Phone Duration Modeling for Greek Emotional Speech Synthesis

Vasiliki Bourna, Alexandros Lazaridis and Nikos Fakotakis

Abstract

In this paper we cope with the task of phone duration modeling for Greek emotional speech synthesis. For this objective, we apply various well established machine learning techniques to an emotional speech database while we construct phone duration prediction models based on morphosyntactic and prosodic features that can be extracted only from text. We employ decision trees, linear regression, lazy learning algorithms and meta-learning algorithms, trained on a database consisting of 4150 words (distributed in five emotions). M5p model trees and meta-learning algorithms using M5p regression trees as base classifier proved to perform better, while we observed that the prediction was more accurate for the emotions with the most uniform distribution of phone durations.

1. Introduction

As the development of Text-to-Speech (TTS) systems is growing rapidly, the realization of better quality of synthesized speech becomes the most important goal in the field of synthetic speech technology. The quality of the systems producing synthetic speech lies upon two major characteristics: the naturalness of the synthetic voice, namely its similarity to the human voice, and its intelligibility, meaning how easy it is for the listener to understand the context of the synthetic speech [1]. Therefore, over the last years, there is an on going research concerning ways to improve the quality of synthetic speech, by implementing several factors that affect human speech.

Prosody refers to the aspects of human speech communication that introduce functions which may not be encoded by grammar, such as emphasis, intent, attitude or emotional state. Prosodic modeling is hence important in the field of speech processing. In speech, prosody is expressed by features such as duration (timing and segmental length), fundamental frequency (pitch variations), and energy (loudness) [1, 3]. In order to build robust prosody models it is essential to study each of these features extensively. In this paper we focus on phone duration modeling, which is a major issue, due to the fact that segmental duration affects the structure of utterances, altering consequently their naturalness and understanding. In this context, the construction of proper phone duration models is crucial for generating a highly natural synthetic speech. In order to achieve this objective, we have to determine the length of phones and specify other features that affect it, which could be extracted from various levels of speech signal representation, such as acoustic, phonetic and morphosyntactic.

Segmental duration modeling can be divided into two categories: rule-based and corpus-based modeling. Rule-based approaches combine linguistic expert knowledge with manual analysis of limited amount of data. The most representative rule-based approach is the one introduced by Dennis Klatt [2], according to which the duration of a segment is modified by rules that are applied sequentially, starting from some intrinsic rule. Rule-based modeling can be overcomplicated by the constant rule inference or conversely, inflexible to the handling of exceptions. On the other hand, corpus-based techniques are using large speech corpora in order to reveal the relation between timing of speech and linguistic features, usually with the aid of machine learning techniques such as Decision Trees (DT) [3, 4, 5, 6, 7], Linear Regression (LR) [8], Sum of Products (SOP) [9, 10, 11], Artificial Neural Networks (ANN) techniques [12, 13, 14, 15], Bayesian models [16, 17] and lazy learning algorithms [18].

To our consideration, in order to take better advantage of the effect of prosodic features in human speech analysis, it is essential to do research not only on the attributes of prosody of neutral speech, but also to examine its features in the context of emotional speech. In this way, it would be possible to add emotional effect on synthesized speech. There are several approaches concerning emotional speech synthesis, such as formant synthesis, diphone concatenation and unit selection. A thorough overview is presented in [19]. These approaches employ prosody modeling in order to synthesize certain emotions or to implement emotional prosody in TTS systems and generate more expressive speech [20, 21, 22]. Nevertheless, we believe that it is important segmental duration of emotional speech to be studied in more detail. The procedure of phone duration modeling in the context of emotional speech, together with the analysis of other prosodic features, takes us one step ahead to the accomplishment both of emotional speech synthesis, and also of more natural and expressive synthetic speech.

In this paper, phone duration modeling for a Greek emotional speech database is performed, using several machine learning techniques. The techniques that we employed can be divided into four categories of data-driven machine learning, which are Decision Trees (DT) [23], Linear Regression (LR) [24], Lazy-learning Algorithms [25] and Meta-learning Algorithms [25]. An emotional speech database has been utilized for the construction of training and testing sets, which was manually annotated according to the Gr-ToBI system [26].

The structure of this paper is as follows. Section 2 describes the emotional speech corpus that was used for the procedure of the duration modeling, along with the feature vector that was generated for the training of the models. In section 3 the machine learning algorithms which were applied for the phone duration modeling are described. Finally, in section 4 we present and discuss the experimental results.

