Journal of Computer Science 6 (3): 341-349, 2010
ISSN 1549-3636

Comparative Evaluation of Phone Duration Models for Greek Emotional Speech

Alexandros Lazaridis, Vasiliki Bourna and Nikos Fakotakis
Wire Communications Laboratory, Department of Electrical and Computer Engineering,
University of Patras, Rion-Patras 26500, Greece

Abstract: Problem statement: In this study we cope with the task of phone duration modeling for Greek emotional speech synthesis. **Approach:** Various well established machine learning techniques are applied for this purpose to an emotional speech database consisting of five archetypal emotions. The constructed phone duration prediction models are built on phonetic, morphosyntactic and prosodic features that can be extracted only from text. We employ model and regression trees, linear regression, lazy learning algorithms and meta-learning algorithms using regression trees as base classifiers, trained on a Modern Greek emotional database consisting of five emotional categories: anger, fear, joy, neutral and sadness. **Results:** Model trees based on the M5' algorithm and meta-learning algorithms using as base classifier regression trees based on the M5' algorithm proved to perform better. **Conclusion:** It was observed that the emotional categories of the speech database with the most uniform distribution of phone durations built the most accurate models.

Key words: Phone duration modeling, statistical modeling, emotional speech, text-to-speech synthesis

INTRODUCTION

The most important goal in the field of synthetic speech technology is the improvement of the quality of synthesized speech. The quality of the synthetic speech lies upon two main characteristics: the naturalness of the synthetic voice and its intelligibility. The former conveys the similarity of the synthetic speech to the human voice (Klatt, 1987). The latter reflects the level of difficulty for the listener to understand the context of the synthetic speech (Klatt, 1987). Consequently, over the last years, there is an ongoing research concerning ways to implement several factors that affect human speech using various techniques for improving the quality of synthetic speech.

Modeling of prosody plays a very important role in the field of speech processing and more specific in speech synthesis. In human speech communication, prosody refers to the introduction of functions and aspects of speech which may not be encoded by grammar, such as emphasis, intent, attitude or emotional state. In speech, prosody is expressed by factors such as duration (timing and segmental length), fundamental frequency (pitch variations) and energy-intensity (loudness) (Klatt, 1987; Mobius and Santen, 1996). For building robust prosody models it is essential to study each of these prosody factors extensively. Consequently, in this study we focus on phone duration modeling, which is a major issue, since segmental duration affects the structure of utterances and therefore alters their naturalness and understanding. In this context, the production of highly natural synthetic speech is highly and directly correlated to the construction of proper phone duration models. In order to achieve this objective, the determination of the length of the phones and the specification of other features that affect it is crucial.

The phone duration modeling approaches are divided in two major categories: The rule-based (Klatt, 1979) and the data-driven methods (Mobius and Santen, 1996; Santen, 1992; Chen *et al*., 1998; Chien and Huang, 2003; Lazaridis *et al*., 2007). In the rule-based methods manually produced rules, extracted from experimental studies on large sets of utterances or based on previous knowledge, are utilized for determining the duration of segments. Expert linguists are required for the extraction of these rules. One of the first and most well known attempts in the filed of rule-based segmental duration modeling is the one proposed by Klatt (1979). Similar models were developed in other languages such as French (Bartkova and Sorin, 1987), Swedish (Carison and Granstrom, 1988 and Greek (Epitropakis *et al*., 1993). The major drawback of the rule-based approaches is the difficulty to represent and

Corresponding Author: Alexandros Lazaridis, Wire Communications Laboratory,
Department of Electrical and Computer Engineering, University of Patras, Rion-Patras 26500, Greece
Tel: +30 2610 996496 Fax: +30 2610 997336

tune manually all the linguistic, phonetic and prosodic factors which influence the segmental duration in speech. Therefore, long-term devotion to this task becomes crucial and mandatory in order to collect all the appropriate (or even enough) rules (Klatt, 1987). Thus, the rule-based duration models are restricted to controlled experiments, where only a limited number of contextual factors are involved, in order to be able to deduce the interaction among these factors and extract these rules (Rao and Yegnanarayana, 2007).

