I.J. Intelligent Systems and Applications, **2012, 8, 1-9**
Published Online July 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2012.08.01

Phone Duration Modeling of Affective Speech Using Support Vector Regression

Alexandros Lazaridis[1,2], Iosif Mporas[1,3], Todor Ganchev[1]

[1]Artificial Intelligence Group, Wire Communications Laboratory, Dept. of Electrical and Computer Engineering, University of Patras, Rion-Patras 26500, Greece

[2]Dept. of Engineering Informatics & Telecommunications, University of Western Macedonia, Kozani, 50100, Greece

[3]Dept. of Informatics and Mass Media, Technological Educational Institute of Patras, Greece

Email:{alaza, imporas, tganchev}@upatras.gr

Abstract—In speech synthesis accurate modeling of prosody is important for producing high quality synthetic speech. One of the main aspects of prosody is phone duration. Robust phone duration modeling is a prerequisite for synthesizing emotional speech with natural sounding. In this work ten phone duration models are evaluated. These models belong to well known and widely used categories of algorithms, such as the decision trees, linear regression, lazy-learning algorithms and meta-learning algorithms. Furthermore, we investigate the effectiveness of Support Vector Regression (SVR) in phone duration modeling in the context of emotional speech. The evaluation of the eleven models is performed on a Modern Greek emotional speech database which consists of four categories of emotional speech (anger, fear, joy, sadness) plus neutral speech. The experimental results demonstrated that the SVR-based modeling outperforms the other ten models across all the four emotion categories. Specifically, the SVR model achieved an average relative reduction of 8% in terms of root mean square error (RMSE) throughout all emotional categories.

Index Terms— Phone Duration Modeling, Statistical Modeling, Support Vector Regression, Emotional Speech, Text-to-speech Synthesis

I. Introduction

Over the past decades, a great variety of techniques for speech synthesis have been developed. Despite the differences in these techniques, they all share one common aim, the improvement of the quality of the synthetic speech. There are two main criteria for measuring the quality of synthetic speech, namely the intelligibility and the naturalness of speech. The intelligibility of the synthetic speech measures the level of difficulty of the listener to understand the semantic contents of the speech [1]. The naturalness of the synthetic speech measures the resemblance between the synthetic speech and the human speech [1]. One of the main aspects for improving the naturalness and intelligibility of synthetic speech is the robust modeling of the prosody. Prosody is shaped by the relative level of the fundamental frequency, the intensity and the duration of the pronounced phones and refers to the introduction of functions and aspects of speech such as emphasis, intent, attitude or emotional state that cannot be encoded by grammar [1,2]. Therefore the accurate modeling of these aspects is mandatory for improving the prosody and consequently the quality of the produced synthetic speech.

In particular, the accurate modeling of phone duration is essential in speech synthesis, as it affects the structure of the utterance and contributes for improving the quality of synthetic speech. The accurate modeling of phone duration can be achieved through the proper modeling of the variables and factors which affect it. In the literature a great variety of factors affecting the duration of phones have been studied and various methods of phone duration modeling have been presented [3-8]. These factors and consequently the variables which are used in phone duration modeling to represent them, mainly belong to categories of speech representation such as the phonetic, the phonological and the morphosyntactic one.

To this end, two categories of methods for phone duration modeling have been developed: (i) the rule-based [9] and (ii) the data-driven methods [2, 8, 10-12]. In the first category (rule-based) the models are based on the use of manually produced rules. In order to produce these rules, experimental studies on large databases of utterances are mandatory. Consequently, the experience and knowledge of expert linguists is binding. The most well known method of rule-based models is the one introduced by Klatt in [9]. Based on this method, other similar models were developed in other language apart from English, such as in Swedish by Carlson and Granstrom [13], in French by Bartkova and Sorin [14] and in Greek by Epitropakis et al. in [15]. The main drawback in these methods is that the large number of the factors, which affect and determine the duration of phones, makes extremely difficult for someone to achieve the proper combination and manual tuning of them building robust phone duration models

[16]. Consequently, long-term devotion to this task becomes obligatory, and sometimes not even adequate, in order to collect all the appropriate rules [17]. These aspects of rule-based modeling restricted their application to controlled and constrained experiments involving a limited number of contextual factors.

