IMPROVING PHONE DURATION MODELLING USING SUPPORT VECTOR REGRESSION FUSION

Alexandros Lazaridis, Iosif Mporas, Todor Ganchev, George Kokkinakis and Nikos Fakotakis

Artificial Intelligence Group, Wire Communications Laboratory,
Department of Electrical and Computer Engineering,
University of Patras, 26500 Rion-Patras, Greece
Tel. +30 2610 996496, Fax. +30 2610 997336
alaza@upatras.gr

ABSTRACT

In the present work, we propose a scheme for the fusion of different phone duration models, operating in parallel. Specifically, the predictions from a group of dissimilar and independent to each other duration models are fed to a machine learning algorithm which yields more precise phone duration predictions. We firstly study the performance of eight independent models. Here, among others, we investigate the effectiveness of Support Vector Regression (SVR), which has not yet been studied on the phone duration prediction task. At the fusion stage, we evaluate the performance of several regression algorithms. The evaluation was carried out on the American-English KED TIMIT and the Greek WCL-1 databases. On both databases, the SVR-based model demonstrated the lowest error rate. When compared to the second-best algorithm, we report a relative reduction of the mean absolute error (MAE) and the root mean square error (RMSE) by 5.5% and 3.7% on KED TIMIT, and 6.8% and 3.7% on WLC-1. The proposed fusion scheme, implemented with SVR fusion, improved the accuracy over the one of the best individual predictor, by 1.9% and 2.0% in terms of relative reduction of the MAE and RMSE on KED TIMIT, and by 2.6% and 1.8% on WCL-1 database.

Index Terms— Duration modelling, parallel fusion scheme, phone duration prediction, support vector regression, text-to-speech synthesis

1. INTRODUCTION

In Text-to-Speech synthesis (TTS) there are two major issues concerning the quality of the synthetic speech, namely the intelligibility and the naturalness (Dutoit, 1997; Klatt, 1987). The former refers to the capability of a sythesized word or phrase to be comprehended by the average listener. The latter represents how close to the human natural speech, the synthetic speech is perceived. One of the most important factors for achieving intelligibility and naturalness in synthetic speech is the accurate modelling of prosody.

Prosody can be regarded as the implicit channel of information in the speech signal that conveys information about expression of emphasis, attitude, assumptions, emotional state of the speaker that provides to the listener clues supporting the recovery of the verbal message (Huang et al., 2001). The accurate modelling and control of prosody in a text-to-speech system leads to synthetic speech of higher quality. Prosody is shaped by the relative level of the fundamental frequency, the intensity (i.e. the ???), the energy of the speech signal and last but not least by the duration of the pronounced phones (Dutoit, 1997; Furui, 2000). The duration of the phones controls the rhythm and the tempo of speech (Yamagishi et al., 2008) and the flattening of the prosody in a speech waveform would result in a monotonous, neutral, toneless and without rhythm synthetic speech, sounding unnatural, unpleasant to the listener or sometimes even scarcely intelligible (Chen et al., 2003). Thus, the accurate modelling of phones' duration is essential in speech processing.

Several areas of speech technology, among which TTS, automatic speech recognition (ASR) and speaker recognition benefit from duration modelling. In TTS, the correct segmental duration contributes to the naturalness of synthetic speech (Chen et al., 1998; Klatt, 1976). In hidden Markov model (HMM)-based ASR, state duration models improve the speech recognition performance (Bourlard et al., 1996; Jennequin and Gauvian, 2007; Levinson, 1986; Mitchell et al., 1995; Pols et al., 1996). Finally, significant improvement of the performance in the speaker recognition task was achieved by Ferrer et al. (2003), when duration-based speech parameters were used for the characterization of the speaker's voice.

Various approaches for segment duration modelling and many factors influencing the segmental duration have been studied in the literature (Bellegarda et al., 2001; Crystal and House, 1988; Edwards and Beckman, 1988; Riley, 1992; Shih and Ao, 1997; van Santen, 1994). The features related to these factors can be extracted from several levels of information, such as the phonetic, the prosodic, the linguistic and the syntactic level. With respect to the way duration models are built, the duration prediction approaches can be divided in two major categories: the rule-based (Klatt, 1976) and the data-driven methods (Campbell, 1992; Chen et al., 1998; Lazaridis et al., 2007; Monkowski et al., 1995; Rao and Yegnanarayana, 2005; Riley, 1992; Takeda et al., 1989; van Santen, 1992).

The rule-based methods use manually produced rules, extracted from experimental studies on large sets of utterances, or based on previous knowledge. The extraction of these rules requires expert phoneticians. In the most prominent attempt in the rule-based duration modelling category, proposed by Klatt (1976), rules which were derived by analyzing a phonetically balanced set of sentences, were used in order to predict segmental duration. These rules were based on linguistic and phonetic information such as positional and prosodic factors. Initially a set of intrinsic (starting) values was assigned on each phone which was modified each time according to the extracted rules. Models of this type and similar to this were developed in many languages such as French (Bartkova and Sorin, 1987), Swedish (Carlson and Granstrom, 1986), German (Kohler, 1988) and Greek (Epitropakis et al., 1993; Yiourgalis and Kokkinakis, 1996), as well as in several dialects such as American English (Allen et al., 1987; Olive and Liberman, 1985) and Brazilian Portuguese (Simoes, 1990). The main disadvantage of the rule-based approaches is the difficulty to represent and tune manually all the linguistic, phonetic and prosodic factors which influence the segmental duration in speech. As a result, it is very difficult to collect all the appropriate (or even enough) rules without long-term devotion to this task (Klatt, 1987). Consequently 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 the corresponding rules (Rao and Yegnanarayana, 2007).

Data-driven methods for the task of phone duration modelling were developed after the construction of large databases (Kominek and Black, 2003). Data-driven approaches overcame the problem of the extraction of manual rules by employing either statistical methods or artificial neural network (ANN) based techniques which automatically produce phonetic rules and construct duration models from large speech corpora. Their main advantage is that this process is semi or fully automated and thus significantly reduces the efforts that have to be spent from phoneticians.

Several machine learning methods have been used in the phone duration modelling task. The linear regression (LR) (Takeda et al., 1989) models are based on the assumption that among the features which affect the segmental duration there is linear independency. These models achieve reliable predictions even with small amount of training data but do not model the dependency among the features. On the other hand, decision tree models (Monkowski et al., 1995) and in particular classification and regression tree (CART) models (Riley, 1992), which are based on binary splitting of the feature space, can represent the dependencies among the features but cannot insert constraints of linear independency for reliable predictions (Iwahashi and Sagisaka, 2000). Another technique which has been used on the phone duration modelling task is the sums-of-products (SOP), where the segment duration prediction is based on a sum of factors and their product terms that affect the duration (van Santen, (1992, 1994)). The advantage of these models is that they can be trained with a small amount of data. Bayesian networks models have also been introduced on the phone duration prediction task. These models incorporate a straightforward representation of the problem domain information and despite their time consuming training phase, they can make accurate predictions even when unknown values come across in some features (Goubanova and King, 2008; Goubanova and Taylor, 2000). Furthermore, instance-based algorithms (Lazaridis et al. 2007) have been used in phone duration modelling. In instance-based approaches the training data are stored and a distance function is employed during the prediction phase in order to determine which member of the training set is closer to the test instance and predict the phone duration. In a recent study (Yamagishi et al., 2008), the gradient tree boosting (GTB) (Friedman, 2001; Friedman, 2002) approach was proposed for the phone duration modelling task as an alternative to the conventional approach using regression trees. The GTB algorithm is a meta-algorithm which is based on the construction of multiple regression trees and consequently taking advantage of them.