Data-driven methods for the task of phone duration modeling were developed after the construction of large databases (Kominek and Black, 2003). Data-driven approaches are based either on statistical methods or Artificial Neural Network (ANN) based techniques that automatically produce phonetic rules and construct duration models from large speech corpora, overcoming in this way the problem of manual rules extraction. Their main advantage, in contrast to the rule-based techniques, is that this process does not depend on linguists. Over the last years various statistical methods have been applied in the phone duration modeling task such as, Linear Regression (LR) (Takeda *et al*., 1989), decisions tree-based models (Mobius and Santen, 1996), Sums-Of-Products (SOP) (Santen, 1992). Artificial Neural Networks (ANN) techniques (Chen *et al*., 1998), Bayesian models (Chien and Huang, 2003) and instance-based algorithms (Lazaridis *et al*., 2007) have also been introduced on the phone duration modeling task. Consequently the data-driven approaches offer us the ability to overcome the time consuming labor of the manual extraction of the rules which are needed in the rule-based approaches.

In phone duration modeling apart from investigating and evaluating different modeling techniques and in order to take better advantage of the effect of prosodic features in human speech analysis, it is essential to investigate not only the attributes of prosody of neutral speech, but also to examine prosodic features in the context of emotional speech. This research can lead to the incorporation of emotional effect on synthesized speech producing expressive synthetic speech. Several approaches introducing emotional speech synthesis have been presented over the years, such as formant synthesis, diphone concatenation and unit selection. In order for these approaches to synthesize certain emotions or to implement emotional prosody in TTS systems and generate more expressive speech, prosody modeling is employed (Jiang *et al*., 2005; Tesser *et al*., 2005; Inanoglu and Young, 2009). However, the task of segmental duration modeling of emotional speech is essential to be studied in more detail. The phone duration modeling task in the context of emotional speech, together with the analysis of other prosodic features, takes us one step ahead to the improvement of the quality of emotional speech synthesis and furthermore to more natural and expressive synthetic speech.

In the present research, several machine learning techniques are employed for the task of phone duration modeling on a Greek emotional speech database. The utilized techniques can be divided into four categories of data-driven machine learning, which are Decision Trees (DT) (Mitchell, 1997), Linear Regression (LR) (Witten and Frank, 2005), lazy-learning algorithms (Witten and Frank, 2005) and meta-learning algorithms (Witten and Frank, 2005). An emotional speech database has been utilized for the construction and evaluation of the phone duration models, which was manually annotated according to the Gr-ToBI system (Arvaniti and Baltazani, 2000).

Firstly, the machine learning algorithms which were applied on the phone duration modeling task are described. Next the emotional speech database that was used for building and evaluating the models along with the feature vector that was used and the performance estimation measures used for the evaluation of the models are described. Finally, we present and discuss the experimental results. This article ends with concluding remarks.

MATERIALS AND METHODS

Duration modeling algorithms: Several machine learning algorithms were applied for the task of phone duration modeling using features that can be extracted only from text. Those methods are classified under four categories, which are the following: decision trees, Linear Regression (LR), lazy learning algorithms and meta-learning algorithms.

Decision trees: Decision trees are predictive models that create a mapping procedure between observations about an item and the conclusions about its target value (Mitchell, 1997). In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to these classifications. In our experimental evaluation trees using the M5' algorithm (Wang and Witten, 1997) utilizing a model (M5p) and a regression (M5p-R) trees were built.

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 (Wang and Witten, 1997). The M5' can be used as a regression tree (M5p-R) or as a model tree (M5p). If a leave, in M5' algorithm's building process, is associated with an average output value of the instances sorted down to it, then the model is called regression tree (Quinlan, 1992). If the tree concludes in its leaves to more complex regression functions of the input variables, then the model is called model tree (Wang and Witten, 1997).

Furthermore the Reduced Error Pruning Trees (REPTrees) (Kaariainen and Malinen, 2004) were used. The REPTrees 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 S 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.

Linear regression: Linear Regression (LR) (Witten and Frank, 2005) algorithm is a classification and prediction algorithm that expresses the class variable as a linear combination of the attributes that are taken into account for constructing 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. Instead of using all the attributes, M5' algorithm can be applied for feature selection (Wang and Witten, 1997). During feature selection the attribute with the smallest standardized coefficient is iteratively removed until no improvement is observed in the error estimation. The error estimation is given by the Akaike Information Criterion (AIC) (Akaike, 1974).

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 (Witten and Frank, 2005).

IBK is an instance based algorithm (Aha *et al*., 1991), 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. In order to locate the instance that is closer to the training instance, it searches among the k nearest neighbors of the test instance. Evaluating this method with different number of neighbors resulted in the adaptation of 12 neighbors (k = 12) since it gave the best results.