Department of Electrical and Computer Engineering, University of Patras, Rion 26500, Greece
Tel. +30 2610 996496, Fax. +30 2610 997336
bmpoyrna@upnet.gr, {alaza, fakotaki}@upatras.gr
2. Database and Feature Set

The speech corpus used for the phone duration modeling of emotional synthetic speech in Modern Greek (MG) was developed in order to be linguistically and prosodically rich, so that it could be used for speech synthesis systems. The final choice of the emotional states to be recorded was based on studies which argue that there are four basic or archetypal emotions [27]. Thus, the selected utterances were recorded expressing sadness, anger, fear and joy as well as a neutral emotional state.

2.1. Corpus Description

The corpus was designed so as for each phone to have multiple instances in various positions in different words (initial, medial, final) in the corpus. Consequently the extraction of them is possible in order to be used as a structural element in a TTS system inventory. The contents of the corpus were extracted from passages, newspapers or were set up by a professional linguist. The final corpus which was utilized for the experiments consisted of 62 utterances. The length of the utterances was ranging from a single word, an utterance, small or large sentence or even a sequence of sentences of fluent speech. All the utterances were uttered separately in all five emotional stages. Furthermore, the context of all sentences was emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactical or semantic means.

The corpus (including all five emotions) consisted of 4.150 words while our phone inventory was composed of 34 phones distributed in 22.045 instances (15.667 voiced and 6.378 unvoiced phones). Moreover, due to the limited amount of training and test data, each vowel class included both stressed and unstressed cases of the corresponding vowel. The experiments consisted of 62 utterances. The length of the utterances was ranging from a single word, an utterance, small or large sentence or even a sequence of sentences of fluent speech. All the utterances were uttered separately in all five emotional stages. Furthermore, the context of all sentences was emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactical or semantic means.

The corpus was designed so as for each phone to have multiple instances in various positions in different words (initial, medial, final) in the corpus. Consequently the extraction of them is possible in order to be used as a structural element in a TTS system inventory. The contents of the corpus were extracted from passages, newspapers or were set up by a professional linguist. The final corpus which was utilized for the experiments consisted of 62 utterances. The length of the utterances was ranging from a single word, an utterance, small or large sentence or even a sequence of sentences of fluent speech. All the utterances were uttered separately in all five emotional stages. Furthermore, the context of all sentences was emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactical or semantic means.

The corpus was designed so as for each phone to have multiple instances in various positions in different words (initial, medial, final) in the corpus. Consequently the extraction of them is possible in order to be used as a structural element in a TTS system inventory. The contents of the corpus were extracted from passages, newspapers or were set up by a professional linguist. The final corpus which was utilized for the experiments consisted of 62 utterances. The length of the utterances was ranging from a single word, an utterance, small or large sentence or even a sequence of sentences of fluent speech. All the utterances were uttered separately in all five emotional stages. Furthermore, the context of all sentences was emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactical or semantic means.

The corpus was designed so as for each phone to have multiple instances in various positions in different words (initial, medial, final) in the corpus. Consequently the extraction of them is possible in order to be used as a structural element in a TTS system inventory. The contents of the corpus were extracted from passages, newspapers or were set up by a professional linguist. The final corpus which was utilized for the experiments consisted of 62 utterances. The length of the utterances was ranging from a single word, an utterance, small or large sentence or even a sequence of sentences of fluent speech. All the utterances were uttered separately in all five emotional stages. Furthermore, the context of all sentences was emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactical or semantic means.

2.2. Feature Set

In order to construct the train and test sets for the phone duration modeling task, we calculated the mean and standard deviations of duration from the entries. In this way, we were able to construct models that predict the segment duration both directly using actual phone duration in seconds and also using the z-score of the phone durations. The z-score is a statistic quantity which indicates how many standard deviations an observation is above or below the mean. The z-scores allow comparison of observations from different normal distributions. After the prediction of z-score, the phone duration is calculated by the following formula:

\[ \text{Dur}_{\text{ph}} = \text{Dur}_{\text{mean}} + (\text{Zscore} \times \text{StdDev}) \]  

Thus, we followed the procedure twice, once using as predicted variable the phone duration and the other using the z-score of each phone.

There are many features that could be extracted from text for the task of training a duration model [3, 6, 9, 11]. The feature set that we decided that is the most appropriate for the task of training duration models in our case includes a variety of phonological, morphological, linguistic and syntactic attributes. For some features, we also applied a window around the investigated phone, in order to take advantage of the information that the neighboring phones convey.

