cluster ...................................................................................... 17

3 RESULTS ................................................................................................................................................................. 17
   3.1 ONE PREDICTION MODEL PER PHONE ........................................................................................................... 17
      3.1.1 Topologies ................................................................................................................................................... 17
      3.1.2 Performance of the Majoritary Rule ......................................................................................................... 20
      3.1.3 Majority Rule Versus ANOVA .................................................................................................................. 23
      3.1.4 Majority Rule Versus CART ..................................................................................................................... 26
   3.2 ONE PREDICTION MODEL PER CLUSTER ....................................................................................................... 28
      3.2.1 Clusterization Tree ..................................................................................................................................... 28
      3.2.2 Topologies ................................................................................................................................................... 36
   3.3 FINAL CONSIDERATIONS ................................................................................................................................. 37
   3.4 SUGGESTIONS FOR FUTURE WORK ................................................................................................................ 38

REFERENCES ............................................................................................................................................................. 38

APENDIX A ............................................................................................................................................................... 39

APENDIX B ............................................................................................................................................................... 42
1 Problem Definition

1.1 Goals

The aim of this work is to present a new method for analysis and prediction of the segmental duration of speech. This new method must perform robust/trustfull mappings from a symbolic space (linguistic features: phonological, syntactical, morphological, ...) into a continuous interval (phone durations). The method developed will be usefull for both purely linguistic analysis and speech synthesis systems such as TTS (Text to Speech Synthesis) and CTS (Concept to Speech Synthesis).

The main techniques used on the development of this new method are:

- Genetic Algorithm.
- Hierarchical Clustering.
- Regression Analysis and Prediction Techniques.
- Principal Component Analysis.
- Boosting and Bagging techniques.
- CART - Classification and Regression Tree
- ANOVA – Analysis of Variance. (MANOVA - Multivariate Analysis of Variance).

Some of these techniques are described on the Appendix A.

1.2 Fundamental Concepts: Factors, Levels and Interactions

All the training examples (used to train the prediction model) present the following format:

\[ O(i) = \langle \text{Phone}, F_1, F_2, F_3, \ldots, F_N, \text{DurPhone} \rangle \]

The term Phone represents the identity of the phone to be model, DurPhone is a scalar number which represents the duration (in milliseconds) of the phone to be model. The terms \( F_i \) are vectors and henceforth will be called factors. The values assumed for each factor \( F_i \) are denominated level. For example, if \( F_i \) represents the factor “Grammatical Category”, then:

\[ \text{Grammatical Category} \iff F_i = \langle \text{Verb, Noun, Adjective, \ldots, SyllablePosition} \rangle \]

The factorial space given by the factors \( F_1, F_2, F_3, \ldots, F_N \) will be defined \( F \). \( F \) will be given by the factorial product:

\[ F = F_1 \times F_2 \times F_3 \times \ldots \times F_N \]
The segmental duration interval, for all possible duration values, will be given by D. The interval D belongs to the axis of real number, \( \mathbb{R} \) and it will be limited by a minimal and a maximal value. In other words:

\[
D \subset \mathbb{R} \quad \text{and} \quad \forall \; \text{dur} \in D, \; \text{dur}_{\text{min}} \leq \text{dur} \leq \text{dur}_{\text{max}}.
\]

The aim of the segmental duration model is to estimate a function DUR able to map any value in the factorial space F into the duration interval D.

\[
\text{DUR} : F \rightarrow D
\]

In the case of an additive model, this function can be decomposed in N terms \( A_i \):

\[
A_i : F_i \rightarrow D
\]

Combining all these terms \( A_i \) we can obtain:

\[
\text{DUR}(F) = \sum_{i=1}^{N} A_i(F_i)
\]

In an simple additive model (or simple linear regression model), each term \( A_i \) is a vector of parameters. The dimension of each vector \( A_i \) is given by the number of level of the factor \( F_i \). For example if the factor \( F_i \) represents the factor Stress = \{HighAccented, Accented, NoAccented\}, then \( A_i(\text{High Accented}) = 20 \text{ms} \), \( A_i(\text{Accented}) = 5 \text{ms} \), \( A_i(\text{NoAccented}) = -25 \text{ms} \). All these parameters can be estimated directly from the training samples. The contribution of each parameter vector \( A_i \) to the model DUR(F) is defined as the effect of the factor \( F_i \).

In an additive model with interactions between factors (or linear regression model with interactions) some extra terms \( A_{ij} \) are added. In the case of considering only interaction between two factors (not between three, four or more factors) we have:

\[
\text{DUR}(F) = \sum_{i=1}^{N} A_i(F_i) + \sum_{i=1}^{J} \sum_{j=1}^{K} A_{ij}(F_i \times F_j)
\]

\( A_{ij} \) is a matrix of parameters, in which the number of components is equal to the number of levels given by the factorial \( F_i \times F_j \). The indexes \( J \) and \( K \) belongs to the interval, \( 1 \leq J \leq N \) and \( 1 \leq K \leq N \).
2 Methodology

2.1 Description, Representation and Analysis of the Data

This section makes some considerations about the data used on the training of the regression models (segmental duration models).

2.1.1 Description

All the data used on the experiments are from an standard american male speaker.

Each example of the training data has the following format

\[ O(i) = \langle \text{Phone}, F_1, F_2, F_3, \ldots, F_N, \text{DurPhone} \rangle \]

The term \( \text{Phone} \) indicates the phone to be model. \( \text{DurPhone} \) represents the duration in miliseconds of the phone to be model. The terms \( F_i \) represents linguistic parameters (syntactic, phonological, morphological, ...) and are denominated \( \text{fatores} \). The \( \text{fatores} F_i \) are vetors, and elements in \( F_i \) are denominated \( \text{niveis} \). The Tabela 1 presents a description of the 14 \( \text{fatores} \) used on the experiments. In some experiments the phoneme identity (\( \text{Ph} \)) will not be used as a \( \text{fator} \). In some other experiments all the 14 \( \text{fatores} \) will be necessary.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Fatores</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ph</td>
<td>( \text{Ph} )</td>
<td>Identity of the phone to be model.</td>
</tr>
<tr>
<td>F1</td>
<td>pos syll</td>
<td>Position of the phone to be model in relation to the accented syllable of the current word.</td>
</tr>
<tr>
<td>F2</td>
<td>Prev ph</td>
<td>Class of the previous phone.</td>
</tr>
<tr>
<td>F3</td>
<td>Next ph</td>
<td>Class of the next phone.</td>
</tr>
<tr>
<td>F4</td>
<td>Next next ph</td>
<td>Class of the next next phone.</td>
</tr>
<tr>
<td>F5</td>
<td>PoS</td>
<td>Part-of-Speech tagging.</td>
</tr>
<tr>
<td>F6</td>
<td>Pitch</td>
<td>Accent level of word which contains the phone to be model.</td>
</tr>
<tr>
<td>F7</td>
<td>NSyll</td>
<td>Number of the syllables in the current word.</td>
</tr>
<tr>
<td>F8</td>
<td>Dist End</td>
<td>Distance in syllable to the end of the current word.</td>
</tr>
<tr>
<td>F9</td>
<td>NAcc</td>
<td>Distance to the next accented syllable (in the next word)</td>
</tr>
<tr>
<td>F10</td>
<td>Next pause</td>
<td>Distance in syllable to the next pause.</td>
</tr>
<tr>
<td>F11</td>
<td>Prev pause</td>
<td>Distance in syllable from the previous pause.</td>
</tr>
<tr>
<td>F12</td>
<td>Chunk</td>
<td>Distance to the end of the accent group in which the phone to be model belongs.</td>
</tr>
<tr>
<td>F13</td>
<td>Pos Syll</td>
<td>Position, in the current word, of the syllable which contains the phone to be model.</td>
</tr>
</tbody>
</table>

The Table 2 presents a list with all the phonemes used with their respective IPA representation.
Table 2: Table with the list of phonemes used and their respective IPA symbols.

