Exploratory Analysis of Linguistic Data based on Genetic Algorithm an its Application to Robust Modeling of Speech Segmental Duration

Edmilson Morais, Fábio Violaro and Alexsandro Meireles

***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 have 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.**

***Keywords* – Speech and language technology, speech prosody, multivariate linear regression, genetic algorithm, hierarchical data clustering.**

I. INTRODUCTION

In Text-to-Speech synthesis (TTS) the modeling and estimation of 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)
- Interaction among linguistic factors

The main goal of this paper is to face the problem of database imbalance. In order to deal with this problem, an exploratory analysis of linguistic data based on Genetic Algorithm - GA [3] is proposed and it is used to improve the performance of an additive Linear Regression model. The additive linear regression model was chosen 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.

Edmilson Morais and Fábio Violaro, School of Electrical and Computer Engineering, University of Campinas, Brazil, E-mails: emorais@decom.fee.unicamp.br and fabio@decom.fee.unicamp.br. Alexsandro Meireles, Visiting scholar, Department of Linguistics at the University of Southern California, E-mail: meirelesalex@gmail.com

The proposed exploratory data analysis is used to answer two 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.

Section 2 of this paper presents a description of the database used in 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 secion 4. The results are on section 5 and finally section 6 presents some final considerations and suggestions for future works.

II. DATABASE DESCRIPTION

All the analysis was carried out on 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 1: List of 14 linguistic factors used in this work. (*Distances are given in number of syllables*).

Factors	Factor description
F0: phID	*Identity of the current phone*
F1: PosInSyll	*Position of the current phone in relation to the accented syllable of the current word*
F2: PrevPh	*Phone class of the previous phone*
F3: NextPh	*Phone class of the next phone*
F4: NNextPh	*Phone class of the next next phone*
F5: PoS	*Part-of-Speech tagging*
F6: ACC	*Degree of accentuation of the current word*

F7: NSyll	*Number of syllables in the current word*
F8: DistEnd	*Distance to the end of the current word*
F9: DistStress	*Distance to the stressed syllable in the next word*
F10:NextPause	*Distance to the next pause*
F11:PrevPause	*Distance from the previous pause*
F12:Chunk	*Distance to the end of the current accent group*
F13:PosInWord	*Position, in the current word, of the syllable that contains the current phone*

Table 2 presents the levels assumed for each of the 14 linguistic factors described in Table 1.

TABLE 2: Levels assumed for each of the 14 linguistic factors.

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

The set of phones used in the evaluations is described in Table 3.

TABLE 3: List of the phones used in this article

Toshiba	IPA	Example	Toshiba	IPA	Example
ii	i:	**ea**se	ai	aɪ	**r**ise
i	ɪ	p**i**t	au	aʊ	h**ou**se
e	ɛ	p**e**t	oi	ɔɪ	n**oi**se
ae	æ	p**a**t	ei	eɪ	r**ai**se
aa	ɑ:	c**a**lm	ou	ou	n**o**se
uh	ʌ	c**u**t	AR	ɑ(r)	f**ar**
oo	ɔ:	c**au**se	OR	ɔ(r)	p**or**t
uu	u:	l**o**se	p	p	**p**in
u	ʊ	p**u**t	t	t	**t**in
ER	ɚ	b**ir**d, moth**er**	k	k	**k**in
@	ə	**a**llow	b	b	**b**in
d	d	**d**in	jh	ʤ	**g**in
g	g	**g**ive	H	h	**h**it
f	f	**f**in	m	m	**m**ock
v	v	**v**an	n	n	**n**ot
s	s	**s**ir	ng	ŋ	doi**ng**
z	z	**z**oo	l	l	**l**eft
sh	ʃ	**sh**in	r	r	**r**ight
zh	ʒ	mea**s**ure	dx	ɾ (tap)	wri**t**er, ri**d**er

Toshiba	IPA	Example	Toshiba	IPA	Example
th	θ	**th**in	w	w	**w**asp
dh	ð	**th**is	y	j	**y**es
ch	ʧ	**ch**in	Q	ʔ	(glottal stop)
ccc	closure (plosives)				