Another lazy learning algorithm that we applied was the Locally Weighted Learning algorithm (LWL) (Atkeson *et al*., 1996). LWL is a general algorithm which assigns weights using an instance-based method-in our case the Linear-NN, which is a nearest neighbor search algorithm-and builds a classifier from the weighted instances. 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 tricube kernel function, while REPTrees were used as classifiers.

Meta-learning algorithms: Meta-learning algorithms (Witten and Frank, 2005) are based on the use of classifiers converting them into more powerful learners. This happens by applying learning algorithms to meta-data, which are data that provide information about other data managed within an application or environment.

Additive Regression (AR) (Stone, 1985) is a meta-learning technique which enhances the performance of a regression algorithm. During the training procedure, the additive regression algorithm builds a regression tree, in each iteration, using the residuals of the previous tree as training data. The regression trees are combined together creating the final prediction function. The addition of the predictions of the next model to the ones of the previous automatically leads to a smaller error in the training data. In our experiments the additive regression algorithm was combined with M5p-R trees (AR-M5p-R) and REPTrees (AR-REPTrees). In these two cases of additive regression meta-classification the shrinkage parameter, ν, indicating the learning rate, was set equal to 0.5 and the number of the regression trees, rt-num, was set equal to 10 after some grid search experiments (ν = {0.1, 0.3, 0.5, 0.7, 0.9}, rt-num = {5, 10, 15, 20}) on a randomly selected subset of the training set, representing the 40% of the size of the full training set.

Moreover the Bagging algorithm (BG) (Breiman, 1996) was used to model the phone duration. In the bagging algorithm, the dataset is split in multi subsets utilizing one regression tree for each one of them. Many of the original instances may be repeated in the resulting training set while other may be left out. The final prediction value is the average of the values predicted from each regression tree. In this case, we also applied M5p-R trees (BG-M5p-R) and REPTrees (BG-REPTrees) as base classifiers. The number of the regression trees, rt-num, was set equal to 10 after some grid search experiments (rt-num = {5, 10, 15, 20}) on the randomly selected subset of the training set, mentioned earlier.

Database and feature set: A Modern Greek (MG) emotional speech database was used for the task of phone duration modeling of emotional speech. This database was developed in order to be linguistically and prosodically rich, so that it could be used for speech synthesis systems. The selected utterances were recorded expressing anger, fear, joy and sadness, as well as a neutral emotional state. The choice of these emotional states to be recorded was based on studies which argue that there are four basic or archetypal emotions (Oatley and Johnson-Laird, 1998).

Database description: During the human speech production procedure the positional and contextual factors of a phone (place in syllable and word) play a very important role in the assessment of its duration (Mobius and Santen, 1996; Santen, 1992). The database was designed following this statement, so as for each phone to have multiple instances in various positions in different words (initial, medial, final) in the database. The database which was utilized for the experiments consisted of 62 utterances, which are pronounced several times with different emotional charge. The utterances of the database were extracted from passages, newspapers or were set up by a professional linguist. The length of the utterances was ranging from a single word, a phrase, small or large sentences or even a sequence of sentences of fluent speech. The context of all sentences was emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactical or semantic means. Moreover all the utterances were uttered separately in the five emotional styles.

The database, including all five emotional states, consisted of 4.150 words. The phone inventory which was used composed by 34 phones distributed in 22.045 instances (15.667 voiced and 6.378 unvoiced phones). Furthermore, each vowel class included both stressed and unstressed cases of the corresponding vowel. The utterances were uttered by a professional, female actress, speaking Modern Greek. She was instructed to read all utterances with one emotion then change it and start over again for the next emotion, ensuring in this way that the speaker would not have to change her emotional state more than five times, expressing anger, fear, joy, sadness and neutral emotion respectively. In addition, she was instructed to express a 'casual' intensity of the chosen states avoiding any theatrical exaggeration.

The recording sessions were held in the anechoic chamber of a professional studio. The recorded speech was sampled directly at 44.1 kHz and then down sampled at the frequency of 16 kHz and a resolution of 16 bit, for the needs of our experiments.