<table>
<thead>
<tr>
<th>Toshiba</th>
<th>IPA</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ii</td>
<td>iː</td>
<td>ease</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
<td>peak</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
<td>pet</td>
</tr>
<tr>
<td>ac</td>
<td>æ</td>
<td>pat, path</td>
</tr>
<tr>
<td>A</td>
<td>a</td>
<td>pot</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>oo</td>
<td>oː</td>
<td>cause</td>
</tr>
<tr>
<td>uu</td>
<td>uː</td>
<td>lose</td>
</tr>
<tr>
<td>u</td>
<td>ū</td>
<td>put</td>
</tr>
<tr>
<td>ER</td>
<td>ɛɹ</td>
<td>bird, mother</td>
</tr>
<tr>
<td>@</td>
<td>o</td>
<td>allow</td>
</tr>
<tr>
<td>ai</td>
<td>ai</td>
<td>rise</td>
</tr>
<tr>
<td>au</td>
<td>aʊ</td>
<td>house</td>
</tr>
<tr>
<td>oi</td>
<td>oɪ</td>
<td>noise</td>
</tr>
<tr>
<td>ei</td>
<td>ei</td>
<td>raise</td>
</tr>
<tr>
<td>ou</td>
<td>oʊ</td>
<td>nose</td>
</tr>
<tr>
<td>AR</td>
<td>a(ɾ)</td>
<td>far</td>
</tr>
<tr>
<td>OR</td>
<td>ɔ(r)</td>
<td>port</td>
</tr>
<tr>
<td>p</td>
<td>p</td>
<td>pin</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>tin</td>
</tr>
<tr>
<td>k</td>
<td>k</td>
<td>kin</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>bin</td>
</tr>
</tbody>
</table>

The Table 3 presents some training samples for the phone @.

Table 3: Format of the training data <Phone, F₁, F₂, F₃, …, F₉, DurPhone>

<table>
<thead>
<tr>
<th>Ph</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>A9</th>
<th>A10</th>
<th>A11</th>
<th>A12</th>
<th>A13</th>
<th>Dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>@</td>
<td>mid</td>
<td>UC</td>
<td>UC</td>
<td>ShortVowel</td>
<td>prep</td>
<td>deac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>@</td>
<td>Aft</td>
<td>VC1</td>
<td>UC</td>
<td>UC</td>
<td>N</td>
<td>High</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td>10</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>@</td>
<td>pre</td>
<td>UC</td>
<td>UC</td>
<td>ShortVowel</td>
<td>N</td>
<td>Acc</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>@</td>
<td>aft</td>
<td>VC1</td>
<td>Uplosive</td>
<td>LongVowel</td>
<td>N</td>
<td>Acc</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>@</td>
<td>mid</td>
<td>UPlosive</td>
<td>UC</td>
<td>VC1</td>
<td>Vi</td>
<td>Acc</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>@</td>
<td>mid</td>
<td>UC</td>
<td>Uplosive</td>
<td>ShortVowel</td>
<td>freq</td>
<td>deac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>Start 48</td>
</tr>
<tr>
<td>@</td>
<td>aft</td>
<td>UPlosive</td>
<td>None</td>
<td>None</td>
<td>N</td>
<td>Acc</td>
<td>2</td>
<td>1</td>
<td>none</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>End 124</td>
</tr>
<tr>
<td>@</td>
<td>aft</td>
<td>VC1</td>
<td>VC1</td>
<td>Uplosive</td>
<td>adv</td>
<td>Acc</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>@</td>
<td>pre</td>
<td>UPlosive</td>
<td>UC</td>
<td>ShortVowel</td>
<td>N</td>
<td>Acc</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>@</td>
<td>mid</td>
<td>UC</td>
<td>UC</td>
<td>LongVowel</td>
<td>prep</td>
<td>deac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>@</td>
<td>aft</td>
<td>VC1</td>
<td>VC1</td>
<td>Closure</td>
<td>adj</td>
<td>Acc</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>End 30</td>
</tr>
<tr>
<td>@</td>
<td>pre</td>
<td>UC</td>
<td>VC1</td>
<td>Diphthong</td>
<td>adv</td>
<td>Acc</td>
<td>2</td>
<td>2</td>
<td>none</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>Start 53</td>
</tr>
<tr>
<td>@</td>
<td>mid</td>
<td>UC</td>
<td>VC1</td>
<td>ShortVowel</td>
<td>freq</td>
<td>deac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>@</td>
<td>aft</td>
<td>UPlosive</td>
<td>VC1</td>
<td>None</td>
<td>nam</td>
<td>Acc</td>
<td>2</td>
<td>1</td>
<td>none</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>End 46</td>
</tr>
<tr>
<td>@</td>
<td>aft</td>
<td>VC1</td>
<td>VC1</td>
<td>UC</td>
<td>nam</td>
<td>Acc</td>
<td>3</td>
<td>2</td>
<td>none</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>Middle 20</td>
</tr>
<tr>
<td>@</td>
<td>mid</td>
<td>UC</td>
<td>VC1</td>
<td>Diphthong</td>
<td>Vb</td>
<td>deac</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>@</td>
<td>pre</td>
<td>UC</td>
<td>Uplosive</td>
<td>VC1</td>
<td>adj</td>
<td>Acc</td>
<td>3</td>
<td>2</td>
<td>none</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>Middle 45</td>
</tr>
<tr>
<td>@</td>
<td>mid</td>
<td>UC</td>
<td>ShortVowel</td>
<td>freq</td>
<td>deac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>start 49</td>
</tr>
<tr>
<td>@</td>
<td>mid</td>
<td>UPlosive</td>
<td>UC</td>
<td>LongVowel</td>
<td>prep</td>
<td>deac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

Campinas, 20 of December, 2004
All the possible levels assumed for each of the 14 factors from Table 1 are described on the Table 4.

Table 4: All possible levels that each of the 14 factors from Table 2 can assume.