III. LINEAR REGRESSION MODELS: QMTI

It was adopted the Hayashi's quantification method type I [2], henceforth QMTI. This method statistically predicts the relationship between a response value and categorical values using the multiple linear regression method as described in equation (3.1):

$$\overline{\overline{durPh(i)}} = \overline{durPh(i)} + \sum_{f}\sum_{l} a_{fl}\cdot\delta_{fl}(i), \quad i = 1,2,\ldots,N \qquad (3.1)$$

where N represents the total number of data samples; $\overline{\overline{durPh(i)}}$ represents the predicted phone duration of the *i-th* sample; $\overline{durPh(i)}$ represents the mean value of all data; a_{fl} represents the regression coefficients; $\delta_{fl}(i)$ represents the characteristic function given in equation (3.2):

$$\delta_{fl}(i) = \begin{cases} 1: & \text{If } i^{th} \text{ datum corresponds to level } l \\ & \text{of factor } f \\ 0: & \text{Otherwise} \end{cases} \qquad (3.2)$$

The regression coefficients a_{fl} can be calculated by minimizing equation (3.3) using a conventional multiple linear regression method

$$Error = \sum_{i=1}^{N} (\, durPh(i) - \overline{durPh(i)}\,)^2 \qquad (3.3)$$

The coefficients a_{fl} can be understood as a tied relationship between the whole linguistic information and the phone duration. From this point of view the predicted phone duration will be equal to the mean duration of the phone plus the effects (offsets) of all linguistic factors considered in the model.

IV. EXPLORATORY DATA ANALYSIS USING GA

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

- *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?

The next subsections, *A* and *B,* present the proposed solution for these two problems.

A. 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 one [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 in Figure 1.