Feature set: In the present research all phone duration models were build twice. On the first time all the models were built so as to model and predict directly the phone durations in milliseconds and the second time were trained based on the z-scores of the durations of the phones so as to model and predict the z-score for each phone. The z-score is a statistic quantity which indicates the number of standard deviations an observation is above or below the mean. The z-score allows comparison of observations from different normal distributions. After the prediction of z-score, the phone duration is calculated by the following formula:

$$\mathrm{Dur}_{ph} = \mathrm{Dur}_{mean} + (\mathrm{Zscore} \times \mathrm{StdDev}) \quad (1)$$

For constructing the models using the z-scores, we calculated the mean and standard deviations of duration from the entries. The z-score has often been used in duration modeling since it allows a certain degree of normalization over different phones. As it is reported in the literature, z-score is a better representation of the segmental durations on the task of duration modeling and usually gives better results (Black and Lenzo, 2000). In order to investigate if this statement is applied on the emotional speech database, all the models were built modeling the phone durations both directly using the actual phone durations in milliseconds and also using the z-scores of the phone durations.

Various features can be extracted from text for the task of phone duration modeling (Mobius and Santen, 1996; Santen, 1992). The feature set implemented for this task includes 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 conveyed by the neighboring phones.

From each utterance we computed 33 features along with the contextual information concerning some of these features, described next:

- Eight phonetic features: The phone type (vowel/consonant), the vowel length (short, long, diphthong or schwa), the vowel height (high, middle or low), the vowel frontness (front, middle or back), the rounded type (lip or rounding), the manner of production (consonant type), the place of articulation (labial, alveolar, palatal, labiodental, dental, velar, glottal), the consonant voicing. Along with the aforementioned features, the information concerning the two previous and the two next instances of these features was also used
- Three segment-level features: The phone name with the information of the neighboring instances (previous, next), the position of the phone in the syllable and the onset-coda type (if the specific phone is before or after the vowel in the syllable)
- Thirteen syllable-level features: The position type of the syllable (single, initial, middle or final) with the information of the neighboring instances (previous, next), the number of all the syllables, the number of the accented syllables and the number of the stressed syllables since the last and to the next phrase break, syllable's onset-coda size (the number of phones before and after the vowel of the syllable) with the information of previous and next instances, the onset-coda type (if the consonant before and after the vowel in the syllable is voiced or unvoiced) with the information of previous and next instances, the position of the syllable in the word and the onset-coda consonant type (the manner of production of the consonant before and after the vowel in the syllable)
- Two word-level features: The part-of-speech (noun, verb and adjective) and the number of syllables of the word
- One phrase-level feature: The syllable break (the phrase break after the syllable) with the information of the neighboring (two previous, two next) instances
- Six accentual features: The ToBI accents and boundary tones with the information of the neighboring (previous, next) instances, the last-next accent (the number of the syllables since the last and to the next accented syllable) and we also included the stressed-unstressed syllable feature (if the syllable is stressed or not) and the accented-unaccented syllable feature (if the syllable is accented or not) with the information of the neighboring (two previous, two next) instances

The overall size of the feature vector, which was used for the task of phone duration modeling, including the aforementioned features and their contextual information is 93.

Performance estimation measures: In order to better utilize the available data, in all the experiments we followed an experimental protocol based on 10-fold cross-validation. The performance of the phone duration prediction models was measured in terms of Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Correlation Coefficient (CC). The RMSE is frequently used as a global measure sensitive to gross errors. The MAE, described as the average magnitude of the errors in a set of predictions, does not consider the direction of the deviations from the ground truth and is not that sensitive to gross errors. The RMSE and the MAE are defined respectively by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(F(x_i) - y_i\right)^2}{N}} \quad (2)$$

and by:

$$MAE = \frac{\sum_{i=1}^{N}\left|F(x_i) - y_i\right|}{N} \quad (3)$$

Where:

N = The number of the test instances
y_i = The actual duration in milliseconds or the z-score of ith instance
$F(x_i)$ = The predicted value for the ith instance

Finally we calculated the Correlation Coefficient (CC) which measures the statistical correlation between the actual and the predicted values of the phone duration directly in milliseconds or the z-scores. The CC is defined by:

$$CC = \frac{\mathrm{cov}\left(F(X), Y\right)}{\sigma_{F(X)}\sigma_Y} = \frac{E\left(\left(F(X) - \mu_{F(X)}\right)\left(Y - \mu_Y\right)\right)}{\sigma_{F(X)}\sigma_Y} \quad (4)$$

Where:

$F(X)$ = The variable of the predicted values

Y = The actual values of the phone durations in milliseconds or of the z-scores

The $\mu_{F(X)}$ and μ_Y are the mean values of the two variables and $\sigma_{F(X)}$, σ_Y are the standard deviations of the variables F(X) and Y respectively. Together these three performance measures offer a good indication about the accuracy of different models.