The feature vector that we used consisted of the following variables:

- **segment duration or z-score**, 
- **name**: the label of the phone, a window of [-1, +1] around the phone was applied, 
- **syl_onsetsize – syl_codaSize**: number of phones before and after the vowel in the specific syllable, 
- **pos_in_word**: position of this phone in the syllable it is related to, 
- **syl_last_accent – syl_next_accent**: number of stressed syllables since the last phrase break and to the next phrase break respectively, 
- **ssyl_in – ssyl_out**: number of stressed syllables since the last and to the next phrase break respectively, 
- **syl_toBI accent – syl_toBi_endtone**: ToBi accent and endtone (boundary tone) related to the syllable, 
- **syl_last_accent – syl_next_accent**: number of syllables after the vowel in the specific syllable, 
- **word_gpos**: function content word distinction, 
- **phonetic_attributes**: description of the articulation position of a phone

3. Duration Modeling Scheme

3.1. Algorithms

For the task of the construction of phone duration prediction models, we applied several machine learning algorithms, using
features that can be extracted only from text. Those methods are classified under four categories: Decision Trees, Linear Regression (LR), Lazy Learning Algorithms and Meta-learning Algorithms and are provided by the WEKA statistical analysis package [25].

3.1.1. Decision Trees

Decision trees are predictive models that create a mapping procedure between observations about an item and the conclusions about its target value [23]. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. There are two types of decision trees, classification trees and regression trees. For our data we built trees using the M5’ algorithm [29] which is implemented in WEKA as classification (M5p) and regression (M5p-R) trees. M5’ algorithm splits the input space progressively based on minimizing the intra-subset variation in the input values down to each branch. In each node, the standard deviation of the output values for the instances reaching a node is taken as a measure of the error of this node and the expected reduction in error is calculated as a result of testing each attribute and all possible split values. This process is applied recursively to all the subsets [29]. M5p are model trees that do the task of classification following the above described procedure. However, there is also the option to build regression trees using the M5’ algorithm, called Mp5-R.

Finally, we built duration models using the Reduced Error Pruning trees (REPTrees) [30]. Reduced error pruning trees use a fast pruning algorithm to produce an optimal pruning of a given tree. The REP algorithm works in two phases: first the set of pruning examples $X$ is classified using the given tree $T$ to be pruned. Counters that keep track of the number of examples of each class passing through each node are updated simultaneously. In the second phase, which is a bottom-up pruning phase, these parts of the tree that can be removed without increasing the error of the remaining hypothesis are pruned away. The pruning decisions are based on the node statistics calculated in the top-down classification phase.

3.1.2. Linear Regression

Linear regression (LR) [24] algorithm is a classification and prediction algorithm that expresses the class variable as a linear combination of the attributes that we take into account in order to construct the prediction model. The training data are used to calculate the weights which will be subsequently applied on the feature set, in order to predict the class. The algorithm is also known as the least squares method, because it performs the standard least-squares linear regression. It can optionally perform attribute selection.

3.1.3. Lazy-learning Algorithms

This category contains algorithms which defer processing of training data until a query needs to be answered. This usually involves storing the training data in memory and finding relevant data in the database to answer a particular query [25].

IBK is an instance based algorithm [31], which belongs to the lazy learning algorithms, using the k-nearest neighbors algorithm (k-NN). At first it stores the training instances verbatim and then searches for the instance that most closely resembles the new instance. This is calculated through the use of a distance function (in our case the Euclidian distance). The instances themselves represent the knowledge and the similarity function defines what is learned. In order to locate the instance that is closer to the training instance, it searches among the k nearest neighbors of the test instance. Testing this method with different number of neighbors, as a parameter, resulted in the adaptation of 12 neighbors ($k=12$) since it gave the best results.

Another lazy learning algorithm that we applied was the Local Weighted Learning algorithm (LWL) [32]. LWL is a general algorithm which assigns weights using an instance-based method and builds a classifier from the weighted instances. The task of the assignment of the weights depends, in this local algorithm, on the position of the training instance in the space of the input variables. The training instances which are located closer to the prediction point receive usually bigger weights. Furthermore, a distance function is also applied. The data weighting takes place either directly or through weighting an error criterion. Weighting the data can be viewed as replicating relevant instances and discarding the irrelevant ones. Moreover, a weighting function or kernel function is used to calculate a weight for a data point from the distance. In our case we used the tri-cube kernel function which is implemented in the WEKA toolkit [25], while we used REPTrees as classifiers.