<table>
<thead>
<tr>
<th>Ph</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
<th>F8</th>
<th>F9</th>
<th>F10</th>
<th>F11</th>
<th>F12</th>
<th>F13</th>
</tr>
</thead>
<tbody>
<tr>
<td>@,AR,ER,H,OR,Q,aa,ae,ai,au,b,ccc,ch,d,dh,dx,e,ei,ii,jh,k,l,m,n,oi,oo,ou,p,r,s,sh,t,th,u,uh,uu,v,w,y,z,zh</td>
<td>pre, mid, aft, non</td>
<td>ShortVowel, LongVowel, Diphthong, VC1, VC2, VPlosive, UPlosive, Closure, UC, Sil, none</td>
<td>ShortVowel, LongVowel, Diphthong, VC1, VC2, VPlosive, UPlosive, Closure, UC, Sil, none</td>
<td>ShortVowel, LongVowel, Diphthong, VC1, VC2, VPlosive, UPlosive, Closure, UC, Sil, none</td>
<td>n, nam, adj, adv, itf, deny, dig, pron2, vi, vs, vt, vb, NULL, w, pnc, nud, int, prep, freq</td>
<td>deacc, acc, high</td>
<td>0,1,2,3,4,5,6,7,8,9,10</td>
<td>0,1,2,3,4,5,6,7,8,9,10</td>
<td>0,1,2,3,4,5,6,7,8,9,10</td>
<td>0,1,2,3,4,5,6,7,8,9,10</td>
<td>0,1,2,3,4,5,6,7,8,9,10</td>
<td>0,1,2,3,4,5,6,7,8,9,10</td>
<td>start, middle, end</td>
</tr>
</tbody>
</table>

If the phoneme identity (Ph) is also considered as a factor, then the factorial space \( F = Ph \times F_1 \times F_2 \times F_3 \times \ldots \times F_{14} \) will have the following dimension

\[
\dim(F) = 44 \cdot 4 \cdot 11 \cdot 11 \cdot 11 \cdot 11 \cdot 11 \cdot 11 \cdot 19 \cdot 3 \cdot 5 \cdot 5 \cdot 5 \cdot 5 \cdot 3 = 7 \cdot 10^{12}
\]

The Table 5 shows the number of training samples for each of the 44 phones.

Table 5: Number of training samples for each of the 44 phones.

<table>
<thead>
<tr>
<th>Phone</th>
<th>Num of training samples</th>
<th>Phone</th>
<th>Num. Of training samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>zh</td>
<td>89</td>
<td>p</td>
<td>1153</td>
</tr>
<tr>
<td>oi</td>
<td>132</td>
<td>ai</td>
<td>1242</td>
</tr>
<tr>
<td>dx</td>
<td>316</td>
<td>w</td>
<td>1353</td>
</tr>
<tr>
<td>AR</td>
<td>320</td>
<td>uh</td>
<td>1360</td>
</tr>
<tr>
<td>ch</td>
<td>323</td>
<td>Q</td>
<td>1459</td>
</tr>
<tr>
<td>jh</td>
<td>331</td>
<td>ccc</td>
<td>1519</td>
</tr>
<tr>
<td>au</td>
<td>387</td>
<td>ER</td>
<td>1581</td>
</tr>
<tr>
<td>th</td>
<td>405</td>
<td>e</td>
<td>1592</td>
</tr>
<tr>
<td>sh</td>
<td>483</td>
<td>k</td>
<td>1622</td>
</tr>
<tr>
<td>g</td>
<td>494</td>
<td>r</td>
<td>1667</td>
</tr>
<tr>
<td>oo</td>
<td>497</td>
<td>m</td>
<td>1711</td>
</tr>
<tr>
<td>OR</td>
<td>515</td>
<td>ae</td>
<td>1770</td>
</tr>
<tr>
<td>y</td>
<td>570</td>
<td>dh</td>
<td>1831</td>
</tr>
<tr>
<td>aa</td>
<td>676</td>
<td>z</td>
<td>1860</td>
</tr>
<tr>
<td>ou</td>
<td>772</td>
<td>d</td>
<td>1940</td>
</tr>
<tr>
<td>uu</td>
<td>815</td>
<td>ii</td>
<td>1941</td>
</tr>
<tr>
<td>u</td>
<td>981</td>
<td>l</td>
<td>2287</td>
</tr>
<tr>
<td>ei</td>
<td>987</td>
<td>s</td>
<td>2771</td>
</tr>
<tr>
<td>H</td>
<td>1034</td>
<td>t</td>
<td>3400</td>
</tr>
<tr>
<td>v</td>
<td>1107</td>
<td>(w)</td>
<td>3730</td>
</tr>
<tr>
<td>f</td>
<td>1137</td>
<td>n</td>
<td>3858</td>
</tr>
<tr>
<td>b</td>
<td>1139</td>
<td>i</td>
<td>4326</td>
</tr>
</tbody>
</table>

Total number of examples: 59483
Figure 1 gives the amount of training samples per phone in percentage. For example, the number of training sample of the phone /@/ is equivalent to almost 6.2% of all training samples in the whole database.

![Figure 1: Amount of training samples per phone, given in percentage.](image)

Figure 2 shows the histograms of each of the 44 phones used. Each histogram below has a legend with following information:
- Phoneme identity,
- Mean,
- Standard deviation,
- Number of samples.

Some important observations about the histograms are:
- Some of them looks like a gamma distribution. See phones “@” and “i” for example.
- Some of them looks like to be bimodal (or even trimodal). See phones “r” and “p” for example.
- Looks like evident that some of them has missing data problem. See phones “ah”, “oi”, “jh”, and “AR” for example.
Exploratory Analysis of Linguistic Data to Optimal Modelling of the Segmental Duration of American English

Campinas, 20 of December, 2004
Figure 2: Histogram of the duration of each of the 44 phones.
2.1.2 Transformation of the Symbolic Space into a Binary Space

Before building the regression models, all the training samples need to be converted in a binary sequence. For example, if example of the training data the level of factor F2 is assumes the value UPlosive then the correspondent binary vector associated to this factor will be:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>F2</th>
<th>B2</th>
<th>ShortVowel</th>
<th>LongVowel</th>
<th>Diphthong</th>
<th>VC1</th>
<th>VC2</th>
<th>Vplosive</th>
<th>UPlosive</th>
<th>Closure</th>
<th>UC</th>
<th>Sil</th>
<th>none</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simbólico</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Numérico</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Figure 3 shows a binary representation of all the factors for all the 676 examples of the training samples of the phone /aa/. In Figure 3, the Factor Ph (identity of the phone) was also expanded to its binary representation (first 44 columns of the Figure).

![Figure 3: Binary representation of the 676 training samples of the phone 'aa'.](image)

Given the binary representation above (which is highly sparse), a simple linear regression model, can be written as:

\[
\begin{bmatrix}
    b_{0} & b_{1} & \cdots & b_{138} \\
    \vdots & \vdots & \ddots & \vdots \\
    b_{676,0} & b_{676,1} & \cdots & b_{676,138}
\end{bmatrix}
\begin{bmatrix}
    a_{0} \\
    a_{1} \\
    \vdots \\
    a_{138}
\end{bmatrix}
= \begin{bmatrix}
    d_{0} \\
    d_{1} \\
    \vdots \\
    d_{676}
\end{bmatrix}
\]

Therefore, the total number of coefficients to be estimated is equal to 138 (total number of levels of each of the 13 factors including the 44 levels representing each phoneme identity)
This simple linear regression model, which initially transforms the symbolic space \( F \) into a binary space \( B \) and then maps this binary space into a point into the durational interval \( D \), will be denominated QMTI (Quantification Method Type I). In other words, the QMTI method consists of two stages:

- To map the symbolic space into the binary space \( F \rightarrow B \)
- To estimate a duration model \( DUR(F) \), using a simple linear regression model (simple additive model): \( DUR(F) \Leftrightarrow DUR(B) = \sum_{i=1}^{N} A_i(B_i) \). Where \( B_i \) is the equivalent binary representation of \( F_i \).