On the task of syllable duration modelling various neural networks have been used, including feedforward neural networks (Campbell, 1992; Rao and Yegnanarayana, 2007) and recurrent neural networks (RNN) (Chen et al., 1998). Furthermore, in the case of syllable duration prediction the SVM regression model has been used in order to perform the function estimation from the training instances using non-linear mapping of the data onto a high-dimensional feature space (Rao and Yegnanarayana, 2005). In (Iwahashi and Sagisaka, 2000) a new scheme for statistical modelling of prosody control in speech synthesis was proposed. It is based on a constrained tree regression scheme that relies on hierarchical combination of different algorithms. It offers a mechanism for evading the disadvantages inherent to one algorithm by benefiting from information provided by another algorithm. This can be explained by the observation that different algorithms perform better in different conditions.

As a result, the task of phone duration modelling based on the data-driven approaches gives the ability to overcome the time consuming labour of the manual extraction of the rules which are needed in the rule-based approaches. However, as shown by van Santen and Olive (1990), these methods are not always satisfactory for the task of phone duration prediction.

All previous studies on phone and syllable duration modelling are restricted to the use of a single linear or non-linear regression algorithm. The only exception to this trend is the work of Iwahashi and Sagisaka (2000), where a hierarchical structure for syllable duration prediction using the outputs of a phone duration model was used. However, this structure is restricted to the post-processing of a single duration prediction model, and no extension to a parallel regression fusion of the duration predictions of multiple models has been studied.

In the present work, aiming at improving the accuracy of the prediction of the segmental durations (i.e. phones), we propose a fusion scheme based on the use of multiple dissimilar phone duration predictors which operate on a common input, and whose predictions are combined using a regression fusion method. The proposed scheme is based on the observation that predictors implemented with different machine learning algorithms perform differently in dissimilar conditions. Hence, we suppose that an appropriate combination of their outputs could result in a new set of more precise phone duration predictions. Thus, an appropriate fusion scheme that can learn how to combine the outputs of a number of individual predictors in a beneficial manner, will contribute to the reduction of the overall prediction error, when compared to the error of each individual predictor.

Based on this assumption, we investigate various implementations of the proposed fusion scheme and study its accuracy for duration prediction on different levels of granularity: vowels/consonants, phonetic category, individual phones, etc. In this connection, initially, we investigate the performance of eight linear and non-linear regression algorithms, five of them already examined in previous studies (Iwahashi and Sagisaka, 2000; Lee and Oh, 1999; Riley, 1992; Takeda et al., 1989; Yamagishi et al., 2008) as individual predictors. These are based on linear regression and decision trees -- model trees, regression trees and pruning decision trees. Furthermore, another two of them (which?) are modifications of algorithms that are already studied in the phone duration prediction task (Yamagishi et al., 2008), and finally, the support vector regression (SVR) algorithm, which to our best knowledge has not yet been employed on the phone duration prediction task. Next, the durations predicted by the individual duration models are fed as inputs to a machine learning algorithm referred to as fusion model, which uses these predictions and produces the final prediction. For the purpose of fusion, we evaluate twelve different (linear and non-linear) regression fusion techniques, among which are linear regression, decision trees, support vector regression, neural networks, etc.

The present study was inspired by the work of Kominek and Black (2004), where a family of acoustic models, providing multiple estimates for each boundary point, was used for segmenting a speech database, creating a more robust unit selection corpus-based synthetic speech. This approach was found more robust than a single estimate, since by taking consensus values large labelling errors are less prevalent in the synthesis catalogue, which improves the resulting synthetic speech. To the extent of our knowledge, a parallel regression fusion of independent models has not yet been studied on the phone duration prediction or on the syllable duration prediction tasks. Furthermore, although SVR models have been used for syllable duration prediction (Rao and Yegnanarayana, 2005), to this end, they have not been employed on the phone duration prediction task.

The remainder of this article is organized as follows. In Section 2 we outline the proposed fusion scheme. In Section 3 we briefly outline the individual phone duration modelling algorithms, the algorithms used in the fusion scheme, the speech databases and the experimental setup used in the evaluation. The experimental results are presented and discussed in Section 4 and finally this work is concluded in Section 5.

2. FUSION SCHEME FOR DURATION MODELLING AND PREDICTION

Phone duration modelling, which mainly relies on regression algorithms, suffers from specific types of errors. The most commonly occurring type of error is the bias (systematic) error (Freedman et al., 2007). This error is a constant shift of the predicted phone durations from the real ones and can be estimated as the difference between the real and predicted mean durations. Other prediction errors that may occur in the phone duration modelling task are small miss-predictions and gross errors (outliers) (Freedman et al., 2007). Small miss-predictions in phone duration, i.e. less than 20 milliseconds, do not significantly affect the quality of the synthetic speech signal, but the other errors degrade the quality of synthetic speech (Wang et al., 2004). Here, we assume that an appropriate combination of the predictions of a number of dissimilar and independent phone duration prediction models will improve the overall phone duration prediction accuracy. This is because different phone duration models will err in a dissimilar manner, and the fusion of their outputs, through a machine learning technique, would be able to learn and compensate some of these errors. Especially, we suppose that such a fusion scheme, apart from improving the overall accuracy of duration prediction, will be able to reduce the outliers.

In Fig. 1, we present the block diagram of the proposed fusion scheme, which relies on the combination of predictions that are produced by multiple dissimilar phone duration models, which operate on a common input. As the figure presents, the predictions of the individual models are introduced into the fusion stage, where a machine learning algorithm uses them for obtaining more precise phone duration predictions. The training and the operational phases of the proposed fusion scheme is discussed in the following subsections.

Figure 1 will be placed here

2.1 *Training of the fusion scheme*

The training of the proposed fusion scheme is an off-line two-step procedure, which relies on two non-overlapping datasets: the training and the development data. The training process can be summarized concisely as follows. During the first training step, the independent phone duration models are created using the training dataset. Subsequently, at the second step, these models are employed to process the development dataset. The outcome of this processing is a set of predictions, which together with the ground truth labels (manually annotated tags) serve as input for training the adjustable parameters of the fusion algorithm. This procedure can be formalized as follows:

Let us define a set of N independent phone duration prediction models, DM_n, with $1 \le n \le N$. The input feature vector, X_j^p, for the jth instance ($1 \le j \le J$) of the phone p, which is used for training the N independent phone duration models, DM_n, is defined as:

$$X_j^p = [\theta_1, \theta_2, ..., \theta_m, ..., \theta_M]^T, \quad j=1,2,\ldots,J, \tag{1}$$

where $X \in \mathbb{R}^M$ ($\mathbb{R}$ is the feature space) and θ_m is the mth feature ($1 \le m \le M$) of the feature vector X_j^p.

Once the independent duration models are trained, they are fed with the development dataset. The outcome of its processing is the set of phone duration predictions, $y_j^{p,n}$, of the nth duration model for the jth instance of the phone p, to be predicted:

$$y_j^{p,n} = f_n^p(X_j^p), \qquad j=1,2,\dots,J, \tag{2}$$

where $y_j^{p,n} \in \mathbb{R}^N$. The vector, Y_j^p, formed by appending the independent phone duration predictions, $y_j^{p,n}$,

$$Y_j^p = \left\{ y_j^{p,n} \right\}^T, \qquad j=1,2,\dots,J, \tag{3}$$

where $1 \le n \le N$, for the jth instance of the phone p, are used in the training of the fusion algorithm. Once the fusion stage is trained, the proposed composite phone duration modelling scheme, shown in Fig. 1, is ready for operation.