3.1.4. Meta-learning Algorithms

Meta-learning algorithms [25] are based on the use of classifiers converting them into more powerful learners. This happens by applying learning algorithms to meta-data (data that provide information about other data managed within an application or environment). Their main goal is to use such meta-data to improve the performance of existing learning algorithms.

Additive Regression [33] is considered as a meta-learning technique as it refers to every prediction generated through addition of the contributions that emerges from other models. The method that is usually followed in additive regression is the following: there is a standard regression model built initially, such as a regression tree. The produced errors of this model are called residues. Subsequently, these errors get corrected by a second model, possibly another regression tree which tries to predict the residues. The addition of the predictions of the second model to the ones of the first automatically lead to a smaller error in the training data. This procedure is followed until no further reduction of the resulting error can be achieved by the system [31]. In our construction of prediction models with the use of additive regression we used M5p-R trees (AR-M5p-R) and REPTrees (AR-REPTree).

Moreover we used the Bagging algorithm (BG) [34] to predict the segmental duration. This method is a bootstrap ensemble of methods (bootstrap aggregation) that creates individual regression models by training the same learning algorithm on a random redistribution of the training set. Each regression model’s training set is generated by randomly drawing, with replacements, N instances, where N is the size of the original training set. Many of the original instances may be repeated in the resulting training set while other may be left out. After the construction of several regression models, taking the average value of the prediction of each regression model gives the final description. In this case, we also applied M5p-R trees (BG-M5p-R) and REPTrees (BG-REPTree) as base classifiers.
3.2. Performance Estimation Measures

Because of the limited amount of data that we had, and in order to build as accurate and independent models as possible, we followed the 10-fold cross validation evaluation method [26]. In order to evaluate the performance of the duration prediction models for each of the above described methods we calculated the root mean square error (RMSE), which is frequently used as a measure of the difference between the predicted by a model values and the actually observed ones. Moreover we calculated the correlation coefficient (CC) which measures the statistical correlation between the actual and the predicted values of the segmental duration or the z-scores, as well as the mean absolute error (MAE) which is described as the average magnitude of the errors in a set of predictions, without considering their direction (i.e. their sign).

### 4. Experimental Results and Discussion

Because of the large amount of different methods that we applied for the construction of the duration prediction models, we distinguished different axes in order to compare and evaluate the performance of the models.

We have to note that, as it was expected, the RMSE and the MAE values were better while the class variable was the z-score rather than when it was directly the segmental duration (measured in seconds). As far as correlation coefficient is concerned, its value turned out to be better for the case of the segmental duration than the one of the z-scores as a class variable. This is also obvious in table 1, where we have calculated the mean values for each case.

This happens because z-scores are a better representation of the duration of phones and give better results for the RMSE and MAE values [35]. Regarding the correlation, the duration models in the z-score domain may not be as good as when training models predicting absolute scores, however, when the scores predicted are convert back into the absolute domain the correlations are better too [35]. Therefore, we will present below only the results of the models constructed for the case of the z-scores as class.

#### 4.1. Experimental Results

In the following tables we present the experimental results for our duration modeling task. In tables 2, 3 and 4 the RMSE, the MAE and the correlation coefficient (CC) values for each different method and emotion are presented respectively.

#### Table 1. Mean values of RMSE, MAE and CC for each emotion for the case of phone duration and z-score as class.

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>MAE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dur</td>
<td>zscore</td>
<td>dur</td>
</tr>
<tr>
<td>sad</td>
<td>0.0230</td>
<td>0.0226</td>
<td>0.0171</td>
</tr>
<tr>
<td>anger</td>
<td>0.0253</td>
<td>0.0241</td>
<td>0.0184</td>
</tr>
<tr>
<td>fear</td>
<td>0.0219</td>
<td>0.0211</td>
<td>0.0159</td>
</tr>
<tr>
<td>joy</td>
<td>0.0218</td>
<td>0.0209</td>
<td>0.0161</td>
</tr>
<tr>
<td>neutral</td>
<td>0.0278</td>
<td>0.0272</td>
<td>0.0187</td>
</tr>
</tbody>
</table>