- To determine the mapping \( A_i \) (regression coefficients) using the Pseudo Inverse technique (See Appendix A).

### 2.1.3 Data Analysis

A dimensionality reduction of the binary space can be obtained using Principal Components Analysis PCA. The PCA can map any vector from the binary space into a new space where the dimension of the vectors can be made smaller than the original one (173 dimension) without any loss of information or with a controlled loss of information (for example 2% or 18% of lost information). The Figure 4 shows a PCA applied to the vectors of the Figure 3 (phone “aa”).

**Figura 4**: Training samples after PCA. The red lines correspond to dimension with 0% of lost (77 coefficients), 2% of lost (55 coefficients) and 18% of lost (27 coefficients).
The application of PCA to the examples of the phone “aa” shows that a significant reduction of the original space can be obtained without any loss of generality. The use of PCA allows a reduction from 173 dimensions to 77 dimensions without any loss of information. Moreover, all the vectors in the PCA domain are orthogonal to each other and this property can be very useful in the prediction model. However, after the use of PCA we do not have the concepts of Factors and Levels anymore. The original factors and levels are transformed in coefficients which have no linguistic meaning anymore. Due to that in this first Report we have decided not to use PCA in our models/experiments.

2.2 Analysis and Estimation of the Linear Regression Models

Two basic strategies were adopted in this work to predict the segmental duration of each of the 44 phone of Table 2.

1. Estimation of 44 regression models. One regression model per phone.
2. Estimation of $N$ regression models, with $1 \leq N \leq 44$. In this case some phone needs to be grouped (clustered) in a simple model.

We can arise four important questions from these two strategies:

1. Type of the model – What is the type of model to be used (linear or no-linear model) ?
2. Topology – Once having defined the regression model to be used, then what factors, level needs to be used for a robust estimation of the segmental durations.? 
3. Clustering – If we decide to use $N$ regression models, $1 \leq N \leq 44$, then, how to define which phone must be grouped in a single model ?
4. Evaluation – How to evaluate the results obtained here?

The next sections presents the solutions and their respective methodologies, proposed for the four questions listed above:

2.2.1 Type of the Regression Model

It was analysed two type of regression model. The first was the linear regression model (simple additive model) including only the main 14 linguistic factors:

$$DUR(F) = \sum_{i=1}^{14} A_i (F_i)$$

The second possibility was the linear regression model with interactions (additive linear model with iteration) including not only the main 14 main linguistic factors but also some interactions between factors:
In this report only the *linear regression model* were analysed. Despite the choice of using the *linear regression model*, all the procedures and results in this work can be, without any lost of generality, extended to more complex model.

In this work all the *linear regression model* where implemented using the QMTI technique.

### 2.2.2 Determination of the Optimal Topology

The determination of the optimal topology for the regression model were done using genetic algorithms (GA). The procedure adopted can be divided in the following steps:

1. Representation of the optimal topologies;
2. Objective function;
3. Fitness function;
4. Operations: Crossover, mutation, selection e reproduction;
5. Majoritary rule, Bagging and Boosting;
6. Computational costs.

#### 2.2.2.a Representation of the Optimal Topologies: One regression model per phone

It was used a very simple binary respresentation. Each cromossomo with 13 dimensions (13 *factors*, it was not considered the phoneme identity). Allelos equal to 1 indicates the presence of the respective factor. Allelos equal to 0 indicates the absence of the respective factor. For example, the following cromossomo :

<table>
<thead>
<tr>
<th>Cromossomo</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatores</td>
<td>F1</td>
<td>F2</td>
<td>F3</td>
<td>F4</td>
<td>F5</td>
<td>F6</td>
<td>F7</td>
<td>F8</td>
<td>F9</td>
<td>F10</td>
<td>F11</td>
<td>F12</td>
<td>F13</td>
</tr>
</tbody>
</table>

It corresponds to the following topology for the regression model.

<table>
<thead>
<tr>
<th>Fatores selecionados</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F10</th>
<th>F11</th>
<th>F12</th>
</tr>
</thead>
</table>

#### 2.2.2.b Objective Function

The objective function was defined as :

\[ 1 - R^2 \]
Where $R^2$ is the sample multiple correlation between the expected segmental duration $d = \{d_1, d_2, ..., d_N\}$ and the predicted duration $\hat{d} = \{\hat{d}_1, \hat{d}_2, ..., \hat{d}_N\}$.

\[
R^2 = \frac{\left(\sum_{i=1}^{n} d_i \cdot \hat{d}_i\right)^2}{\left(\sum_{i=1}^{n} d_i^2\right) \cdot \left(\sum_{i=1}^{n} \hat{d}_i^2\right)}
\]

In an attempt to improve the generalization capabilities of the model, the data was split into a training set (80% of the data) and a cross-validation set (20% of the data). It was used a different splitting between (training X cross-validation) for each training epoch of the GA.

### 2.2.2.e Fitness Function

The fitness function is a transformation of the objective function into a linear function. This linear function varies from 0 to 2, in other words, the smallest objective value is mapped to 0 and the largest value is mapped to 2. Therefore, the fitness function performs a kind of smoothing of the objective function. This smoothing decreases the selective pressure of individuals of the population (solutions to the optimal topology of the regression model).

### 2.2.2.d Crossover, mutation, selection and reproduction

It was used a single point crossover with a rate of 80%. It was used a binary mutation with a rate of 0.01%. The selection method was the roulette wheel. The reproduction rate was 60%.

### 2.2.2.e Majoritary Rules

In the item 2.2.2.b, it was said that each epoch used a different partition between training and cross-validation data. The aim of using different splittings was to verify if the GA could generate different optimal topologies for each splitting. This solution for each splittings will be called henceforth, intermediate solutions. However, want we want is a single final topology as solution to the problem, therefore it was necessary to define a method to combine all this intermediate topologies into a single one. It was defined a simple rule called, majoritary rule, which selects only the factors that appears in more than 50% of the intermediate solutions. This majoritary rule can be understood as a sort of leveraging or bagging technique. This procedure is explained bellow:
<table>
<thead>
<tr>
<th>Nº of epochs</th>
<th>Intermediate solutions : (different training and validation sets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 0 0 1 1 1 0 0 0 0 0 1 1 1 0 1</td>
</tr>
<tr>
<td>2</td>
<td>0 1 1 1 1 0 0 0 0 0 0 1 1 0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>0 1 0 1 1 0 0 0 0 0 0 1 1 1 0 0</td>
</tr>
<tr>
<td>4</td>
<td>0 0 1 1 0 0 0 1 0 0 1 1 1 0 1 0</td>
</tr>
<tr>
<td>5</td>
<td>0 0 1 1 1 0 0 0 0 0 0 1 1 1 1 0</td>
</tr>
<tr>
<td>6</td>
<td>1 1 0 1 1 0 0 0 0 0 0 1 1 1 0 1</td>
</tr>
<tr>
<td>7</td>
<td>0 0 1 1 1 0 0 0 0 0 0 1 1 1 0 1</td>
</tr>
<tr>
<td>Sum</td>
<td>1 3 5 7 6 0 0 1 0 6 7 6 1 5</td>
</tr>
<tr>
<td>Mean</td>
<td>0.14 0.42 0.71 0.85 0.0 0.14 0.0 0.85 1 0.85 0.14 0.71</td>
</tr>
<tr>
<td>Majority rule.</td>
<td>0 0 1 1 1 0 0 0 0 1 1 1 1 0 1</td>
</tr>
</tbody>
</table>

2.2.2.f Computacional Cost

The major computational cost of the proposed algorithm is associated to the estimation/evaluation of the fitness function for each individual (candidate to the optimal topology of the model). The main reason for this so high cost is the estimation of the objective function. The evaluation of the objective function requires the estimation of the prediction coefficients by the solution of the Pseudo-Inverse method.