2.2 Operation of the fusion scheme

In the operational mode, the input vector, X_j^p, for the jth instance of the phone p of the test dataset, appears as input to the N individual phone duration prediction models, DM_n, with $1 \le n \le N$ (refer to Fig.1). Their outputs, $y_j^{p,n}$, as computed in eq. 2, form the vector of predictions Y_j^p, which serves as input for the fusion stage. At the fusion stage the vector Y_j^p is processed by the fusion algorithm, which computes the final phone duration prediction for the jth instance as:

$$O_j^p = g^p(Y_j^p), \qquad j=1,2,\dots,J, \tag{4}$$

with $O_j^p \in \mathbb{R}$.

The fusion of multiple different predictions is expected to contribute to the reduction of the types of errors described above (first paragraph in section 2), and thus contributes to the decrease of the overall error rate. This expectation is based on the observation that different predictors, which rely on different machine learning algorithms, err in a dissimilar manner. Employing an appropriate fusion scheme, which is capable to learn the proper mapping between a set of noisy predictions and the true phone duration values, could turn out beneficial in terms of improved accuracy.

3. EXPERIMENTAL SETUP

To investigate the practical usefulness of the proposed approach, we trained several individual phone duration models, and then employed them in the fusion scheme described in Section 2. The various individual phone duration models and the fusion algorithms involved in the phone duration prediction fusion scheme, as well as the speech databases used in the experiments and the experimental protocol that was followed, are described in the following subsections.

3.1 Individual phone duration models

In the present work we consider eight different machine learning algorithms for phone duration modelling, the outputs of which are then fed to the fusion model. These algorithms are well known and have successfully been used over the years, in different modelling tasks. One exception is the support vector regression (SVR) based modelling, which to this end has not been employed on the phone duration modelling task. In brief, the eight individual phone duration modelling algorithms that we consider here are:

(i) the linear regression (LR) (Witten and Frank, 1999) using Akaike's Information Criterion (AIC) (Akaike, 1974) in backward stepwise selection (BSS) (Kohavi and John, 1997) procedure eliminating unnecessary variables of the training data,

(ii) the m5p model tree, using a linear regression function on each leaf, and the m5pR regression tree, using a constant value on each leaf node instead (Quinlan, 1992; Wang and Witten, 1997).

(iii) two additive regression algorithms (Friedman, 2002) and two bagging algorithms (Breiman, 1996) were used, by using two different regression trees (m5pR and REPTrees) (Kaariainen and Malinen, 2004; Quinlan, 1992; Wang and Witten, 1997) as base classifiers in each case. The latter four algorithms are meta-learning algorithms (Vilalta and Drissi, 2002) using regression trees as base classifiers.

During the training process, 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. In these two cases of additive regression meta-classification, the shrinkage parameter, v, indicating the learning rate, was set equal to 0.5 and the number of the regression trees, *rt-num*, was set equal to ten. These values were selected after grid-search experiments (v={0.1, 0.3, 0.5, 0.7, 0.9}, *rt-num*={5, 10, 15, 20}) on a randomly selected subset of the training data, of size approximately equal to 20% of the size of the training set.

In the bagging algorithm, the dataset was split in multiple subsets using a regression tree for each of them. The final prediction value is the average of the values predicted from each regression tree. In a similar manner, the number of the regression trees (*rt-num*) was set equal to ten after a number of grid-search experiments (*rt-num*={5, 10, 15, 20}) on the randomly selected subset of the training data.

(iv) Finally, the support vector regression (SVR) model (Platt, 1999), which employs the sequential minimal optimization (SMO) algorithm for training a support vector classifier (Smola and Scholkopf, 1998), was used. Many kernel functions have been used in SVR such as the polynomial, the radial basis function (RBF) and the Gaussian functions (Scholkopf and Smola, 2002). In our experiments the RBF kernel was used as mapping function. The ε and C parameters, where $\varepsilon \geq 0$ is the maximum deviation allowed during training and $C > 0$ is the penalty parameter for exceeding the allowed deviation, were set equal to 10^{-3} and 10^{-1} respectively. This was done after a grid search (ε={10^{-1}, 10^{-2}, …, 10^{-5}}, C={0.05, 0.1, 0.3, 0.5, 0.7, 1.0, 10, 100}) on a randomly selected subset of the training set, representing 20% of the size of the full training set.

Our motivation to select these algorithms was based on previous research (Iwahashi and Sagisaka, 2000; Lee and Oh, 1999; Riley, 1992; Takeda et al., 1989; Yamagishi et al., 2008), where these algorithms were reported successful on the segmental duration modelling task. Along with the phone duration prediction task, many of these algorithms have also been used in syllable duration prediction task, supporting different languages and databases.

3.2 Fusion algorithms

In order to select the most advantageous fusion method, we evaluated ten different machine learning algorithms for numerical prediction. These are the eight algorithms outlined in Section 3.1, as well as (i) the radial basis function neural network (RBFNN) with Gaussian kernel (Park and Sandberg, 1993), and (ii) the instance-based algorithm (IBK) (Aha and Kibler, 1991), which is a *k*-nearest neighbours classifier.

In the case of RBFNN, the *k*-means algorithm is used as a first step in the training process, for the estimation of the centres of the radial basis units in the hidden layer on the network. The outputs of the hidden layers are combined with linear regression. The number of clusters (*num-cl*) for the *k*-means to generate and the minimum standard deviation (*cl-std*) for the clusters were set equal to 135 and 10^{-2} respectively. These parameters were determined after a grid search (*num-cl*={5,10, …, 200}, *cl-*

std={0.001, 0.01, 0.1, 0.5}) on a randomly selected subset of the training set, consisting of approximately 20% of the set.

In the case of IBK a linear nearest neighbours search algorithm was used, employing the Euclidean distance as a distance function. Leave-one-out cross-validation was used to select the best value for k, under the restriction, $k \leq 35$, i.e. an upper limit of 35 nearest neighbours. The predictions from the k nearest neighbours were weighted according to the inverse distance.

Furthermore, in addition to the ten machine learning methods described so far, we implemented two well-known linear combination fusion schemes, which serve as intuitive reference points in this study. These are: (i) the average of all individual predictions, in which the mean value of the prediction estimations of all the independent duration models was regarded as the final phone duration prediction value (average linear combination) and (ii) the selection of the best prediction among the individual phone duration models for each instance (best-case selection). The selection of the best duration model per instance can be performed for different categories of clusters, such as voiced/unvoiced, vowels/consonants, phonetic category, individual phones, tri-phones, etc.

In the best-case selection fusion scheme, we relied on the root mean square error (RMSE) of each phone prediction algorithm over the development data as the criterion for the selection of the best model for each case (Chen et al., 1998; Goubanova and King, 2008; Yamagishi et al., 2008). Specifically, the duration model prediction with the lowest RMSE for the cluster (vowels/consonants, phonetic category and individual phones) of each instance was selected.

3.3 Databases and feature set

In the evaluation experiments we used two databases: the American-English speech database CSTR US KED TIMIT (CSTR, 2001) and the Modern Greek speech prosodic database, WCL-1 (Zervas et al., 2008). KED TIMIT consists of 453 phonetically balanced sentences (3400 words approximately) uttered by a Native American male speaker. The WCL-1 prosodic database consists of 5500 words distributed in 500 paragraphs, each one of which may be a single word, a short sentence, a long sentence, or a sequence of sentences uttered by a female professional radio actress. The final corpus includes 390 declarative sentences, 44 exclamation sentences, 36 decision questions and 24 “wh” questions.