RESULTS

As it was expected, the RMSE and the MAE values were lower-meaning more accurate models-while using the z-scores as prediction variable rather than when directly the phone duration (measured in milliseconds) was predicted by the models. Concerning the correlation coefficient, its values turned out to be higher-showing higher statistical correlation-for the case of using directly the phone duration in milliseconds rather than when the z-scores were used as prediction variable. This can be seen in Table 1, where the mean values of all the models of both prediction class variables (z-scores and durations in milliseconds) for all the emotions and all the performance estimation measurements are shown. This was expected since the z-scores are a better representation of the duration of phones offering a certain degree of normalization over different phones resulting in more accurate phone duration models (Black and Lenzo, 2000). Regarding the CC, the duration models in the z-score domain may not be as high as when training models predicting directly the duration of the phones, however, if the predicted z-scores are converted back into the absolute domain the correlations are better too (Black and Lenzo, 2000).

In the Table 2-4 we present the experimental results for the phone duration modeling task. In Table 2-4 the RMSE, the MAE and the CC values for each algorithm and emotion of both prediction class variables (z-scores and durations in milliseconds) are presented respectively.

Table 1: Mean values of RMSE, MAE and CC for each emotion for the case of phone duration measured in milliseconds and z-score as prediction variable class

	RMSE		MAE		CC	
	Z-scores	Duration	Z-scores	Duration	Z-scores	Duration
Anger	23.87	24.96	17.49	18.29	0.68	0.77
Fear	20.63	21.91	15.29	16.15	0.64	0.65
Joy	20.72	21.40	15.34	15.80	0.65	0.71
Neutral	26.83	27.29	17.78	18.24	0.56	0.63
Sadness	22.26	22.51	16.83	17.02	0.60	0.69

Table 2: Root Mean Square Error (RMSE) for all the emotional categories and all the applied methods

	Z-scores					Durations (ms)				
RMSE	Anger	Fear	Joy	Neutral	Sadness	Anger	Fear	Joy	Neutral	Sadness
AR-M5p-R	22.0	19.5	19.1	25.7	20.4	22.1	20.1	19.0	26.3	20.6
AR-REPTrees	23.2	20.9	20.3	26.6	21.4	23.8	21.3	20.8	26.7	22.1
BG-M5p-R	22.4	20.0	19.8	26.0	21.0	23.3	20.9	20.4	26.7	21.4
BG-REPTrees	26.2	21.4	21.1	27.5	24.0	28.2	22.5	22.8	27.6	24.3
IB12	23.2	20.7	20.6	26.3	21.9	24.7	21.8	22.2	27.5	20.6
LWL	26.8	23.0	22.2	28.7	24.5	28.6	24.4	23.4	28.9	25.7
LR	22.7	20.9	19.8	26.4	20.9	22.8	22.0	19.8	26.4	20.8
M5p	21.7	19.6	19.4	25.2	20.9	21.7	20.2	19.5	26.2	20.9
M5pR	22.9	20.5	20.4	26.4	22.0	24.1	21.6	21.6	27.2	22.1
REPTrees	27.6	19.8	24.5	29.5	25.6	30.3	24.3	24.5	29.4	26.6

Table 3: Mean Absolute Error (MAE) for all the emotional categories and all the applied methods

	Z-scores					Durations (ms)				
MAE	Anger	Fear	Joy	Neutral	Sadness	Anger	Fear	Joy	Neutral	Sadness
AR-M5p-R	16.0	14.5	14.3	17.2	15.5	16.3	14.9	14.0	17.5	15.6
AR-REPTrees	17.0	15.5	15.0	17.6	16.4	17.5	15.7	15.3	17.8	16.8
BG-M5p-R	16.4	14.8	14.7	17.2	15.9	17.1	15.4	15.1	17.7	16.2
BG-REPTrees	19.2	15.8	15.6	18.2	18.0	20.5	16.5	16.7	18.6	18.1
IB12	16.6	15.0	14.9	17.4	16.4	18.0	15.8	16.4	18.4	15.6
LWL	19.4	17.2	16.3	18.6	18.3	20.5	18.2	17.0	19.3	19.0
LR	17.2	15.2	14.9	17.7	16.1	17.1	16.0	14.9	17.7	16.1
M5p	16.0	14.7	14.5	16.6	15.8	16.1	15.0	14.8	17.1	16.0
M5pR	16.8	15.3	15.3	17.7	16.7	17.6	16.0	15.9	18.2	16.8
REPTrees	20.3	14.9	17.9	19.6	19.2	22.2	18.0	17.9	20.1	20.0