In the implemented algorithm, the following operation were performed for each algorithm:

- 50 epochs - Each epoch used a different splitting of the training and cross-validation data.
- 15 generations - The algorithm presented a quick fast convergency.
- 50 individuals - Some tests were done to define 50 individuals as a good number.

Therefore, taking 44 phones, the number of necessary evaluation of the fitness function is equal to:

\[ 44 \times 50 \times 15 \times 50 = 1,650,000. \]

Meanwhile the combinatorial space given by all possible topologies for 44 phones and 50 epochs is given by:

\[ 44 \times 50 \times 2^{13} = 18,022,400 \]

This numbers show that the space searched by our algorithms is equivalent to 0.15% of the whole combinatorial space. This results may not look like so expressive, but if one consider that our algorithm took 2 days to run this experiment, then we can assume that an algorithm explore the whole space in running in the sample computational platform could take 22 days.

Another important consideration is related to the increasing of the number of factors. Instead of using 13 \textit{fatores} it was used 20 \textit{factores}, then the combinatorial space of all possible topologies, for the 44 phones and 50 epochs should became:

\[ 44 \times 50 \times 2^{20} = 2,306,867,200 \]
In this case the searched space corresponds to only 0.0715% of the whole combinatorial space and instead of taking 2 days running the algorithm, one could take 2,797 days.

2.2.3 Clusterization

Given the regression problem treated in this work, we can make two suppositions about the training data and the behaviour of the segmental durations:

- *First suposition*: Maybe the training samples for all phones do not present a good coverage (uniform coverage) of the typical linguistic space and of the durational interval of the american language. In other words, maybe there are some “holes” in important regions of the available distributions. For example, maybe some typical and important durational observations of the american English for the $i^{th}$ phone, are not present in the training data. It is important not to forget that the size of the whole factorial linguistic space is of the order of $10^{12}$.

- *Second suposition*: Maybe some phones can present similar durational behaviours. For example, maybe the process of mapping from the linguistic space to the durational interval for two phones $fone_i$ and $fone_j$ can be similar.

Based on the two above suposition, maybe it is possible to minimize the missing data problem (*first suposition*) grouping and modeling some phones with similar durational behaviour (*second suposition*) together. In order to verify such suposition, it was proposed a method base on GA to automatically group some phones together. This grouping consiste on a Top-Down hierarchical clustering. The main steps of the algorithm are described bellow:

1. Representation;
2. Objective function;
3. Fitness function;
4. Operations: Crossover, mutation, selection e reproduction;
5. Selection of the better cluster;
6. Determination of the optimal topologies for each cluster.

2.2.3.a Representations

It was adopted a Top-Down hierarchical clustering. Moreover, the technique adopted only allows each cluster to be splited up into two other clusters. The procedure starts with all data in a single cluster and then split it into two clusters. After that this two clusters are then divided into two other cluster. This process of splitting/duplication continues untill each terminal cluster has only one phone. The representation adopted was the following:

<table>
<thead>
<tr>
<th>Fone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>44</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cromossomo</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
If the allelo of $i^{th}$ position is equal to 0, then the $i^{th}$ phone must be classified in the class ZERO. If the allelo of $i^{th}$ position is equal to 1, then $i^{th}$ phone must be classified in the class ONE.

### 2.2.3.b Objective Function

The objective function used in each division (splitting), was given by the average of the complementary correlation function $1 - R^2$ for the derived clusters

$$Error(Cluster_1, Cluster_2) = \frac{(1 - R^2_{Cluster_1}) + (1 - R^2_{Cluster_2})}{2} = 1 - \frac{(R^2_{Cluster_1} + R^2_{Cluster_2})}{2}$$

Where $R^2$ is the the sample multiple correlation between the expected segmental duration $d = \{d_1, d_2, \ldots, d_N\}$ and the predicted duration $\hat{d} = \{\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_N\}$.

$$R^2 = \frac{\left( \sum_{i=1}^{n} d_i \cdot \hat{d}_i \right)^2}{\left( \sum_{i=1}^{n} d_i^2 \right) \cdot \left( \sum_{i=1}^{n} \hat{d}_i^2 \right)}$$

Different from the method used on the optimal topology estimation (optimal factors for the regression model), the clusterization method did not use any splitting of the data into training and cross-validation. It was used all the available data both for training and cross-validation. In other words, all the models were evaluated using the same data used for training.

### 2.2.3.c Fitness Function

Similar to the item 2.2.2.c, it was used a fitness function that maps the values from the objective function into values over a linear function. The linear function used for fitness varies from 0 to 2, in other words, the smaller fitness value is mapped to 0 and the largest value is mapped to 2. Therefore, the fitness function performs a kind of smoothing of the objective function. This smoothing decrease the selective pressure of individuais of the population (solutions to the optimal topology of the regression model).

### 2.2.3.d Crossover, Mutation, Selection and Reproduction

It was used crossover of single point with and rate of 70 %. It was used binary mutation with a rate of 0.01% and the selection uses rollete weel. The reproduction rate was 60 %.

### 2.2.3.e Selection of the Best Cluster

It was used 100 individuals, 50 epoch and 15 generation per each splitting of the classes.
2.2.3.3 Optimal Topology Determination per Cluster

The process of determining the optimal topologies per cluster was the same used to generate the linear regression model per phone on section 2.2.2. In summary:

- Use of QMTI + GA + Pseudo-inversa to generate the models;
- Division of the data into training and cross-validation data. Each division generates an intermediate topology;
- Application of the Majority rule to determine the optimal model.

3 Results

In this section the results are organized in the following manner:

- One prediction model per phone
  - Topologies obtained using QMTI + Majority Rule and QMTI + ANOVA
  - Performance: QMTI + Majority Rule X QMTI Full Model
  - Performance: QMTI + Majority Rule X QMTI +ANOVA X QMTI Full model
  - Performance: QMTI + Majority Rule X Decision Trees

- Hierarchical Clustering
  - Cluster tree obtained using GA
  - Topologies of the models using QMTI + Cluster + Majority rule
  - Performance: QMTI + Cluster + Majority Rule X QMTI + Majority Rule.

3.1 One Prediction Model per Phone

3.1.1 Topologies

The Figures 5 and 6 show the model topologies obtained with the use of GA plus Majority rule. The factors which make part of the topologies for each model (phone) is indicated with a black color (white color indicates that such factor doesn’t make part of the model) For example, for the phone /zh/ we have:

\[ \text{DUR}(zh) = a_1 \cdot \text{posSyll} + a_2 \cdot \text{nextPh} \]
Figura 5: Topologies obtained using GA + Majoritary Rule.