On the other hand, data-driven methods, which were developed after the emergence of large databases [18], are based on statistical methods and Artificial Neural Network (ANN) techniques. These approaches exploit large databases in order to automatically extract the phonetic rules and consequently produce phone duration models. In this way, the problem of manual rules extraction mentioned earlier, is overcome, leading to a more efficient use of the expert linguists' efforts and experience. Some of the most widely used statistical approaches, which have been introduced in data-driven phone duration modeling over the last decades, are the Linear Regression (LR)-based technique [19], sums-of-products (SOP) [8] and decisions tree-based models [2]. Moreover, Artificial Neural Networks (ANN) techniques [10], Bayesian models [11] and instance-based algorithms [12] have also been introduced on this task.

Over the last years, the interest in emotional/affective speech synthesis is increasing continuously. Emotional speech synthesis mainly has been following the developments in the field of speech synthesis of neutral speech. Murray and Arnott in [20] and Burkhardt and Sendlmeier in [21] developed emotional speech synthesizers based on formant synthesis techniques, diphone concatenation approach was used in [22,23], while unit selection corpus based methods were implemented in [24,25]. In all these attempts prosody models are implemented in TTS systems in order to synthesize certain categories of emotional speech or produce more expressive speech [26-28]. Despite the progress in the recent years, we deem that further research investigation on the aspects related to phone duration modeling, and particularly in the context of emotional speech, has the potential to improve the quality of synthetic emotional speech.

In the present work, we offer a comparative evaluation of ten phone duration modeling techniques in the context of emotional speech. These algorithms have been used successfully on the task of phone and syllable duration modeling and belong to four categories of algorithms, (i) decision trees (DT) [29-31], (ii) lazy-learning algorithms [32,33], (iii) meta-learning algorithms [34,35] and (iv) linear regression (LR) [36]. Furthermore, we introduce, the support vector regression (SVR)–based phone duration modeling, in the context of emotional speech, and compare with the abovementioned traditional methods.

The remaining of this article is organized as follows. In Section 2, we present the support vector regression algorithm. In Section 3, we outline the emotional speech database and the feature set used in the present study. Moreover, we overview ten traditional phone duration modeling techniques, which are evaluated in this work, and briefly describe the performance evaluation metrics used in the evaluation of the phone duration models. The experimental results concerning the phone duration models are discussed in Section 4. In Section 5, we conclude this paper with a brief summary of the work.

II. Support Vector Regression (SVR)

The wide-spread use of SVR in statistical learning methods is mainly due to its good generalization performance, the absence of local minima and the sparse representation of solution [37,38]. In contrast to other traditional methods, which implement the Empirical Risk Minimization (ERM) principle, the SVMs implement the Structural Risk Minimization (SRM) principle [37,38]. The SRM principle seeks to minimize an upper bound of the generalization error rather than minimize the training error, which results in better generalization performance in SVMs. Training an SVM, is equivalent to solving a linearly constrained Quadratic Programming (QP) problem resulting in a unique and global optimum. In addition, SVMs are characterized by the sparse representation of the solution requiring less storage space and time for actual prediction since only the support vectors, which are a subset of the training data, are memorized after the training procedure.