For the experiments on the KED TIMIT database, we adopted the phone set provided with the database (CSTR, 2001) which consists of 44 phones. For the experiments using the WCL-1 database we adopted the phone set provided with the database (Zervas et al., 2008) consisting of 34 phones. In all experiments, the manually labelled phone durations were used as the ground truth (reference) durations. In this work, a number of features, which have been reported successful in the literature (Crystal and House, 1988; Campbell, 1992; Klatt, 1987; Goubanova and King, 2008; Riley, 1992; van Santen, 1994), are considered for the task of phone duration prediction. From each utterance we computed 33 features along with the contextual information concerning some of these features, described next:

(i) eight phonetic features: the phone class (consonants/non-consonants), the phone types (short vowels, long vowels, diphthongs, schwa, consonants), the vowel height (high, middle or low), the vowel frontness (front, middle or back), the lip rounding (rounded/unrounded), the manner of production (consonant type), the place of articulation (labial, labio-dental, dental, alveolar, palatal, 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.

(ii) three segment-level features: the phone name with the information of the neighbouring 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).

(iii) thirteen syllable-level features: the position in the word type of the syllable (single, initial, middle or final) with the information of the neighbouring instances (previous, next), the number of all the syllables in the word, the number of the accented syllables and the number of the stressed syllables since the last and to the next phrase break (i.e. the ???), 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).

(iv) two word-level features: the part-of-speech (noun, verb, adjective, etc) and the number of syllables of the word.

(v) one phrase-level features: the syllable break (i.e. the phrase break after the syllable) with the information of the neighbouring (two previous, two next) instances.

(vi) six accentual features: the ToBI accents and boundary tones with the information of the neighbouring (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 neighbouring (two previous, two next) instances.

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

In all experiments we followed an experimental protocol based on 10-fold cross-validation. Specifically, in each fold the training data were split in two portions, the training dataset and the development dataset. The former, amounting to approximately 60% of the full dataset, was used for the training of the individual phone duration predictors, and the latter, amounting to approximately 30% of the full dataset, for the training of the fusion algorithm. Furthermore, the test dataset, amounting to approximately 10% of the full dataset, was used for evaluating the performance of the eight independent duration prediction algorithms, as well as the performance of the fusion scheme.

3.4 Performance Metrics

The experimental results were evaluated using the two most commonly used figures of merit, namely the mean absolute error (MAE) and the root mean squared error (RMSE), between the predicted duration and the actual (reference) duration of each phone (Chen et al., 1998; Goubanova and King, 2008; van Santen, 1992; Yamagishi et al., 2008). Due to the squaring of values in the RMSE, large errors (outliers) are weighted heavily, which makes this figure of metric more sensitive to outliers than the MAE (Witten and Frank, 1999). This sensitivity of the RMSE makes it a more illustrative measurement concerning the outliers, e.g. the gross errors, when compared to the MAE.

4. EXPERIMENTAL RESULTS

In the present work, we consider clustering of the instances on the basis of (i) vowels/consonants categorization, (ii) phonetic categories and (iii) individual phones. This offers different degree of detail and allows us to gain insights about the advantages and disadvantages of each algorithm. The same clustering of the instances is used in the best-case selection fusion model. In this fusion model, as mentioned in Section 3.2, the criterion for the selection of the best model for each case is the RMSE of each phone duration prediction algorithm over the development data, as in (Chen et al., 1998; Goubanova and King, 2008; Yamagishi et al., 2008). Specifically, the phone duration model prediction with the lowest RMSE per cluster (vowels/consonants, phonetic categories, individual phone) of each instance is selected.

4.1 Duration prediction with individual phone duration models

As a first step we examined the performance of the eight individual algorithms on both databases using the entire feature set described in Section 3. The RMSE, the MAE and the standard deviation of the absolute error (STD of AE) for all individual algorithms specified in Section 3.1 are shown in Table 1, where Table 1 (*a*) presents the results obtained on KED TIMIT, and Table 1 (*b*) the ones on the WCL-1 database. The results of the best performing model, among the eight individual prediction models, are in bold. As can be seen, on both databases, the proposed support vector regression (SMOreg) model, implemented with the SMO regression algorithm, outperforms all the other models. Specifically, on the KED TIMIT database the SMOreg model outperformed the second-best model, i.e. the meta-classifier additive regression using m5pR model, by approximately 5.5% and 3.7% in terms of MAE and RMSE respectively. On the WCL-1 database the SMOreg model outperformed the second-best model, i.e. the Linear Regression model, by approximately 6.8% and 3.7% in terms of MAE and RMSE respectively. This advantage of the SMOreg models, on both databases, is owed to the advantage of SVMs to cope better with high-dimensional feature spaces (Vapnik, 1995; Vapnik, 1998), when compared to the other classification and regression algorithms.

Table 1 (a)

Table1 (b)

In Table 2 we present the performance per phonetic category as well as for the vowel/consonant categorization of the eight independent phone duration models, implemented by different algorithms, on the KED TIMIT (Table 2 (*a*)) and the WCL-1 (Table 2 (*b*)) databases. As can be seen, the SMOreg model demonstrates the lowest RMSE on both databases, in all cases except for the Affricates on KED TIMIT, where the lowest RMSE is observed for the REPTrees and the SMOreg achieves the second-best performance.

Table 2 (a)

Table 2 (b)

In Table 3 the phone duration prediction results obtained on the level of individual phones are presented. Specifically, Table 3 (*a*) shows the RMSE for the 44 phone set of the KED TIMIT database and Table 3 (*b*) for the 34 phone set of the WCL-1 database. The results for the best performing algorithm are in bold. As shown in the tables, despite the fact that the SMOreg model demonstrates the highest overall performance on both databases (refer to Table 1), in one phonetic category (Affricates in Table 2 (*a*)) and in some particular phones, such as *ch*, *ay*, etc (Table 3), other models offer a higher phone duration prediction accuracy. For instance, on the KED TIMIT database, the highest accuracy for the phone *ch* is observed for the Linear Regression model, while for the phone *ay* the highest accuracy is for the m5p model (refer to Table 3 (*a*)). These specific results, and other similar cases shown in the Table 3, are in support of our observation that different algorithms perform better in different phonetic categories and phones. This indicates that an appropriate fusion of the outputs of the individual phone duration prediction models could be beneficial for reducing the overall error rate. Experimental results for various implementations of the fusion stage are presented in Section 4.2.

Table 3 (a)

Table 3 (b)

4.2 Duration prediction with the proposed fusion scheme

In the following, we report the evaluation results for the twelve fusion algorithms outlined in Section 3.3. In Table 4, we present the results obtained in the evaluation of the *average* linear combination and the *best-case selection* techniques for different cases: overall performance, vowel/consonant categories, phonetic categories and individual phones. The results for the best individual phone duration prediction model (SMOreg) are duplicated from Table 1 for the purpose of direct comparison. We consider the *average* and the *best-case selection* techniques as intuitive reference points against which the performance of the other ten fusion algorithms, evaluated here, is compared. As can be seen in Table 4 (*a*), for KED TIMIT, and in Table 4 (*b*), for the WCL-1 database, in the cases of vowel/consonant and phonetic categories the best-case selection algorithm did not offer advantage over the best individual model, SMOreg, because the SMOreg outperforms the other individual predictors in all categories (Table 2). As discussed above the only exception is Affricates on KED TIMIT, where the additive regression with REPTrees algorithm is the best, but is not sufficient for significant advantage of the fusion scheme.

Concerning the clustering according to individual phones, the best-case selection method slightly outperformed the best individual duration model on both databases. In detail, the best-case selection, outperformed the SMOreg model by approximately 0.2% and 0.9% in terms of MAE and RMSE on the KED TIMIT database, and by approximately 0.6% and 0.5% on the WCL-1 database. As can be seen in Table 3, this is owed to the fact that, for both databases, in approximately 35-40% of the phones the best performing algorithm is not the SMOreg. Consequently, the best-case selection fusion scheme outperforms the best individual phone duration prediction model (SMOreg).