Figura 6: Frequency of occurrence of each of the 13 factors for all the 44 models of the Figure 5. The three most frequent ones are prevPh, nextPh and Pitch.
The Figures 7 and 8 show the model topologies obtained with the use of ANOVA. The factors which make part of the topologies of each model (phone) are indicated with a black color (white color indicates which factors don’t make part of the model) For example, for the phone /zh/ we have:

\[ \text{DUR(zh)} = a_1 \cdot \text{posSyll} + a_2 \cdot \text{prevPh} + a_3 \cdot \text{nextPh} \]

**Figura 7:** Topologies obtained using ANOVA.

**Figura 8:** Frequency of occurrence of each of the 13 factors for all the 44 models of the Figure 7. The three most frequent ones are: nextPh, prevPh and posSyll.
3.1.2 Performance of the Majoritary Rule

The Figures 9, 10 and 11 present the 50 intermediate topologies obtained for each the phones /AR/, /b/ and /@/. Each of these 50 topologies were obtained using a different partitions of the training and validation data set. The aim was to maximize the diversity of possible topologies. Each of these intermediate topologies are biased for their specific splitting training/validation. The majoritary rule was applied on these 50 intermediate topologies in order to generate a single optimal topologies which is more robust to missing data than any of the 50 intermediate topologies.

Figura 9: Topologies obtained for the phone /AR/. Each topology corresponds to a different partition of the data set (training and validation sets).

Applying the majoritary rule we have:

\[ \text{DUR}(\text{AR}) = a_1 \cdot \text{nextPh} + a_6 \cdot \text{nextnextPh} + a_8 \cdot \text{PoS} + a_9 \cdot \text{Pitch} + a_{11} \cdot \text{NSyll} + a_{13} \cdot \text{DistEnd} + a_{13} \cdot \text{PrevPause} + a_{13} \cdot \text{PosSyll} \]

Figura 10: Topologies obtained for the phone /b/. Each topology corresponds to a different partition of the data set (training and validation sets).

Applying the majoritary rule we have:

\[ \text{DUR}(\text{b}) = a_2 \cdot \text{prevPh} + a_5 \cdot \text{nextPh} + a_6 \cdot \text{Pitch} + a_{11} \cdot \text{NSyll} + a_{13} \cdot \text{DistEnd} \]
Figura 11: Topologies obtained for the phone /@/. Each topology corresponds to a different partition of the data set (training and validation sets).

Applying the majoritary rule we have:

\[
DUR@ = a_1 \cdot posSyll + a_2 \cdot prevPh + a_3 \cdot nextPh + a_4 \cdot nextnextPh + a_5 \cdot Pitch + a_6 \cdot NSyll + a_7 \cdot DistEnd + a_8 \cdot NAcc + a_9 \cdot nextPause + a_{10} \cdot prevPause + a_{11} \cdot PosSyll
\]

Figures 12, 13 and 14 show the performance of the topology obtained with the use of majoritary rule in comparison with the full topology (using all the factors). In Figures 12, 13 and 14 the models were evaluated 30 times. Each of them corresponds to a different partition of the data set (in terms of training and validation sets). The red line corresponds to the majoritary rule topology. The green line corresponds to the full topology. The blue line corresponde to interval where the results of each of the 50 intermediate topologies, takes place. The top of the blue line corresponds to the best result among the 50 intermediate topologies and the botton of the blue line corresponds to the worst result among the 50 intermediate topologies. Through these Figures we can visualize two important results:

- The performance of the Majoritary rule topology is superior than the performance of the full topology.
- The performance of the Majoritary rules topology is, in general, superior than the average of the performance of the 50 intermediate topologies used in the Majoritary rule. Moreover, the Majoritary rule performance in many times is similar to the performance of the best topologies among the 50 intermediate topologies (It can be seen where the red line almost touch the top of the blue lines).

The majoritary rules have also presented a better performance than the Full model, for all the other 41 phones.
**Report:** Exploratory Analysis of Linguistic Data to Optimal Modelling of the Segmental Duration of American English

**Campinas, 20 of December, 2004**

**Figura 12:** Performance: Majority rule topology vs Full topology. (Phone /AR/)

**Figura 13:** Performance: Majority rule topology vs Full topology. (Phone /b/)
3.1.3 Majoritary Rule Versus ANOVA

Figures 15,16,17,18 and 19 are similar to Figures 12, 13 and 14, but them also includes the results obtained with topologies optimized using ANOVA. The red line corresponds to the majority rule topology. The green line corresponds to the full topology. The blue line corresponds to ANOVA topology. The vertical blue lines correspond to the interval where the results of each of the 50 intermidiate topologies takes place. Through these Figures we can visualize two important results:

- The performance of the Majority rule topology is superior than the performance of the ANOVA topology.
- The performance of the ANOVA topology is, in general, superior than the performance of the Full topology.

The majoritary rules also presentes a better performance than the ANOVA model for all the other 39 phones.
Figura 15: Performance: Majoritary rule topology X ANOVA X Full topology. (Phone /AR/)  

Figura 16: Performance: Majoritary rule topology X ANOVA X Full topology. (Phone /b/)  

Campinas, 20 of December, 2004
Figure 17: Performance: Majority rule topology X ANOVA X Full topology. (Phone: @)

Figure 18: Performance: Majority rule topology X ANOVA X Full topology. (Phone: jh)
3.1.4 Majoritary Rule Versus CART

Figures 20 and 21 show the performance of the topology obtained with the use of majoritary rule in comparison with the CART (Classification and Regression Trees). Figures 20 presents the results for all 44 phones when the training set was equal to 80% of the data and the validation set was equal to the remaining 20% of the data. Figure 21 shows the case when the training set was equal to the validation set. Through these Figures we can visualize two important results:

- The performance of the QMTI + Majoritary rule topology is superior than the performance of the CART when the validation set is different from the training set. In other words the QMTI + Majoritary rule is better to deal with unseen data than the CART model.

- The performance of the CART is superior than the performance of the Majoritary rule topology when the training set is equal to the validation set. In other words the CART is better to fit the training data, but is not good to generalize unseen data.

**Figura 19:** Performance: Majoritary rule topology X ANOVA X Full topology. (Phone /y/)

CAMPINAS, 20 DE DEZEMBRO, 2004
**Report**: Exploratory Analysis of Linguistic Data to Optimal Modelling of the Segmental Duration of American English

Campinas, 20 of December, 2004

---

**Figure 20**: Performance: Majoritary rule topology X CART for all 44 phones. Training set equal to 80% of the data and validation set equal to 20% of the data.

**Figure 21**: Performance: Majoritary rule topology X CART for all 44 phones. Training set equal to validation set (both are equal to 100% of the data).
3.2 One prediction model per Cluster

3.2.1 Clusterization Tree

Table 6 shows the clusters obtained for all the 44 phones:

List of the 44 phone:

- @ H Q b ccc d dh dx e g i l m n r t u uh v w y z zh AR ER OR aa ae ai au ch ei f ii jh k oi oo ou p s sh th uu.