Figure 1: Algorithm for optimal topology estimation

```
Algorithm for optimal topology estimation
for each phone class (one or more phones per class)
  BEGIN
    for i = 1 to N_i  (N_i: number of intermediate topologies)
      BEGIN
        Partition of the database (training X validation)
        Estimation of the i^th intermediate topology
      END
    Estimation of the optimal topology by majority rule
  END

Routine for estimation of the i^th 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
        Reprodution 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.

A.1. Cromossomic representation of each individual

A very simple binary representation was used. In this representation each cromossome corresponds to a 14 dimensional binary vector. If an allele is equal to 1, then its correspondent factor should make part of the topology; otherwise its correspondent factor should not make part of the topology. Table 4 shows an example:

TABLE 4: Cromossomic representation of the topologies

Cromossome	0	1	1	0	…	0	1	1
Selected factors	-	F1	F2	-	…	-	F12	F13

A.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. 80% of the data was used for training and 20% 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.

A.3. Objective function and 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].

A.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.

B. Phone Clustering

In order to deal with problems of sparse 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 into 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 in Figure 2.

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

B.1. Cromossomic representation of each individual

Table 5 shows an example describing the representation used. If the allele of the i^{th} position is equal to *0,* then the i^{th} phone will be classified in class *L* (*left*) ; otherwise the i^{th} phone will be classified in class *R* (*Right*).

TABLE 5: Cromossomic representation for cluster selection

Phone	1	2	3	...	43	44	45
Cromossome	**0**	**1**	**1**	**...**	**0**	**1**	**1**

Figure 2: Algorithm for phone clustering

Algorithm for phone clustering
Initialization: Start with all phones in a single cluster
$l = 1$ *(Initialize l equal to the first level of the tree)*
Repeat Until each cluster in the tree has only a single phone
BEGIN
for $i = 1$ to N_l *(N_l : Number of clusters in level l)*
BEGIN
Split the i^{th} cluster in level l
END
$l = l + 1$ *(Go to next level of the tree)*
END
Routine for spliting the i^{th} cluster in level l
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
Reprodution and selection of the individuals
END
IndivEpoch (j) =Best individual in generation j
END
BestSpliting =Best individual over all IndivEpoch (j)

B.2. Objective function

At each division of $Cluster_i$ into two son clusters, $Cluster_i^L$ and $Cluster_i^R$, the following objective function was used:

$$ObjF(Cluster_i^L, Cluster_I^R) = 1 - \frac{Corr(Cluster_i^L) + Corr(Cluster_i^R)}{2} \quad (4.1)$$