The basic principle of SVM is the mapping of the training data from the input space onto a higher dimensional feature space using a function Φ, constructing subsequently a separating hyperplane with maximum margin in the feature space. Consider a training set of data, $(x_1, y_1), (x_1, y_1), \ldots, (x_i, y_i), \ldots,$ (x_N, y_N) , where each $x_i \in X \subseteq \Re^n$, denotes the input space of the sample and has a corresponding target value $y_i \in Y \subseteq \Re$, for i = 1, …, N , and N is the total number of the training samples. The regression problem is based on the determination of a linear regression function $f(x)$ that can approximate future values accurately, defined as:

$$f(x) = w^T \Phi(x_i) + b \tag{1}$$

where $w \subset \Re^n, b \subset \Re$ and Φ maps the training data to a higher dimensional space. This leads to the following optimization problem:

$$\underset{w,\xi_i,\xi_i^*}{\arg\min} \left\{ \frac{1}{2} w^T w + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \right\} \tag{2}$$

subject to:

$$\left\{ \begin{array}{l} y_i - \left(w^T \phi(x_i) + b \right) \le \varepsilon + \xi_i^* \\ \left(w^T \phi(x_i) + b \right) - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, \iota \in [0, \mathrm{N}] \end{array} \right\}, \tag{3}$$

where $\varepsilon \geq 0$ is the maximum deviation allowed during training and $C > 0$ is the penalty parameter for exceeding allowed deviation. The, ξ_i, ξ_i^* are the slack variables for exceeding the target value more or less than ε, respectively. The first term of Eq. (2) controls the function capacity and the second term is the empirical error.

Utilizing the method of Lagrange multipliers, α_i, α_i^*, a different strategy for finding the maximum or minimum of the function subject to constraints is used, the dual optimization problem is now defined as:

$$\arg\min_{\varepsilon, a_i, a_i^*} \left\{ \begin{array}{l} \frac{1}{2}\sum_{i,j}\left(a_i - a_i^*\right)\left(a_j - a_j^*\right)\left\langle x_i \cdot x_j \right\rangle \\ \varepsilon\sum_i\left(a_i - a_i^*\right) - \sum_i y_i\left(a_i - a_i^*\right) \end{array} \right\} \quad (4)$$

subject to:

$$\left\{ \begin{array}{l} \sum_i\left(a_i - a_i^*\right) = 0 \\ \left(a_i, a_i^*\right) \in \left[0, C\right] \end{array} \right\}. \quad (5)$$

The main advantage of the dual optimization problem is the easy expandability to a non-linear mapping function. In this case the input data are mapped to a higher dimensionality space by a non-linear function Φ. Consequently the term $\left\langle x_i \cdot x_j \right\rangle$ of Eq. 4 is replaced by the kernel function $k\left(x_i, x_j\right)$. Various kernel functions have been used successfully in non-linear SVR, such as the polynomial, radial basis function (RBF) or Gaussian functions. Unfortunately, the size of a SVR model depends on the number of the training samples. Thus, for large data sets, algorithm with low memory consumption is desired. The most widely used algorithm in this category is the Sequential Minimal Optimization (SMO) algorithm [29] applied in regression problems in [39]. A characteristic of the SMO algorithm is its ability to break the QP problems into a sequence of smaller possible QP problems, reducing the amount of memory required.

III. Experimental Setup

A Modern Greek database of emotional speech was used for the creation and the evaluation of the eleven phone duration models considered here. This database was purposely designed in support of research on speech synthesis. The database is linguistically and prosodically rich, containing speech of the four archetypal emotional categories [40]: anger, fear, joy and sadness, as well as neutral speech. In the following subsections the database along with the feature set used for building the phone duration models are introduced. Furthermore, the algorithms used for building the phone duration models and the performance metrics, which were used in order to evaluate and compare the models, are described.

A. The Modern Greek emotional speech database

The Modern Greek emotional speech database contains data for four archetypal emotion categories: anger, fear, joy and sadness along with the neutral category. The database was designed in such a way so as for each phone to have multiple instances in different positions (initial, medial, final) in various words in the database. This is a very important aspect of the database since the positional and contextual factors of a phone (place in syllable, word etc) play a very important role in the assessment of its duration [2,8,41,42]. The sentences and phrases in the database were extracted from passages, newspapers or designed by a linguist. The database consists of 62 utterances, which are pronounced several times with different emotional charge. The length of the utterances ranges from a single word, a phrase, short and long sentence or even a sequence of sentences of fluent speech, summing up to a total of 4150 words in 310 utterances throughout all the emotion categories. A phone inventory of 34 phones was used, with a total of 22045 instances consisting of 15667 voiced and 6378 unvoiced phone occurrences. Moreover, each vowel class included both stressed and unstressed cases of the corresponding vowel. In this point it should be pointed out that the context of all sentences is emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactic or semantic means.