Table 4 (a)

Table 4 (b)

In Table 5, we present results for the remaining ten fusion algorithms (refer to Section 3.2): the Linear Regression, the m5p model tree, the m5pR regression tree, the additive regression algorithms based on m5pR and REPTrees, the bagging algorithms based on m5pR and REPTree, the instance based learning (IBK), the support vector regression (SVR) which implements the sequential minimal optimization (SMO) algorithm, and the radial basis function neural network (RBFNN). The best fusion result is shown in bold. For the reason of comparison, in Table 5 we duplicate the results for the best individual phone duration model, SMOreg. As can be seen from the results on the KED TIMIT (Table 5 (a)) and the WCL-1 (Table 5 (b)) databases, the SMOreg fusion model outperformed all the other fusion models that were evaluated here. It is also noteworthy to mention that only the SMOreg fusion model outperformed the best individual duration prediction model. Specifically, the SMOreg fusion model outperformed the individual SMOreg predictor by approximately 1.9% and 2.0% in terms of MAE and RMSE on KED TIMIT, and by approximately 2.6% and 1.8% on the WCL-1 database, respectively. Furthermore, we should point out that the SMOreg fusion model apart from reducing the overall error also reduced the outliers. Specifically, in comparison to the best individual predictor, i.e. the SMOreg model, the SMOreg fusion model reduced the STD of AE by approximately 2.1% on KED TIMIT and by approximately 1.2% on the WCL-1 database, respectively.

Table 5 (a)

Table 5 (b)

Finally, in order to investigate the statistical significance of the difference between the results of the best individual phone duration model (SMOreg) and the results for the best fusion scheme (fusion with SMOreg algorithm) the Wilcoxon test (Wilcoxon, 1945) was carried out. The Wilcoxon test showed that on both databases, the difference between the results for the best individual model and these for the fusion scheme is statistically significant. Specifically, for a significance level of 0.05 the Wilcoxon test estimated a *p*-value equal to $5.77e^{-09}$ and $3.5e^{-11}$ on KED TIMIT and WCL-1, respectively. Consequently, the fusion scheme contributes to the improvement of the accuracy of phone duration prediction, when compared to best predictor among all evaluated individual phone duration prediction models.

5. SUMMARY AND CONCLUSIONS

In this work we studied the accuracy of various machine learning algorithms on the task of phone duration prediction. The experimental results showed that on the task of phone duration prediction, Support Vector Machines (SVM), as a regression model, outperforms various other machine learning techniques. Specifically, in terms of relative decrease of the mean absolute error and root mean square error, the SMO regression model outperformed the second-best model by approximately 5.5% and 3.7% on KED TIMIT, and by approximately 6.8% and 3.7% on the WCL-1 database, respectively.

Furthermore, the proposed fusion scheme, which combines predictions from multiple independent phone duration models, operating on a common input, takes advantage of the observation that different prediction algorithms perform better in different conditions. The experimental validation demonstrated that the fusion scheme improves the accuracy of phone duration prediction. The SVM-based fusion algorithm was found to outperform all other fusion techniques. Specifically, the fusion scheme based on the SVM regression algorithm outperformed the best individual predictor (SVM regression) by approximately 1.9% and 2.0% in terms of relative reduction of the mean absolute error and root mean square error on the KED TIMIT database, and by 2.6% and 1.8% on the WCL-1 database, respectively.

6. REFERENCES

Aha, D., Kibler, D., 1991. Instance-based learning algorithms. Machine Learning. 6, 37-66.

Allen, J., Hunnicutt, S., Klatt, D.H., 1987. From text to speech: The MITalk system. Cambridge University Press, Cambridge.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. on Automatic Control. 19(6), 716-723, doi:10.1109/TAC.1974.1100705.

Bartkova, K., Sorin, C., 1987. A model of segmental duration for speech synthesis in French. Speech Communication. 6, 245-260.

Bellegarda, J.R., Silverman, K.E.A., Lenzo, K., Anderson, V., 2001. Statistical prosody modelling: from corpus design to parameter estimation. IEEE Trans. on Speech and Audio Processing. 9(1), 52-66.

Breiman, L., 1996. Bagging predictors. Machine Learning. 24(2), 123-140.

Bourlard, H., Hermansky, H., Morgan, N., 1996. Towards increasing speech recognition error rates. Speech Communication. 18(3), 205-231.

Campbell, W.N., 1992. Syllable based segment duration, in: Bailly, G., Benoit, C., Sawallis, T.R. (Eds.), Talking Machines: Theories, Models and Designs, Elsevier, Amsterdam, Netherlands, pp. 211-224.

Carlson, R., Granstrom, B., 1986. A search for durational rules in real speech database. Phonetica. 43, 140-154.

Chen, S.H., Hwang, S.H., Wang, Y.R., 1998. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. IEEE Trans. on Speech and Audio Processing. 6(3), 226-239.

Chen, S.H., Lai, W.H., Wang, Y.R., 2003. A new duration modeling approach for Mandarin speech. IEEE Trans. on Speech and Audio Processing. 11(4), 308-320.

Crystal, T.H., House, A.S., 1988. Segmental durations in connected-speech signals: Current results. Journal of the Acoustical Society of America. 83(4), 1553-1573.

CSTR, 2001. CSTR US KED TIMIT. University of Edinburgh, http://www.festvox.org/dbs/dbs_kdt.html.

Dutoit, T., 1997. An Introduction to Text-To-Speech Synthesis, Kluwer Academic Publishers, Dordrecht.

Edwards, J., Beckman, M.E., 1988. Articulatory timing and the prosodic interpretation of syllable duration. Phonetica. 45, 156-174.

Epitropakis, G., Tambakas, D., Fakotakis, N., Kokkinakis, G., 1993. Duration modelling for the Greek language. In Proc. of EUROSPEECH-1993, pp. 1995-1998.

Ferrer, L., Bratt, H., Gadde, V.R.R., Kajarekar, S.S., Shriberg, E., Sonmez, K., Stolcke, A., Venkataraman, A., 2003. Modeling duration patterns for speaker recognition. In Proc. of EUROSPEECH-2003, pp. 2017-2020.

Freedman, D., Pisani, R., Purves, R., 2007. Statistics, fourth ed. W.W. Norton & Company, New York London.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 29(5), 1189-1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Statist. Data Anal. 38(4), 367-378.

Furui, S., 2000. Digital Speech Processing, Synthesis, and Recognition, 2nd ed. Marcel Dekker, New York.

Goubanova, O., King, S., 2008. Bayesian networks for phone duration prediction. Speech Communication. 50(4), 301-311.

Goubanova, O., Taylor, P., 2000. Using Bayesian belief networks for model duration in text-to-speech systems. In Proc. of ICSLP-2000.

Huang, X., Acero, A., Hon H.W., 2001. Spoken Language Processing: a guide to theory, algorithm, and system development. Prentice Hall.

Iwahashi, N., Sagisaka, Y., 2000. Statistical modeling of speech segment duration by constrained tree regression. IEICE Trans. Inform. Systems. E83-D(7), 1550-1559.

Jennequin, N., Gauvain, J.L., 2007. Modeling duration via lattice rescoring. In Proc. of ICASSP-2007, Honolulu, Hawaii, 641-644.

Kaariainen, M., Malinen, T., 2004. Selective Rademacher penalization and reduced error pruning of decision trees. Journal of Machine Learning Research. 5, 1107-1126.