<table>
<thead>
<tr>
<th>Table 6: Cluster Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.1.1 ( \rightarrow @ H Q b ccc d dh dx e g i l m n r t u uh v w y z zh ) &amp; C.5.16 ( \rightarrow th )</td>
</tr>
<tr>
<td>C.1.2 ( \rightarrow AR ER OR aa ae ai au ch ei f ii jh k oi oo ou p s sh th uu ) &amp; C.5.17 ( \rightarrow ER )</td>
</tr>
<tr>
<td>C.2.1 ( \rightarrow H Q b e g i l m n r t u uh z zh ) &amp; C.5.18 ( \rightarrow oo )</td>
</tr>
<tr>
<td>C.2.2 ( \rightarrow @ ccc d dh dx r u v w y ) &amp; C.5.19 ( \rightarrow aa jh k p sh )</td>
</tr>
<tr>
<td>C.2.3 ( \rightarrow ER aa ae f ii jh k oo p s sh th ) &amp; C.5.20 ( \rightarrow f )</td>
</tr>
<tr>
<td>C.2.4 ( \rightarrow AR OR ai au ch ei ou uu ) &amp; C.5.21 ( \rightarrow ch ei ou )</td>
</tr>
<tr>
<td>C.3.1 ( \rightarrow H Q b n t z ) &amp; C.5.22 ( \rightarrow au )</td>
</tr>
<tr>
<td>C.3.2 ( \rightarrow e g i l m uh zh ) &amp; C.5.23 ( \rightarrow AR oi )</td>
</tr>
<tr>
<td>C.3.3 ( \rightarrow @ ccc d u v w y ) &amp; C.5.24 ( \rightarrow OR )</td>
</tr>
<tr>
<td>C.3.4 ( \rightarrow dh dx r ) &amp; C.5.25 ( \rightarrow H )</td>
</tr>
<tr>
<td>C.3.5 ( \rightarrow ae ii s th ) &amp; C.6.2 ( \rightarrow Z )</td>
</tr>
<tr>
<td>C.3.6 ( \rightarrow ER aa f jh k oo p sh ) &amp; C.6.3 ( \rightarrow Q )</td>
</tr>
<tr>
<td>C.3.7 ( \rightarrow au ch ei ou uu ) &amp; C.6.4 ( \rightarrow b )</td>
</tr>
<tr>
<td>C.3.8 ( \rightarrow AR OR ai oi ) &amp; C.6.5 ( \rightarrow i )</td>
</tr>
<tr>
<td>C.4.1 ( \rightarrow H t z ) &amp; C.6.6 ( \rightarrow zh )</td>
</tr>
<tr>
<td>C.4.2 ( \rightarrow Q b n ) &amp; C.6.7 ( \rightarrow m uh )</td>
</tr>
<tr>
<td>C.4.3 ( \rightarrow g i zh ) &amp; C.6.8 ( \rightarrow e )</td>
</tr>
<tr>
<td>C.4.4 ( \rightarrow e l m uh ) &amp; C.6.9 ( \rightarrow u )</td>
</tr>
<tr>
<td>C.4.5 ( \rightarrow @ ccc u v w y ) &amp; C.6.10 ( \rightarrow w )</td>
</tr>
<tr>
<td>C.4.6 ( \rightarrow d ) &amp; C.6.11 ( \rightarrow @ v y )</td>
</tr>
<tr>
<td>C.4.7 ( \rightarrow dh dx ) &amp; C.6.12 ( \rightarrow ccc )</td>
</tr>
<tr>
<td>C.4.8 ( \rightarrow r ) &amp; C.6.13 ( \rightarrow aa jh p sh )</td>
</tr>
<tr>
<td>C.4.9 ( \rightarrow ae ii ) &amp; C.6.14 ( \rightarrow k )</td>
</tr>
<tr>
<td>C.4.10 ( \rightarrow s th ) &amp; C.6.15 ( \rightarrow ch ou )</td>
</tr>
<tr>
<td>C.4.11 ( \rightarrow ER oo ) &amp; C.6.16 ( \rightarrow ei )</td>
</tr>
<tr>
<td>C.4.12 ( \rightarrow aa f jh k p sh ) &amp; C.6.17 ( \rightarrow AR )</td>
</tr>
<tr>
<td>C.4.13 ( \rightarrow au ch ei ou ) &amp; C.6.18 ( \rightarrow oi )</td>
</tr>
<tr>
<td>C.4.14 ( \rightarrow uu ) &amp; C.7.1 ( \rightarrow m )</td>
</tr>
<tr>
<td>C.4.15 ( \rightarrow AR OR oi ) &amp; C.7.2 ( \rightarrow uh )</td>
</tr>
<tr>
<td>C.4.16 ( \rightarrow ai ) &amp; C.7.3 ( \rightarrow v y )</td>
</tr>
</tbody>
</table>
Figure 20 shows the hierarchical binary cluster tree which corresponds to the Table 6. The tree has 9 levels which are differentiated by colors (see correspondence with Table 6). In Figure 20 each node of the tree has the following information:

- **First line**: Cluster index, as indicated in table 6 above.
- **Second line**: Sample mean (for all samples in that cluster).
- **Third line**: Sample standard deviation (for all the samples in the cluster).
- **Fourth line**: Correlation coefficient.
Figura 22.a: First part of the hierarchical binary cluster tree.
Figure 22.b: Second part of the hierarchical binary cluster tree.
Figura 22.c: Third part of the hierarchical binary cluster tree.
3.2.2 Topologies

The hierarchical binary cluster tree, Figure 22, was used only to indicate potential phone candidates to be clustered. The procedure to derive which clusters must be used in our final models is the following:

1. Derived the Majority rule topology for all the nodes of Figure 22.
2. Calculate the correlation factor for all the nodes of the Figure 22 using the Majority rule models.
3. Use a bottom-up approach to select the best cluster. Started from the terminal nodes (bottom of the tree) and follow to the root of the tree looking for the cluster (nodes) which have higher correlation factor than the average of the correlation factor of its correspondent fones separately.

The selected clusters were the following:

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
</tr>
</thead>
<tbody>
<tr>
<td>g + I + zh</td>
<td>dh + dx</td>
<td>H + z</td>
<td>Q + b</td>
<td>AR + oi</td>
<td>aa + sh</td>
<td>ae + ii</td>
</tr>
</tbody>
</table>

Figure 23 shows the correlation factor obtained for the selected clusters versus the average of the correlation factor obtained for their correspondent phones separately.

![Figure 23: Correlation factor](image)

**Figure 23:** Erros de predição dos métodos QMTI original, QMTI original + cluster, QMTI + GA e QMTI + GA + cluster.

Figure 24 shows the results obtained for all clusters selected and and the remaining phones. These topologies corresponds to the Majority rule topology. The factors to be used on the topologies (phone or cluster) are indicated with a black color (white color indicates factors not used) For example, for the cluster “dh+dx” we have:

\[
\text{Dur(dh+dx)} = a_2 \cdot \text{phID} + a_1 \cdot \text{posSyll} + a_2 \cdot \text{prevPh} + a_3 \cdot \text{nextPh} + a_4 \cdot \text{Pitch}
\]
3.3 Final Considerations

This work presents a new method for analysis and prediction of the segmental duration of speech. This new method of prediction performs robust mappings from the linguistic space (symbolic space) to the limited and continuous interval of the segmental durations (phone durations).