where $Corr(Cluster_i^L)$ and $Corr(Cluster_i^R)$ are the *Pearson* correlation coefficients for clusters $Cluster_i^L$ and $Cluster_i^R$ respectively.

V. RESULTS

A. 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/ contains only the factors: *F3* (*PosInSyll*) and *F4* (*NextPh*):

$$\overline{\overline{dur(zh)}} = \overline{dur(zh)} + a_1 \cdot PosInSyll + a_3 \cdot NextPh$$

Figure 3: Topologies - One phone per model

B. Analysis of the Majority Rule

In order to evaluate the performance of the majority rule, the following experiment was designed. For each phone, its correspondent majority rule topology and its intermediate topologies (50 different intermediate topologies) were evaluated 25 times using different partitions of the database into training and validation sets. Figure 4 shows the result for the phone /AR /.

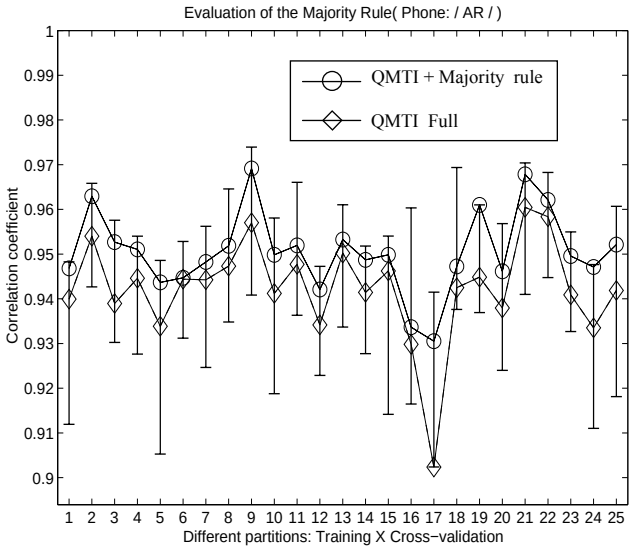

Figure 4: Evaluation of the Majority rule topology (legend ○) versus the Full topology (legend ◇) and the Intermediate topologies (vertical lines).

The vertical lines in Figure 4 show the correlation coefficient ranges for each intermediate topology along the 25 experiments. It is possible to see from Figure 4 that the results given by the majority rule are, in general, higher than or equal to the results given by the full topology. Moreover, the majority rule results are, in general, higher than the mean value given by the intermediate topologies and in some cases the majority rule results are almost equal to best intermediate topology results.

TABLE 6: Clusters at each level of the clusterization tree.

Clusters in the First Level	
C.1.1: @ H Q b ccc d dh dx e g i l m n ng 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	
Clusters in the Second Level	
C.2.1: @ H Q b ccc d dh dx e g i l m n ng r t u uh v w y z zh	
C.2.2: AR ER OR aa ae ai au ch ei f ii jh k oi oo ou p s sh th uu	
Clusters in the Third Level	
C.3.1: H Q b e g i l m n ng t uh z zh	
C.3.2: @ ccc d dh dx r u v w y	
C.3.3: ER aa ae f ii jh k oo p s sh th	
C.3.4: AR OR ai au ch ei oi ou uu	
Clusters in the Fourth Level	C.6.19: aa jh k p sh
C.4.1: H Q b n ng t z	C.6.20: f
C.4.2: e g i l m uh zh	C.6.21: ch ei ou
C.4.3: @ ccc d u v w y	C.6.22: au
C.4.4: dh dx r	C.6.23: AR
C.4.5: ae ii s th	C.6.24: OR
C.4.6: ER aa f jh k oo p sh	C.6.25: oi
C.4.7: au ch ei ou uu	C.6.26: ai
C.4.8: AR OR ai oi	**Clusters in the Seventh Level**
Clusters in the Fifth Level	C.7.1: H
C.5.1: H t z	C.7.2: z
C.5.2: Q b n ng	C.7.3: Q
C.5.3: g i zh **(C07)**	C.7.4: b ng
C.5.4: e l m uh	C.7.5: i
C.5.5: @ ccc u v w y	C.7.6: zh
C.5.6: d	C.7.7: m uh
C.5.7: dh dx **(C05)**	C.7.8: e
C.5.8: r	C.7.9: aa jh p sh
C.5.9: ae ii **(C08)**	C.7.10: k
C.5.10: s th	C.7.11: ch ou
C.5.11: ER oo	C.7.12: ei
C.5.12: aa f jh k p sh	**Clusters in the Eighth Level**
C.5.13: au ch ei ou	C.8.1 : b
C.5.14: uu	C.8.2 : ng
C.5.15: AR OR **(C04)**	C.8.3 : m
C.5.16: ai oi **(C09)**	C.8.4 : uh
Clusters in the Sixth Level	C.8.5 : a jh sh
C.6.1: H z **(C03)**	C.8.6 : p
C.6.2: t	C.8.7 : ch
C.6.3: Q b ng **(C02)**	C.8.8: ou