A professional female actress, speaking Modern Greek, was employed for uttering all the sentences of the database. All the recordings of each specific emotional category were recorded in series, before proceeding with the other emotional categories. All recording sessions were held in the anechoic chamber of a professional studio so as to ensure the quality of the audio and noise-free conditions of the recordings. Speech was sampled at 44.1 kHz, and a resolution of 16 bit per speech sample. For the needs of our experiments we down-sampled all the recordings to sampling rate of 16 kHz.

B. Feature set

A number of features which have been reported successful on the task of phone duration modeling were used in our experiments [2,4-8,12,17,19,41,43-50].

Since the input to a speech synthesis system is only text, the linguistic features composing the feature set were extracted only from text. In more specific, for each utterance 33 features were extracted per phone instance. In addition, the syntagmatic neighbors of some of these features, defined on the level of the respective feature, i.e. phone-level, syllable-level, word-level, were extracted from the utterances. The features composing the feature set are presented in Table 1. After including the aforementioned features along with their syntagmatic neighbors information as reported above (one or two previous and next instances on the level of the respective feature, phone-level, syllable-level, word-level), the overall size of the feature set sums up to 93.

TABLE I. FEATURE SET USED FOR BUILDING THE PHONE DURATION MODELS

Feature set
eight phonetic features:
the phone class (consonants/non-consonants), along with the information of the neighboring (two previous, two next) instances.
the phone types (short vowels, long vowels, diphthongs, schwa, consonants), along with the information of the neighboring (two previous, two next) instances.
the vowel height (high, middle, low), along with the information of the neighboring (two previous, two next) instances.
the vowel frontness (front, central, back), along with the information of the neighboring (two previous, two next) instances.
the lip rounding (rounded/unrounded), along with the information of the neighboring (two previous, two next) instances.
the manner of production (plosive, fricative, affricate, liquids, nasal), along with the information of the neighboring (two previous, two next) instances.
the place of articulation (labial, labio-dental, dental, alveolar, palatal, velar, glottal), along with the information of the neighboring (two previous, two next) instances.
the consonant voicing, along with the information of the neighboring (two previous, two next) instances.
three segment-level features:
the phone name with the information of the neighboring instances (previous, next),
the position of the phone in the syllable,
the onset-rhyme type (onset: if the specific phone is before the vowel in the syllable, rhyme: if the specific phone is the vowel or it is after the vowel in the syllable),
thirteen syllable-level features:
the position type of the syllable (single, initial, middle or final) in the word along with the information of the neighboring instances (previous, next),
the number of all the syllables in the utterance,
the number of accented syllables and the number of stressed syllables since the last and to the next phrase break (i.e. the break index tier of ToBI (Silverman et al., 1992) with values: 0, 1, 2, 3, 4),
syllable's onset-coda size (the number of phones before and after the vowel of the syllable) along with the information of the previous and next instances,
the onset-coda type (if the consonant before and after the vowel in the syllable is voiced or unvoiced) along with the information of the previous and next instances,
the position of the syllable in the word,
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, adjective, etc),
the number of syllables in the word,
one phrase-level feature:
the syllable break (the phrase break after the syllable) along with the information of the neighboring (two previous, two next) instances. The syllable break feature is based on the break index tier (0, 1, 2, 3, 4) of ToBI (Silverman et al., 1992),
six accentual features:
the ToBI accents and boundary tones along 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),
the stressed-unstressed syllable (if the syllable is stressed or not),
the accented-unaccented syllable (if the syllable is accented or not) with the information of the neighboring (two previous, two next) instances.