Table 4: Correlation Coefficient (CC) for all the emotional categories and all the applied methods

	Z-scores					Durations				
CC	Anger	Fear	Joy	Neutral	Sadness	Anger	Fear	Joy	Neutral	Sadness
AR-M5p-R	0.74	0.69	0.70	0.62	0.69	0.83	0.72	0.78	0.66	0.75
AR-REPTrees	0.70	0.62	0.65	0.56	0.63	0.79	0.67	0.73	0.65	0.70
BG-M5p-R	0.73	0.67	0.68	0.60	0.67	0.81	0.70	0.75	0.66	0.73
BG-REPTrees	0.60	0.60	0.62	0.52	0.51	0.70	0.62	0.66	0.62	0.63
IB12	0.71	0.64	0.65	0.58	0.63	0.78	0.66	0.69	0.63	0.75
LWL	0.59	0.52	0.58	0.51	0.51	0.70	0.55	0.65	0.59	0.59
LR	0.72	0.63	0.68	0.57	0.67	0.81	0.66	0.76	0.66	0.74
M5p	0.75	0.69	0.70	0.63	0.67	0.83	0.72	0.77	0.67	0.74
M5pR	0.72	0.64	0.65	0.57	0.62	0.79	0.66	0.70	0.63	0.70
REPTrees	0.55	0.67	0.60	0.44	0.43	0.65	0.55	0.60	0.57	0.54

Comparison among the algorithms: As it is shown in the Table 2-4, all the algorithms which were applied in the task of the phone duration modeling built models with satisfactory performance, yielding RMSE between 19.1 and 29.5 and MAE between 14.3 and 20.3 when the z-scores were used as prediction class variable and RMSE between 19.0 and 30.3 and MAE between 14.0 and 22.2 when the phone durations in milliseconds were predicted directly. Regarding the CC, the models achieved performance between 0.43 and 0.75 and between 0.54 and 0.83 when the z-scores and when the phone durations in milliseconds were used as prediction class variable respectively, which is a considerably good outcome.

As can be seen in Table 2-4 and as was mentioned above, in almost all the models and all the emotional categories, the models which were built using the z-scores as prediction variable achieved better performance than the respective ones predicting directly the phone duration in milliseconds. Furthermore, the methods with the overall best performance were the M5p model trees, as well as the meta-learning algorithms which used M5p-R regression trees as base classifiers (AR-M5p-R, BG-M5p-R). Moreover, it can be noticed that LR had a very satisfactory performance too, together with M5p regression trees (M5p-R). It is interesting to remark that between the two lazy learning methods which were applied, IB12 rather than LWL performed better in all cases. Furthermore in the case of Sadness emotional category the IB12 model predicting phone duration directly in milliseconds had the best performance along with the AR-M5p-R model. Finally, the REPTrees models appear to have the lowest performance, comparing to the others, both as single prediction method and as base-classifier for the case of AR and BG algorithms.

Comparison among the emotions: Here, it is interesting to compare the experimental results on the basis of the performance of the models among the emotional categories. As can be seen in the Table 2 and 3, a tendency in the case of some emotions to show lower RMSE and MAE values than others exists. This means that the same algorithms managed to achieve lower errors in some emotional categories than in others.

Joy and Fear emotional categories presented the lowest values for RMSE and MAE, independently of which algorithm was applied. In Joy category the RMSE did not overcome 24.5 and MAE had a maximum of 17.9 when the z-scores where used as prediction class variable and 24.5 and 17.9 respectively when the phone durations in milliseconds were predicted directly (both for the case of REPTrees). In Fear category the RMSE did not overcome 23.0 and MAE had a maximum of 17.2 when the z-scores where used as prediction class variable and 24.4 and 18.2 respectively when the phone durations in milliseconds were predicted directly (both for the case of LWL). Sadness and Anger emotional categories had slightly higher errors. Finally the Neutral category achieved the lowest performance, with maximum RMSE of 29.5 and MAE of 19.6 when the z-scores where used as prediction class variable and 29.4 and 20.1 respectively when the phone durations in milliseconds were predicted directly (both for the case of REPTrees). The same tendency is shown on the CC, where the highest values were for the emotional category of Anger and then followed by that of the Joy, Fear, Sadness and Neutral categories.