Klatt, D.H., 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustic Society of America. 59, 1209-1221.

Klatt, D.H., 1987. Review of text-to-speech conversion for English. Journal of the Acoustical Society of America. 82(3), 737-793.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence. 97(1-2), 273-324.

Kohler, K.J., 1988. Zeistrukturierung in der Sprachsynthese. ITG-Tagung Digitalc Sprachverarbeitung, vol.6, pp.165-170.

Kominek, J., Black, A.W., 2003. CMU ARCTIC databases for speech synthesis. CMU-LTI-03-177, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Kominek, J., Black, A.W., 2004. A family-of-models approach to HMM-based segmentation for unit selection speech synthesis. In Proc. of Interspeech-2004, pp. 1385-1388.

Lazaridis, A., Zervas, P., Kokkinakis, G., 2007. Segmental duration modeling for Greek speech synthesis. In Proc. of IEEE ICTAI-2007, vol. 2, pp. 518-521.

Lee, S., Oh, Y., 1999. CART-based modelling of Korean segmental duration. In Proc. of Oriental COCOSDA'99, pp. 109-112.

Levinson, S., 1986. Continuously variable duration hidden Markov models for speech analysis. In Proc. of ICASSP-86, pp.1241-1244.

Mitchell, C., Harper, M., Jamieson, L., Helzermam, R., 1995. A parallel implementation of a hidden Markov model with duration modeling for speech recognition. Digital Signal Processing. 5, 43-57.

Monkowski, M.D., Picheny, M.A., Rao, P.S., 1995. Context dependent phonetic duration models for decoding conversational speech. In Proc. of ICASSP-95, vol. 1, pp. 528-531.

Olive, J.P., Liberman, M.Y., 1985. Text to speech - an overview. Journal of the Acoustical Society of America, 78 (Suppl. 1):S6.

Park, J., Sandberg, I.W., 1993. Approximation and radial-basis-function networks. Neural Computation. The MIT Press. 5, 305-316.

Platt, J., 1999. Fast training of support vector machines using sequential minimal optimization, in: Scholkopf, B., Burges, C., Smola, A. (Eds.), Advances in kernel methods: Support vector learning, Cambridge, MA: MIT Press, pp. 185-208.

Pols, L.C.W., Wang, X., ten Bosch, L.F.M., 1996. Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR. Speech Communication. 19, 161-176.

Quinlan, R.J., 1992. Learning with continuous classes. In Proc. of 5th Australian Joint Conference on Artificial Intelligence, Singapore, pp. 343-348.

Rao, K.S., Yegnanarayana, B., 2005. Modeling syllable duration in Indian languages using support vector machines. In Proc. of Intelligent Sensing and Information Processing, pp. 258-263.

Rao K.S., Yegnanarayana, B., 2007. Modeling durations of syllables using neural networks. Computer Speech & Language. 21(2), 282-295.

Riley, M., 1992. Tree-based modelling for speech synthesis, in: Bailly, G., Benoit, C., Sawallis, T.R. (Eds.), Talking Machines: Theories, Models and Designs, Elsevier, Amsterdam, Netherlands, pp. 265-273.

Scholkopf, B., Smola, A.J., 2002. Learning with Kernels, MIT Press.

Shih, C., Ao, B., 1997. Duration study for the Bell Laboratories Mandarin text-to-speech system, In: van Santen, J., Sproat, R., Olive, J., Hirschberg, J. (Eds.), Progress in Speech Synthesis, Springer, New York, pp. 383-399.

Simoes, A.R.M., 1990. Predicting sound segment duration in connected speech: An acoustical study of Brazilian Portugese. In Proc. of Workshop on Speech Synthesis, pp.173-176.

Smola, A.J., Scholkopf, B., 1998. A tutorial on support vector regression. Royal Holloway College, London, U.K., NeuroCOLT Tech. Rep. TR 1998-030.

Takeda, K., Sagisaka, Y., Kuwabara, H., 1989. On sentence-level factors governing segmental duration in Japanese. Journal of Acoustic Society of America. 86(6), 2081-2087.

van Santen, J.P.H., 1992. Contextual effects on vowel durations. Speech Communication. 11(6), 513-546.

van Santen, J.P.H., 1994. Assignment of segmental duration in text-to-speech synthesis. Computer Speech & Language. 8(2), 95-128.

van Santen, J., Olive, J., 1990. The analysis of contextual effects on segmental duration. Computer, Speech & Language. 4(4), 359-390.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, N.Y..

Vapnik, V., 1998. Statistical Learning Theory. New York: Wiley.

Vilalta, R., Drissi, Y., 2002. A perspective view and survey of meta-learning. Artificial Intelligence Review. 18(2), 77-95.

Wang, Y., Witten, I.H., 1997. Induction of model trees for predicting continuous classes. In Poster papers of the 9th European Conference on Machine Learning.

Wang, L., Zhao, Y., Chu, M., Zhou, J., Cao, Z., 2004. Refining segmental boundaries for TTS database using fine contextual-dependent boundary models. In Proc. of ICASSP-2004, vol. 1, pp. 641-644.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics, 1, 80-83.

Witten, H.I., Frank, E., 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kauffman Publishing.

Yamagishi, J., Kawai, H., Kobayashi, T., 2008. Phone duration modeling using gradient tree boosting. Speech Communication. 50(5), 405-415.

Yiourgalis, N., Kokkinakis, G., 1996. A TtS system for the Greek language based on concatenation of formant coded segments. Speech Communication. 19(1), 21-39.

Zervas, P., Fakotakis, N., Kokkinakis, G., 2008. Development and evaluation of a prosodic database for Greek speech synthesis and research. Journal of Quantitative Linguistics. 15(2), 154-184.

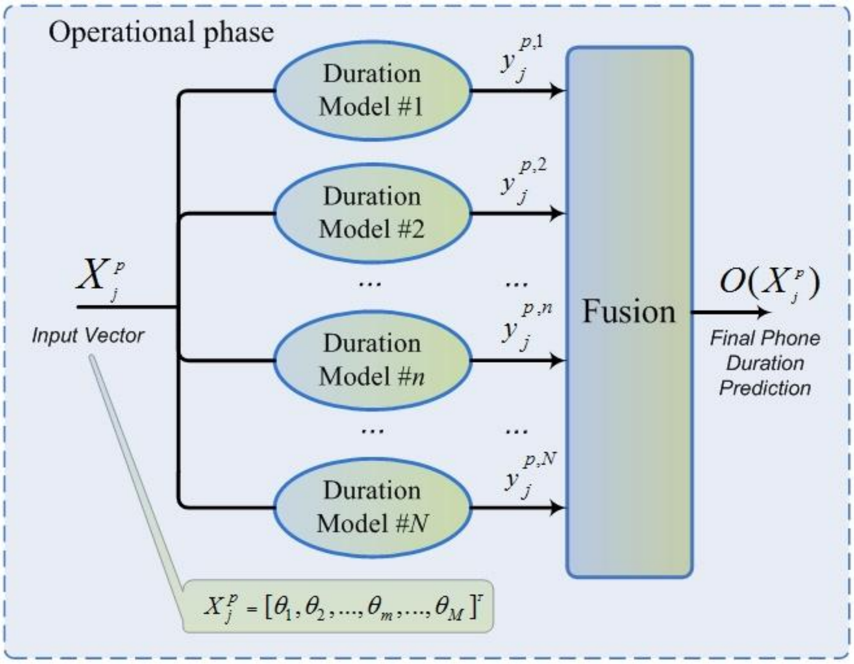

Fig. 1. Block diagram of the proposed fusion scheme, which exploits multiple dissimilar phone duration predictors, operating on a common input.