#### Table 2. RMSE values for all the emotional states and all the applied methods.

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>MAE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dur</td>
<td>zscore</td>
<td>dur</td>
</tr>
<tr>
<td>sad</td>
<td>0.0209</td>
<td>0.0217</td>
<td>0.0196</td>
</tr>
<tr>
<td>anger</td>
<td>0.0222</td>
<td>0.0229</td>
<td>0.0205</td>
</tr>
<tr>
<td>fear</td>
<td>0.0256</td>
<td>0.0276</td>
<td>0.0198</td>
</tr>
<tr>
<td>joy</td>
<td>0.0209</td>
<td>0.0229</td>
<td>0.0209</td>
</tr>
<tr>
<td>neutral</td>
<td>0.0222</td>
<td>0.0234</td>
<td>0.0228</td>
</tr>
<tr>
<td></td>
<td>0.0245</td>
<td>0.0268</td>
<td>0.0231</td>
</tr>
<tr>
<td>AR-REPtree</td>
<td>0.0241</td>
<td>0.0257</td>
<td>0.0227</td>
</tr>
<tr>
<td>AR-M5p-R</td>
<td>0.021</td>
<td>0.022</td>
<td>0.0198</td>
</tr>
<tr>
<td>BG-REPtree</td>
<td>0.024</td>
<td>0.0262</td>
<td>0.0214</td>
</tr>
<tr>
<td>BG-M5p-R</td>
<td>0.021</td>
<td>0.0224</td>
<td>0.02</td>
</tr>
</tbody>
</table>

#### Table 3. MAE values for all the emotional states and all the applied methods.

<table>
<thead>
<tr>
<th></th>
<th>MAE</th>
<th>MAE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dur</td>
<td>zscore</td>
<td>dur</td>
</tr>
<tr>
<td>sad</td>
<td>0.0158</td>
<td>0.016</td>
<td>0.0147</td>
</tr>
<tr>
<td>anger</td>
<td>0.0167</td>
<td>0.0168</td>
<td>0.0153</td>
</tr>
<tr>
<td>fear</td>
<td>0.0192</td>
<td>0.0203</td>
<td>0.0149</td>
</tr>
<tr>
<td>joy</td>
<td>0.0161</td>
<td>0.0172</td>
<td>0.0152</td>
</tr>
<tr>
<td>neutral</td>
<td>0.0164</td>
<td>0.0168</td>
<td>0.0168</td>
</tr>
<tr>
<td></td>
<td>0.0183</td>
<td>0.0194</td>
<td>0.0172</td>
</tr>
<tr>
<td>AR-REPtree</td>
<td>0.0181</td>
<td>0.0188</td>
<td>0.0168</td>
</tr>
<tr>
<td>AR-M5p-R</td>
<td>0.016</td>
<td>0.0162</td>
<td>0.0149</td>
</tr>
<tr>
<td>BG-REPtree</td>
<td>0.018</td>
<td>0.0192</td>
<td>0.0158</td>
</tr>
<tr>
<td>BG-M5p-R</td>
<td>0.0159</td>
<td>0.0164</td>
<td>0.0148</td>
</tr>
</tbody>
</table>

#### Table 4. Correlation coefficient values for all the emotional states and all the applied methods.

<table>
<thead>
<tr>
<th></th>
<th>CC</th>
<th>CC</th>
<th>CC</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dur</td>
<td>zscore</td>
<td>dur</td>
<td>zscore</td>
</tr>
<tr>
<td>sad</td>
<td>0.6725</td>
<td>0.7513</td>
<td>0.6863</td>
<td>0.6986</td>
</tr>
<tr>
<td>anger</td>
<td>0.6169</td>
<td>0.7156</td>
<td>0.6341</td>
<td>0.6482</td>
</tr>
<tr>
<td>fear</td>
<td>0.4313</td>
<td>0.5477</td>
<td>0.6737</td>
<td>0.6044</td>
</tr>
<tr>
<td>joy</td>
<td>0.6666</td>
<td>0.7156</td>
<td>0.6325</td>
<td>0.6754</td>
</tr>
<tr>
<td>neutral</td>
<td>0.6255</td>
<td>0.7058</td>
<td>0.545</td>
<td>0.6377</td>
</tr>
<tr>
<td></td>
<td>0.5091</td>
<td>0.5912</td>
<td>0.5205</td>
<td>0.5796</td>
</tr>
<tr>
<td>AR-REPtree</td>
<td>0.5636</td>
<td>0.6426</td>
<td>0.5739</td>
<td>0.6023</td>
</tr>
<tr>
<td>AR-M5p-R</td>
<td>0.66</td>
<td>0.7402</td>
<td>0.671</td>
<td>0.6912</td>
</tr>
<tr>
<td>BG-REPtree</td>
<td>0.5138</td>
<td>0.598</td>
<td>0.6017</td>
<td>0.6174</td>
</tr>
<tr>
<td>BG-M5p-R</td>
<td>0.6675</td>
<td>0.7347</td>
<td>0.6741</td>
<td>0.6799</td>
</tr>
</tbody>
</table>