DISCUSSION

All the applied algorithms managed to build models which perform adequately on the task of phone duration modeling. The M5p models accomplished the best performances 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.

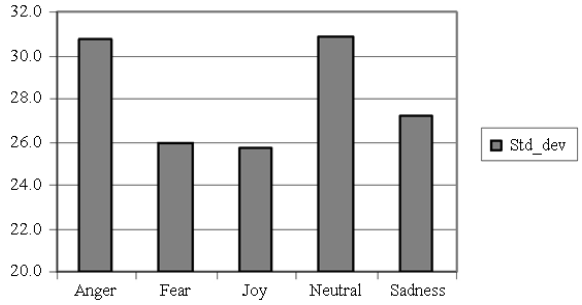

Fig. 1: Weighted average values of standard deviations in milliseconds of phone durations for all the emotion

The M5' algorithm minimizes the error at each interior node recursively until all or almost all of the instances are correctly predicted. In this way, although the computational cost increases, very robust models are constructed. Moreover, as it was expected, the models based on the meta-learning algorithms managed to build robust models by taking advantage of the information that is produced from other methods, as they process meta-data. However, it must be pointed out that the Additive Regression and the Bagging models performed better when were combined with a robust prediction method such as M5p-R, while they didn't perform that well when the REPTrees were used as a base classifier instead. This leads to the conclusion that the choice of the appropriate classifiers is an important issue when meta-algorithms are applied. Moreover, it should be noticed that lazy learning methods or methods that apply a more strict strategy of 'pruning' built models with lower computational cost, but achieve lower performance.

Finally it is interesting to point out that the emotional categories, the phones of which had the lowest values of standard deviation, namely the ones which had more uniform distribution of the mean duration of each phone, were the ones with the lowest prediction errors. As shown in Fig. 1, for the categories of Joy and Fear the weighted average of standard deviations of the phones was the lowest and therefore the phone duration models performed better.

CONCLUSION

In this research, we coped with the task of phone duration modeling on Greek emotional speech implementing various machine learning techniques such as: model and regression trees, linear regression, lazy learning algorithms and meta-learning algorithms. The emotional speech database which was used on this task consisted of five archetypal emotional categories: anger, fear, joy, neutral and sadness. The results showed that all the machine learning algorithms managed to build robust phone duration models. The model trees based on the M5' algorithm and the meta-learning algorithms using regression trees based on the M5' algorithm as base classifier, achieved the best performances. Finally the models built using the emotional categories with the most uniform distributions of the phone durations achieved the best performances.

ACKNOWLEDGEMENT

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

REFERENCES

Aha, D., D. Kibler and M. Albert, 1991. Instance-based learning algorithms. J. Mach. Learn., 6: 37-66. DOI: 10.1023/A:1022689900470

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control, 19: 716-723. DOI: 10.1109/TAC.1974.1100705

Arvaniti, A. and M. Baltazani, 2000. Greek ToBI: A system for the annotation of Greek speech corpora. Proceeding of the 2nd International Conference on Language Resources and Evaluation, May 31-June 2, European Language Resources Association, Athens, Greece, pp: 555-562. http://ling.ucsd.edu/~arvaniti/A&B-LREC.pdf.

Atkeson, C.G., A.W. Moorey and S. Schaal, 1996. Locally weighted learning. Artifi. Intel. Rev., 11: 11-73. DOI: 10.1023/A:1006559212014

Bartkova, K. And C. Sorin, 1987. A model of segmental duration for speech synthesis in French. Speech Commun., 6: 245-260. DOI: 10.1016/0167-6393(87)90029-X

Black, A. and K. Lenzo, 2000. Building voices in FESTIVAL speech synthesis system. http://festvox.org/bsv/bsv.pdf

Breiman, L., 1996. Bagging predictors. J. Mach. Learn., 24: 123-140, DOI: 10.1023/A:1018054314350.