Table 1. Mean Absolute Error (MAE), standard deviation of absolute error (STD of AE) and Root Mean Square Error (RMSE) (in milliseconds) for the eight individual phone duration prediction algorithms on: (*a*) the KED TIMIT database, and (*b*) the WCL-1 database.

(*a*) results on the KED TIMIT database

Individual models (KED TIMIT database)	***MAE*** **(*ms*)**	***STD of AE*** **(*ms*)**	***RMSE*** **(*ms*)**
SMOreg	**14.95**	**14.11**	**20.56**
Add. Reg. m5pR (Yamagishi et al., 2008)	15.82	14.34	21.35
Add. Reg. REPTrees	16.29	15.06	22.19
Bagging m5pR (Lee and Oh, 1999)	16.51	14.76	22.14
m5p (Iwahashi and Sagisaka, 2000)	16.62	14.77	22.23
Bagging REPTrees	16.69	15.89	23.04
m5pR (Riley, 1992)	16.93	15.16	22.72
Linear Regression (Takeda et al., 1989)	17.15	15.16	22.89

(*b*) results on the WCL-1 database

Individual models (WCL-1 database)	***MAE*** **(*ms*)**	***STD of AE*** **(*ms*)**	***RMSE*** **(*ms*)**
SMOreg	**16.78**	**18.81**	**25.21**
Linear Regression (Takeda et al., 1989)	18.00	19.02	26.19
Add. Reg. REPTrees	18.08	19.97	26.94
Add. Reg. m5pR (Yamagishi et al., 2008)	18.13	19.16	26.38
Bagging m5pR (Lee and Oh, 1999)	18.14	19.63	26.72
m5p (Iwahashi and Sagisaka, 2000)	18.31	20.08	27.17
Bagging REPTrees	18.93	20.32	27.77
m5pR (Riley, 1992)	19.07	20.10	27.71

Table 2. Root Mean Square Error (in milliseconds) per phonetic category for the eight individual phone duration prediction algorithms on: (*a*) the KED TIMIT database, and (*b*) the WCL-1 database.

(*a*) results on the KED TIMIT database

KED TIMIT database	***LR***	***m5p***	***m5pR***	***Additive Regression***		***Bagging***		***SMOreg***
				m5pR	***REPTrees***	***m5pR***	***REPTrees***	
Vowel	24.56	24.18	25.46	23.67	24.87	24.78	26.34	**22.72**
Consonant	21.72	20.86	20.74	19.69	20.24	20.24	20.60	**19.02**
Phonetic category	***LR***	***m5p***	***m5pR***	***Additive Regression***		***Bagging***		***SMOreg***
				m5pR	***REPTrees***	***m5pR***	***REPTrees***	
Vowel	24.56	24.18	25.46	23.67	24.87	24.78	26.34	**22.72**
Affricate	22.44	24.41	23.48	22.86	**21.72**	22.96	23.34	21.88
Approximant	22.23	22.44	23.09	21.77	22.56	22.59	24.07	**20.42**
Fricative	22.51	21.67	21.10	20.19	20.69	20.63	20.96	**19.63**
Lateral	21.16	20.98	21.18	20.29	21.16	20.52	21.89	**19.77**
Nasal	18.59	17.88	17.80	17.11	16.94	17.28	17.57	**16.53**
Stop	23.39	22.07	21.62	20.26	20.97	21.04	20.80	**19.61**

(*b*) results on the WCL-1 database

WCL-1 database	***LR***	***m5p***	***m5pR***	***Additive Regression***		***Bagging***		***SMOreg***
				m5pR	***REPTrees***	***m5pR***	***REPTrees***	
Vowel	24.22	24.68	26.04	24.51	25.18	24.91	26.62	**23.12**
Consonant	27.86	29.25	29.13	27.97	28.44	28.27	28.77	**26.57**
Phonetic category	***LR***	***m5p***	***m5pR***	***Additive Regression***		***Bagging***		***SMOreg***
				m5pR	***REPTrees***	***m5pR***	***REPTrees***	
Vowel	24.22	24.68	26.04	24.51	25.18	24.91	26.62	**23.12**
Fricative	25.67	27.04	26.93	25.79	26.23	25.57	26.45	**23.95**
Liquid	19.46	19.19	19.55	18.84	17.83	18.47	18.02	**16.38**
Nasal	22.44	22.94	23.11	22.27	22.15	22.18	22.27	**20.62**
Stop	34.09	36.22	35.88	34.46	35.46	35.35	35.98	**33.53**

Table 3 (*a*). Root Mean Square Error (in milliseconds) per phone for the eight individual phone duration prediction algorithms on the KED TIMIT database

KED TIMIT database	*LR*	*m5p*	*m5pR*	*Additive Regression*		*Bagging*		*SMOreg*
				m5pR	*REPTrees*	*m5pR*	*REPTrees*	
aa	27.81	25.57	28.01	**24.57**	27.27	26.71	29.22	25.64
ae	31.40	30.97	31.67	29.75	31.19	31.64	33.13	**29.23**
ah	19.67	22.34	22.27	20.31	21.50	20.88	22.27	**19.38**
ao	32.79	**29.21**	32.95	30.54	32.11	32.66	33.29	29.65
aw	33.46	**32.89**	37.35	34.55	38.49	37.25	40.02	33.07
ax	16.10	15.66	16.06	15.16	15.56	15.54	15.93	**14.80**
ay	37.12	**32.78**	38.37	34.43	34.04	36.64	37.23	34.51
b	23.89	22.36	23.42	22.24	23.03	22.52	**21.19**	21.33
ch	**19.69**	23.34	21.17	20.48	20.43	19.89	22.36	20.57
d	20.77	19.36	19.66	19.12	20.05	19.32	20.54	**18.26**
dh	17.56	16.03	15.72	15.19	**14.57**	15.16	15.30	15.14
dx	11.08	10.38	11.30	9.99	**8.78**	9.54	8.86	9.63
eh	20.94	20.39	22.50	21.41	21.44	21.32	22.61	**19.05**
el	21.39	27.24	21.52	20.79	**18.98**	19.97	21.05	22.32
em	13.61	15.31	10.51	10.44	**10.13**	10.28	13.99	11.58
en	22.26	24.60	25.01	23.18	**20.67**	22.44	21.80	21.01
er	28.41	29.28	28.73	27.09	27.77	28.15	29.87	**25.29**
ey	27.76	26.99	29.43	28.12	29.90	28.72	31.36	**26.73**
f	22.84	23.90	21.52	20.09	21.05	21.08	22.43	**18.91**
g	18.23	17.14	18.73	17.04	17.65	17.88	17.62	**16.22**
hh	19.13	18.79	18.82	18.52	18.73	18.28	18.73	**17.54**
ih	19.38	19.76	20.16	19.09	19.81	19.82	20.86	**17.53**
iy	23.04	23.06	23.87	22.05	24.93	23.27	25.39	**20.99**
jh	24.36	25.22	25.14	24.56	**22.68**	25.08	24.07	22.85
k	22.18	21.82	20.62	18.65	18.63	19.94	18.93	**17.64**
l	21.13	20.18	21.14	20.24	21.39	20.58	21.98	**19.47**
m	16.07	15.32	16.20	15.45	16.19	15.81	17.04	**14.38**
n	18.69	17.65	17.29	16.70	**16.18**	16.80	16.32	16.19
ng	22.38	20.86	20.13	**19.90**	20.88	20.61	22.41	20.91
ow	28.12	28.98	28.93	27.20	28.85	27.73	30.68	**25.54**
oy	**25.45**	30.16	34.58	28.81	30.61	33.13	34.72	31.19
p	25.06	24.90	22.50	21.05	21.32	21.94	21.25	**20.45**
r	19.20	18.84	20.18	19.28	20.11	19.92	21.18	**18.25**
s	26.37	24.47	24.31	23.46	24.45	24.36	24.54	**23.21**
sh	19.71	21.72	19.28	18.30	20.53	18.49	20.27	**16.41**
t	28.18	25.60	25.06	23.64	25.14	24.72	24.93	**23.37**
th	24.09	26.39	29.14	25.58	**21.31**	25.21	22.59	22.05
uh	20.64	20.61	23.10	20.45	25.35	22.68	26.16	**19.88**
uw	27.65	27.73	29.05	28.00	30.35	29.40	33.64	**24.97**
v	17.26	17.31	16.72	16.93	17.15	16.66	17.34	**16.26**
w	20.28	20.09	22.35	19.81	20.93	20.89	22.59	**19.12**
y	18.36	19.08	18.85	18.80	19.42	19.22	20.56	**16.34**
z	22.38	20.42	19.94	19.07	19.37	19.24	19.10	**18.99**
zh	25.60	28.40	25.25	**22.62**	26.38	23.95	27.28	24.66