C. Phone Duration Models

Along with the SVR which were introduced and described in the previous section, ten different phone duration models belonging to four categories of machine learning algorithms (decision trees, linear regression, lazy-learning and meta-learning algorithms) are built and evaluated in this work. In the following, we briefly outline these algorithms:

i. Three decision trees were used. Two of them, namely the M5p model tree [31] and the M5pR regression tree [30], are based on the M5' algorithm [31]. The third algorithm is a regression tree -- the Reduced Error Pruning trees (REPTrees) [29].

ii. In the lazy-learning category, two different algorithms were implemented: the Instance based learning IBK [32] and the Locally Weighted Learning (LWL) algorithm [33]. The IBK uses the k-nearest neighbors algorithm (k-NN). The IBK algorithm, in order to locate the instance that is closer to the training instance, searches among the k nearest neighbors of the instance. Evaluating this method with different number of neighbors resulted in the adaptation of 12 neighbors (k = 12), since it gave the best results. The LWL assigns weights using an instance-based method. In this case, the kernel function which is used to calculate weights for the data points was the tricube kernel function, while REPTrees were used as classifiers.

iii. Furthermore, two meta-learning algorithms were used, namely the Additive Regression (AR) [35] and the Bagging algorithm (BG) [34]. Both of these algorithms were implementing using two different regression trees (M5pR and REPTrees) as base classifiers. In the two cases of additive regression meta-learning algorithm 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 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. In the two cases of the bagging algorithm, 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.

iv. Moreover, in our experiments, the linear regression (LR) [36] algorithm was used. This algorithm is a classification and prediction algorithm that expresses the class variable as a linear combination of the features. The error estimation in LR algorithm is given by the Akaike Information Criterion (AIC) [51].

v. Finally, for the SVR model, the radial basis function (RBF) was chosen as kernel function and ε and C parameters, where $\varepsilon \geq 0$ and $C > 0$, were set equal to 10^{-2} and 1.0 respectively, 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 the randomly selected subset of the training dataset mentioned above.

D. Performance metrics

Two of the most commonly used figures of merit were used in order to evaluate the phone duration modes: (i) the mean absolute error (MAE) and (ii) the root mean squared error (RMSE) between the predicted and the actual duration of each phone [8,10,45,46]. The RMSE is considered to be a metric more sensitive to outliers (large errors), weighing the heavily, due to the squaring of values [36]. This sensitivity of the RMSE makes it a more illustrative measurement concerning the outliers, e.g. the gross errors, in comparison to the MAE. Furthermore, the Correlation Coefficient (CC) was calculated. The CC measures the statistical correlation between the actual and the predicted values of the phone duration.

Finally, an experimental protocol based on 10-fold cross-validation was applied in all the experiments in order to exploit in the best way the available data.

IV. Experimental Results

In Tables II, III and IV the performance evaluation results of the eleven phone duration models are presented. All values of the RMSE and MAE are in milliseconds. As can be seen throughout all the emotion categories, the phone duration models implemented with the support vector regression algorithm, SVR, outperformed all the other models. The second-best accuracy was observed for the M5p trees and the meta-learning Additive Regression and Bagging algorithms using M5pR regression trees as base classifiers (AR.M5pR and BG.M5pR). In details, in terms of RMSE, the SVR outperformed the respective second-best model, presenting a relative reduction of 9.2% compared to the M5p model for category Anger, 9% compared to AR.M5pR model in category Fear and 8% compared to M5p model in category Neutral. A smaller reduction was achieved in the other two emotion categories, presenting a relative reduction of 6.8% and 7.3% compared to the second-best model AR.M5pR in categories Joy and Sadness, respectively.