Carison, R. and B. Granstrom, 1988. A search for durational rules in real speech database. Phonetica, 43: 140-154. DOI: 10.1159/000261766

Chen, S.H., S.H. Hwang and Y.R. Wang, 1998. An RNN-based prosodic information synthesizer for mandarin text to speech. IEEE Trans. Speech Audio Process., 6: 226-239. DOI: 10.1109/89.668817

Chien, J.T. and C.H. Huang, 2003. Bayesian learning of speech duration models. IEEE Trans. Speech Audio Process., 11: 558-567. DOI: 10.1109/TSA.2003.818114

Epitropakis, G., D. Tambakas, N. Fakotakis and G. Kokkinakis, 1993. Duration modeling for the Greek language. Proceeding of the 3rd European Conference on Speech Communication and Technology, Sept. 22-25, ISCA, Berlin, Germany, pp: 1995-1998. http://www.isca-speech.org/archive/eurospeech_1993/e93_1995.html

Inanoglu, Z. and S. Young, 2009. Data-driven emotion conversion in spoken English. Speech Commun., 51: 268-283. DOI: 10.1016/j.specom.2008.09.006

Jiang, D.N., W. Zhang, L. Shen and L.H. Cai, 2005. Prosody analysis and modeling for emotional speech synthesis. Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Process Mar. 18-23, IEEE Xplore Press, USA., pp: 281-284. DOI: 10.1109/ICASSP.2005.1415105

Kaariainen, M. And T. Malinen, 2004. Selective rademacher penalization and reduced error pruning of decision trees. J. Mach. Learn. Res., 5: 1107-1126. http://www.jmlr.org/papers/volume5/kaariainen04a/kaariainen04a.pdf

Klatt, D.H., 1979. Synthesis by Rule of Segmental Durations in English Sentences. In: Frontiers of Speech Communication Research, Lindlom, B. and S. Ohman (Eds.). Academic Press, New York, USA., ISBN: 10:0124498507, pp: 287-300.

Klatt, D.H., 1987. Review of text to speech conversion for English. J. Acoustical Soc. Am., 82: 737-793. DOI: 10.1121/1.395275

Kominek, J. and A.W. Black, 2003. CMU ARCTIC databases for speech synthesis, CMU-LTI-03-177. Language Technologies Institute, School of Computer Science, Carnegie Mellon University. http://festvox.org/cmu_arctic/cmu_arctic_report.pdf

Lazaridis, A., P. Zervas and G. Kokkinakis, 2007. Segmental duration modeling for Greek speech synthesis. Proceeding of the 19th IEEE International Conference on Tools with Artificial Intelligence, Oct. 29-31, IEEE Computer Society, Washington DC., USA., pp: 518-521. DOI: 10.1109/ICTAI.2007.163

Mitchell, T., 1997. Decision Tree Learning, In: Machine Learning, Mitchell, T. (Ed.). McGraw-Hill Companies, Inc., ISBN: 10: 0071154671, pp: 52-78.

Mobius, B. and P.H.J. Santen, 1996. Modeling segmental duration in German text to speech synthesis. Proceeding of the International Conference on Spoken Language, Oct. 3-6, IEEE Xplore Press, Philadelphia, PA., pp: 2395-2398. DOI: 10.1109/ICSLP.1996.607291

Oatley, K. and P. Johnson-Laird, 1998. The Communicative Theory of Emotions. In: Human Emotions: A Reader, Jenkins, J., K. Oatley and N. Stein (Eds.). Oxford, Blackwell, ISBN: 0631207473, pp: 84-87.

Quinlan, R.J., 1992. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Nov. 16-18, World Scientific Pub Co Inc., pp: 343-348.

Rao, K.S. and B. Yegnanarayana, 2007. Modeling durations of syllables using neural networks. Comput. Speech Language, 21: 282-295. DOI: 10.1016/j.csl.2006.06.003

Santen, J.P.H., 1992. Contextual effects on vowel durations. Speech Commun., 11: 513-546. DOI: 10.1016/0167-6393(92)90027-5

Stone, C.J., 1985. Additive regression and other nonparametric models. Ann. Stat., 13: 689-705. DOI: 10.1214/aos/1176349548

Takeda, K., Y. Sagisaka and H. Kuwabara, 1989. On sentence-level factors governing segmental duration in Japanese. J. Acoust. Soc. Am., 86: 2081-2087. DOI: 10.1121/1.398467

Tesser, F., P. Cosi, C. Drioli and G. Tisato, 2005. Emotional festival-mbrola TTS synthesis. http://www.istc.cnr.it/doc/75a_2006041311625t_tf-INTERSPEECH2005-03.pdf

Wang, Y. and I.H. Witten, 1997. Induction of model trees for predicting continuous classes. Proceeding of the Poster Papers of the European Conference on Machine Learning (ECML'97), Springer, Prague, Czech Republic, pp: 128-137.

Witten, H.I. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufmann Publishing, ISBN: 10: 0-12-088407-0, pp: 525.