Table 3 (*b*). Root Mean Square Error (in milliseconds) per phone for the eight individual phone duration prediction algorithms on the WCL-1 database

WCL-1 database	*LR*	*m5p*	*m5pR*	*Additive Regression*		*Bagging*		*SMOreg*
				m5pR	*REPTrees*	*m5pR*	*REPTrees*	
K	45.75	44.50	45.47	43.94	46.86	45.73	45.82	**43.28**
i	24.17	24.30	25.54	24.27	24.68	24.68	26.98	**23.09**
t	34.70	36.84	34.98	34.31	36.09	34.96	36.04	**34.07**
a	24.25	25.85	25.76	24.07	24.57	24.83	26.07	**22.71**
ks	**22.50**	24.00	42.80	39.97	26.32	42.34	27.10	23.16
e	25.05	25.11	26.69	25.62	26.79	25.71	26.48	**24.05**
m	22.90	22.82	23.74	22.56	23.46	22.67	23.96	22.27
s	26.93	27.65	27.28	26.10	24.75	25.47	25.11	**23.49**
o	23.70	23.81	25.99	24.18	25.08	24.36	25.85	**22.72**
u	23.24	**22.63**	27.51	24.78	25.11	25.37	29.45	23.12
l	19.95	19.34	20.63	19.85	20.17	19.65	20.96	**18.31**
N	26.99	33.30	36.39	33.22	24.13	34.11	24.37	**21.26**
p	29.65	32.71	31.57	**28.65**	30.51	29.80	30.04	28.24
n	21.76	22.14	21.58	21.21	21.07	20.96	20.88	**19.38**
D	**22.64**	23.08	25.08	24.24	24.39	24.00	26.25	**22.64**
k	**42.27**	46.15	44.65	43.31	44.58	43.90	47.61	43.61
r	18.64	17.53	17.30	17.06	14.94	16.41	14.53	**13.85**
G	30.72	37.75	37.14	31.68	31.56	33.82	33.99	**29.89**
d	**19.33**	20.40	23.54	21.01	21.44	21.39	24.61	20.10
Q	**23.08**	25.82	25.22	23.49	24.99	23.82	26.83	23.85
z	23.05	22.38	23.31	22.64	23.13	22.98	24.68	21.58
y	20.77	20.35	22.98	21.03	21.35	21.38	21.64	**19.68**
Y	**26.68**	28.56	29.65	28.08	27.41	28.37	28.39	26.82
X	22.75	24.38	26.33	24.45	23.44	25.03	25.95	**21.44**
v	**23.87**	24.11	26.09	25.80	34.70	25.56	27.08	24.86
b	21.05	24.66	24.41	21.84	22.05	22.33	22.53	**20.20**
f	30.13	34.11	30.41	29.95	33.13	**29.50**	31.94	29.56
w	**20.83**	25.71	40.93	42.47	25.92	42.98	29.62	23.66
x	**20.33**	24.82	26.45	23.35	21.87	24.98	25.06	21.58
g	34.43	38.79	34.94	35.14	40.05	34.30	37.80	**33.85**
c	24.16	28.40	25.43	22.48	**20.29**	26.62	23.62	20.85
L	**24.98**	32.86	32.93	29.20	28.64	29.98	29.43	26.64
h	24.73	25.91	26.39	24.69	24.87	23.78	25.88	**23.50**
j	25.65	26.18	23.13	21.75	**20.52**	24.39	23.48	21.50

Table 4. Mean Absolute Error (MAE), standard deviation of absolute error (STD of AE) and Root Mean Square Error (RMSE) (in milliseconds) for the fusion scheme, implemented with the average linear combination and the best-case selection fusion algorithms on: (*a*) the KED TIMIT database, and (*b*) the WCL-1 database.

(*a*) results on the KED TIMIT database

Fusion algorithms on the KED TIMIT database	***MAE* (*ms*)**	***STD of AE* (*ms*)**	***RMSE* (*ms*)**
Overall (*average* linear combination)	15.32	14.01	20.76
Vowel/consonant (best-case selection)	14.95	14.11	20.56
Phonetic category (best-case selection)	14.94	14.11	20.54
Phone (best-case selection)	14.92	13.87	20.37
No fusion – best individual model, SMOreg	14.95	14.11	20.56

(*b*) results on the WCL-1 database

Fusion algorithms on the WCL-1 database	***MAE* (*ms*)**	***STD of AE* (*ms*)**	***RMSE* (*ms*)**
Overall (*average* linear combination)	16.91	18.72	25.29
Vowel/consonant (best-case selection)	16.78	18.81	25.21
Phonetic category (best-case selection)	16.78	18.81	25.21
Phone (best-case selection)	16.68	18.65	25.08
No fusion – best individual model, SMOreg	16.78	18.81	25.21

Table 5. Mean Absolute Error (MAE), standard deviation of absolute error (STD of AE) and Root Mean Square Error (RMSE) (in milliseconds) for the various fusion techniques on: (*a*) the KED TIMIT database, and (*b*) the WCL-1 database.

(*a*) results on the KED TIMIT database

KED TIMIT database	***MAE (ms)***	***STD of AE (ms)***	***RMSE (ms)***
SMOreg	**14.66**	**13.82**	**20.14**
IBK	15.19	14.69	21.02
Linear Regression	15.49	14.45	21.18
RBFNN	15.53	14.49	21.24
m5p	15.56	14.60	21.34
Add. Regr. m5pR	15.72	14.94	21.69
Add. Regr. REPTrees	15.79	14.94	21.74
Bagging m5pR	15.81	15.09	21.86
Bagging REPTrees	15.88	15.15	21.95
m5pR	15.97	15.28	22.10
No fusion – best individual model, SMOreg	14.95	14.11	20.56

(*b*) results on the WCL-1 database

WCL-1 database	***MAE (ms)***	***STD of AE (ms)***	***RMSE (ms)***
SMOreg	**16.35**	**18.59**	**24.76**
IBK	16.98	18.85	25.47
RBFNN	17.34	19.51	26.10
Add. Regr. m5pR	17.69	19.84	26.58
Bagging m5pR	17.72	19.84	26.60
m5p	17.84	20.51	27.18
m5pR	17.91	20.00	26.85
Bagging REPTrees	17.99	20.45	27.23
Add. Regr. REPTrees	18.00	20.56	27.32
Linear Regression	18.32	20.19	27.26
No fusion – best individual model, SMOreg	16.78	18.81	25.21