The proposed method uses clusterization techniques plus genetic algorithm to identify which phones must be model together and after that uses QMTI + GA to determine the optimal topology of the regression models. In fact the proposed method consist of a new technique to perform an exploratory analysis of the available data and to determine the optimal estimator for the segmental duration of one specific data base (How many models and which are their optimal topologies).

Several of the performed experiments confirm the superiority of the proposed technique when compared to exploratory data analysis using ANOVA (Analysis of Variance) and regression techniques using CART (Classification and Regression Trees).

Our results confirm some initial supositions:

- CART is a no-linear method (despite been a peace-wise linear method) therefore is was expected that it presents a better performance to fit the training data (training set equal to validation set).
Pruning techniques for CART, mainly when applied for regression problems are not very efficiently. Due to that, the generalization capabilities of the CART are not good enough to deal with missing data problems.

When trained and evaluated with all the data (training set equal to validation set) the QMTI with full model (all factors) are superior than the QMTI + GA (with selected factors).

3.4 Suggestions for Future Work

To use regression models with interaction between factors, for example:

\[ \text{DUR}(\mathbf{F}) = \sum_{i=1}^{N} A_{ij}(F_i) + \sum_{i=1}^{J} \sum_{j=1}^{K} A_{ij}(F_i \times F_j) \]

Where \( A_{ij} \) is a matrix of parameters which the number of components (levels) is equal to the factorial \( F_i \times F_j \). The indexes \( J \) and \( K \) belongs to the interval, \( 1 \leq J \leq N \) e \( 1 \leq K \leq N \).

The clusterization technique presented in this work doesn’t use the majoritary rule during the construction of the binary cluster tree. The binary cluster tree is used only to indicate possible candidates to be clustered. The clustering process is performed in a bottom-up fashion using the majoritary rule models. A more interesting solution could be the use the majoritary rule model directly during the construction of the tree.

To try to build a superpositional rhythmic of the speech combining the technique proposed in this work with the concepts of “stress group” and “VV” units proposed by Barbosa [Barbosa, 2002]

References


APENDIX A

Genetic Algorithms

Genetic Algorithms (GA) are stochastic search algorithms inspired on natural evolution process and it is largely used in combinatorial optimization problems. GA consiste of a population of individuals (solutions to an optimization problem) which are evolves through generations. A measure of fitness of each individual (which depends on the problems to be treated) is used to select each individual will survive to the next generation. Operations of crossover and mutation are also applied to the individuals in order to guarantee the diversity of the population. It is expected that after each generation the fitness of the whole population will increase and that one of the individuals, the one with higher fitness, will represent an optimal solution to the optimization problem.

Good references to GA can be found on:

CART – Classification and Regression Trees

A very natural and intuitive way to perform pattern classification (or symbolic regression) is by means of a sequence of well elaborated questions. This type of classification/regression is particularly usefull for data which are no-numeric (symbolic). In this case it is possible to elaborate questions like: yes/no, true/false, belong-to/no-belong-to, etc.

A sequence of questions can be organized in a form of a decision/regression tree. The algorithm CART (Classification and Regression Tree) is a procedure to “grow” decision/regression trees automatically from data. Some important questions about the algorithm CART are:

- How many splittings should occurs at each node;
- Which are the optimal questions to be applied to each node;
- If the tree gets to big, how to make it simpler. In other words, how to prune the tree;
- How to deal with missing data problems.

In this work it was used the CART algorithm from the Matlab toolbox. It was also used standard Matlab techniques to perform the appropriated pruning of the trees.

Figure 25 shows the CART tree obtained for the phone /zh/.
Figure 25: Regression tree for the phone /zh/.

Pseudo Inverse Matrix

The treatment of multiple linear solutions can be simplified by using matricial notation:

$$\begin{bmatrix} y_{10} & y_{11} & \cdots & y_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ y_{n0} & y_{n1} & \cdots & y_{nd} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \Leftrightarrow Y \cdot a = b$$

If $Y$ is no singular, then $a$ can be written as $a = Y^{-1} \cdot b$. However, if $Y$ is not squared, for example, if the number of equations is higher than the number of variables, then $a$ is “overdetermined”, and in general do not existe exact solution to the problem. What can be done in this case is to look for a vector of weights $a$ which minimize some error function between $Y \cdot a$ and $b$. Defining an error vector $e$ as:

$$e = Y \cdot a = b$$

Then, a possible solution is the quadratic euclidian length of the vector $e$. In other words, a possible criteria is the minimization of the sum of the squared errors.

$$J_s(a) = \|Y \cdot a - b\|^2 = \sum_{i=1}^{n} (a' \cdot y_i - b_i)^2$$

The problem is minimizing the sum of quadratic errors is a classical problem. A closed solution to the problem can be obtained by taking the gradient:
\[ \nabla J_x = \sum_{i=1}^{n} 2 \cdot (a' \cdot y_i - b_i) \cdot y_i - 2 \cdot Y' \cdot (Y \cdot a - b) \]

Making the above equation equal to zero, we have the solution:

\[ Y' \cdot Y \cdot a = Y' \cdot b \]

We note that the problem \( Y \cdot a = b \) can be converted to \( Y' \cdot Y \cdot a = Y' \cdot b \). This solution have the property that the matrix \( Y' \cdot Y \) is squared and generally is no simetric. If \( Y' \cdot Y \) is no singular then there is a unique solution to \( a \):

\[ a = \left( Y' \cdot Y \right)^{-1} \cdot Y' \cdot b \Leftrightarrow a = \tilde{Y} \cdot b \]

With \( \tilde{Y} = \left( Y' \cdot Y \right)^{-1} \cdot Y' \). \( \tilde{Y} \) is denominated pseudo inverse of \( Y \). Note that if \( Y \) is squared and no singular, then a pseudo inverse is similar to the standard inverse matrix of \( Y \).
APENDIX B

The Figures B.1 show the correlation factor for each of the 44 phones. The correlation factor was estimated as described in the section 2.2.3.b.

The Figure B.2 shows the mean squared error for all 44 phones. It is important to note that despite the mean correlation factor to be equal to 0.9491 (taking all phones together), the mean squared error given in milliseconds is very high. The mean of the squared error, taking all phones together is equal to 18.15 ms, for the phone “uu” this value is equal to 28.61 ms and for the phone “zh” this value is only 3.62 ms.

The Figure B.3 shows the mean squared error divided by the mean of the phone. This normalization was done assuming that shorter phones (phone with smaller mean) are more sensitive to higher mean squared error than longer phones (phone with higher mean). This value was multiplied by 100 and expressed in percentage.

The last 44 Figures in this Appendix show the histogram for each 44 phone followed a graph showing the predicted durations versus the expected duration (the corrected duration). The green line corresponds to the corrected mean of the phone and the red line corresponds to the identify function (points where the error of estimation is equal to zero: predicted duration is equal to expected duration). In these Figures is possible to see that for some phone the prediction models are very poor. For example, the model of the phone “@” estimates almost only the mean of the phone. The models of the phone “ii” and “l”, among others, are not able to estimate well high values of duration. The models for the phones “r” and “z” explicitly concentrate on some bi(tri)modalites of the distributions, …

Figure B.1: Correlation factor per phone.
Figure B.2: Mean squared error for all 44 phones.

Figure B.3: Normalized mean squared error for all 44 phones.