C.6.4: n	**Clusters in the Nineth Level**
C.6.5: i zh	C.9.1: aa sh **(C01)**
C.6.6: g	C.9.2: jh
C.6.7: e m uh	**Clusters in the Tenth Level**
C.6.8 : l	C.10.1: aa
C.6.9: ccc u v w y **(C06)**	C.10.2: sh
C.6.10: @	
C.6.11: dh	
C.6.12: dx	
C.6.13: ae	
C.6.14: ii	
C.6.15: s	
C.6.16: th	
C.6.17: ER	
C.6.18: oo	

C. Optimal Topologies: Phone Classes

Table 6 shows the clusters (group of phones) at each level of the clusterization tree. In order to identify which phones should be grouped and modeled together into 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 5 shows the topologies obtained for each phone class. 33 phone classes were obtained, 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)

Figure 5: Topologies – Phone classes

D. Overall Performance of the Method

Figure 6 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 in Figure 6 was specifically designed to evaluate the performance of the proposed method to deal with database imbalance problems. In order to obtain the results showed in Figure 6, 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 in Figure 6 are an average over the 50 experiments.

E. Analysis of the Linguistic Factor Effects

One of the main advantages of the proposed method is that it allows a straightforward evaluation of the influence of each linguistic factor into the segmental duration of each phone or phone class. Figure 7 shows an example for the phone / OR /. The vertical bars in Figure 7 represent the effect of each linguistic factor. These effects correspond to deviations around the mean value of the mean duration of the phone /OR/.

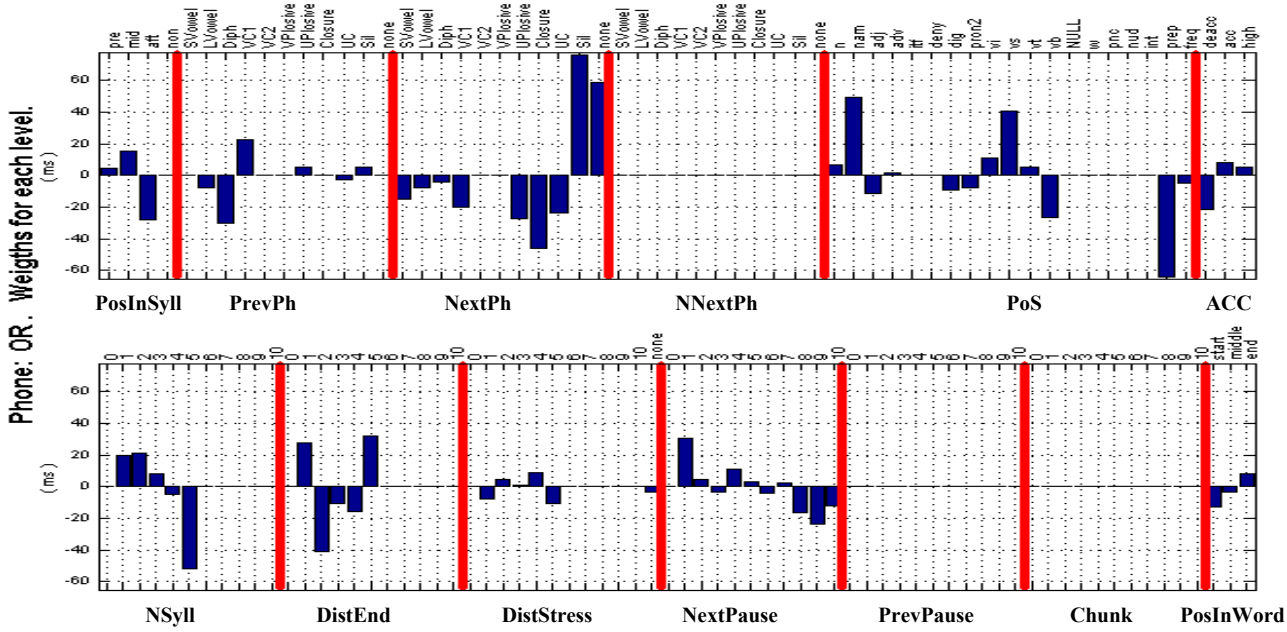

Figure 7: Regression coefficients of the QMTI+GA model for the phone /OR/ (weights/effects for each level of the linguistic factors).

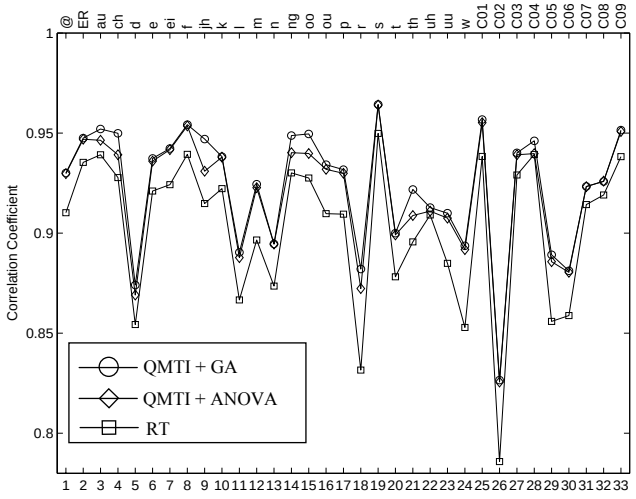

Figure 6: Overall results: QMTI+GA, QMTI+ANOVA and RT.

VI. 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 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 estimate 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 in this work, and in special to Dr. Kate Knill.

REFERENCES

[1] Sproat, Richard, *Multilingual Text-To-Speech Synthesis*, Spring-Verlang, New York, 1997.

[2] Hans., C., Sagisaka, Y., "Analysis of Segmental Duration for Thai Speech Synthesis", *Speech Prosody 2004*, Nara, Japan, March 23-26, 2004.

[3] Mitchel, M., *An Introduction to Genetic Algorithms*, The MIT Press, Cambridge, Massachusetts, 1998.

[4] Van Santen, J., "Assignment of Segmental Duration in Text-to-Speech Synthesis", *Computer, Speech and Language*, 8, 1994.

[5] Van Santen, J., Sproat, H.R., et al., (Eds.), *Progress in Speech Synthesis*. Springer, Berlin, 1997.

[6] Bernd M., van Santen J.,: "Modeling segmental duration in German text-to-speech synthesis". *ICSLP*, Philadelphia, PA, 1996.

[7] Breiman, L., "Bagging Predictors", *Technical Report No 421*, Univ. Berkeley, CA, September, 1994

[8] Jobson J.,D., *Applied Multivariate Data Analysis*, Volume I, Spring-Verlang, New York, 1991.