#### 4.1.1. Comparison between methods

As it is clearly shown in the tables, all the methods which were applied had a satisfactory performance, yielding RMSE values between 0.0193 and 0.0295 and MAE values between 0.0145 and 0.0203. Regarding the correlation coefficient, our models succeeded values from 0.4313 to 0.7513, which is a considerably good outcome.
In the tables the lowest RMSE and MAE values and the highest correlation coefficient values for each emotion are shown in bold. As we can notice, the methods with the overall best performance were the M5p trees, as well as the meta-learning algorithms which used M5p regression trees as base classifiers (AR-M5p-R, BG-M5p-R).

Going further in the comparison between methods, we notice that LR had a very satisfactory performance too, together with M5p Regression trees (M5p-R). Moreover, it is interesting to remark that between the two local learning methods that we applied, IB12 performed better in all cases than LWL.

On the contrary, REPTrees appear to have the worst performance, comparing to the others, both as a single prediction method, and as a base-classifier for the case of AR and BG algorithms. We discuss possible reasons for these performances below.

4.1.2. Comparison between emotions

At this point, it is interesting to compare the experimental results on another basis, the one that concerns the performance between emotions. According to the results presented in the tables, there seems to be a tendency in the case of some emotions to show lower RMSE and MAE values than others, which means that their duration prediction models perform better.

The cases where we observed the lowest values for RMSE and MAE are joy and fear, in which cases, independently of which method we applied, the RMSE values do not overcome 0.0245 and MAE has a maximum of 0.0179 (both for the case of REPTrees). Sadness and anger have slightly worse evaluation values and neutral speech is the state for which we noticed the worst performance, with maximum RMSE of 0.0295 and MAE of 0.0196. The same tendency exists for the case of correlation coefficient, where the highest values were for the case of anger and then follow joy, fear, sadness and neutral speech.

4.2. Discussion

As it is clear from the above tables, all the applied methods worked well for our data and built models which perform adequately for the task of phone duration modeling. M5p trees accomplished the best results due to the fact that they adopt a greedy algorithm which constructs a model tree with a non-fixed structure by using a certain stopping criterion. M5’ minimizes error at each interior node, one node at a time, and does this process recursively until all or almost all of the instances are correctly classified. In this way, it carries out the task of prediction in longer time, but constructs very robust models.

Moreover, the meta-learning algorithms take advantage of the information that is produced from other methods, as they process meta-data. Thus, it was expected that they would perform well. However, as we can notice, Additive Regression and Bagging performed better when combined with a robust prediction method such as M5p-R, while they didn’t perform that well for the case of REPTrees. This leads to the conclusion that the choice of the appropriate classifiers is an important issue when this method is applied. Regarding the other methods, we noticed that local learning methods or methods that apply a more strict strategy of “pruning” might perform faster, but they do not yield the best results.

In this point it is interesting to point out that the emotions, the phones of which had the lower values of standard deviation, namely the ones that have more uniform distribution of the mean duration of each phone, are the ones with the lower error values. As shown in fig. 1, for the case of joy and fear the weighted average of phone standard deviations is lower, and therefore the phone duration model performed better.

5. Conclusion – Future Work

In this study, we presented various phone duration models for Greek emotional speech, using Decision Trees, linear regression, lazy learning algorithms and meta-learning algorithms. A speech database of five emotional states was used, while the selected corpus consisted of 4150 words. Our evaluation results show that all the duration models that we built make robust predictions. Nevertheless, M5p model trees and meta-learning algorithms perform better. As future work, we are interested in the exploitation of the phone duration modeling not only in the field of emotional speech synthesis, but also for the achievement of more expressive and natural synthetic speech.

6. Acknowledgement

This work was supported by the PlayMancer project (FP7-ICT-215839-2007), which is co-funded by the European Commission.

7. References


Bayesian Network for Phone

Chien J. T. and Huang C.H. “Bayesian Learning of

An RNN-based