Table II. RMSE values in milliseconds of the eleven PDMs for the different categories of emotional speech

	Anger	Fear	Joy	Neutral	Sadness
SVR	**19.7**	**18.3**	**17.7**	**24.1**	**19.1**
AR.M5pR	22.1	20.1	19.0	26.3	20.6
AR.R.Tr.	23.8	21.3	20.8	26.7	22.1
BG.M5pR	23.3	20.9	20.4	26.7	21.4
BG.R.Tr.	28.2	22.5	22.8	27.6	24.3
IB12	24.7	21.8	22.2	27.5	20.6
LWL	28.6	24.4	23.4	28.9	25.7
LR	22.8	22.0	19.8	26.4	20.8
M5p	21.7	20.2	19.5	26.2	20.9
M5pR	24.1	21.6	21.6	27.2	22.1
R.Tr.	30.3	24.3	24.5	29.4	26.6

Table III. MAE values in milliseconds of the eleven PDMs for the different categories of emotional speech

	Anger	Fear	Joy	Neutral	Sadness
SVR	**14.4**	**13.6**	**13.2**	**15.4**	**14.6**
AR.M5pR	16.3	14.9	14.0	17.5	15.6
AR.R.Tr.	17.5	15.7	15.3	17.8	16.8
BG.M5pR	17.1	15.4	15.1	17.7	16.2
BG.R.Tr.	20.5	16.5	16.7	18.6	18.1
IB12	18.0	15.8	16.4	18.4	15.6
LWL	20.5	18.0	17.0	19.3	19.0
LR	17.1	16.0	14.9	17.7	16.1
M5p	16.1	15.0	14.8	17.1	16.0
M5pR	17.6	16.0	15.9	18.2	16.8
R.Tr.	22.2	18.2	17.9	20.1	20.0

Table IV. CC values of the eleven PDMs for the different categories of emotional speech

	Anger	Fear	Joy	Neutral	Sadness
SVR	**0.86**	**0.77**	**0.81**	**0.73**	**0.79**
AR.M5pR	0.83	0.72	0.78	0.66	0.75
AR.R.Tr.	0.79	0.67	0.73	0.65	0.70
BG.M5pR	0.81	0.70	0.75	0.66	0.73
BG.R.Tr.	0.70	0.62	0.66	0.62	0.63
IB12	0.78	0.66	0.69	0.63	0.75
LWL	0.70	0.55	0.65	0.59	0.59
LR	0.81	0.66	0.76	0.66	0.74
M5p	0.83	0.72	0.77	0.67	0.74
M5pR	0.79	0.66	0.70	0.63	0.70
R.Tr.	0.65	0.55	0.60	0.57	0.54

Regarding the MAE and CC, reduction was achieved in all emotion categories. In terms of MAE, the SVR outperformed the respective second-best model, presenting a relative reduction of 10.6%, when compared to the M5p model in category Anger, 8.2% compared to the AR.M5pR model in category Fear, and 9.9% compared to the M5p model in category Neutral. A slightly smaller reduction was achieved in the other two emotion categories, presenting a relative reduction of 5.7% and 6.4% compared to the second-best model AR.M5pR in categories Joy and Sadness, respectively. Finally concerning CC, the SVR model outperformed the respective second-best model, presenting a relative increase of 6.9% compared to the AR.M5pR model in category Fear and 9% compared to the M5p model in category Neutral. A smaller increase was observed in the other three emotion categories, presenting a relative increase of 3.6% compared to the M5p model in category Anger, 3.5% and 5.3% compared to the second-best model, AR.M5pR, in categories Joy and Sadness, respectively.

The overall second-best accuracy of phone duration modeling was observed for the M5p trees and the meta-learning, Additive Regression and Bagging, algorithms using M5pR regression trees as base classifiers (AR.M5pR and BG.M5pR). Furthermore, even though the simple LR model showed higher error rates in respect to the above-mentioned models, it still performed close to the M5pR regression trees. Concerning the two local learning algorithms, it should be pointed out that the models implemented with IB12 rather than the LWL managed to achieve a higher performance in all emotion categories. Finally, REPTrees (R.Tr.) demonstrated the lowest accuracy among all evaluated methods, both as a single model, and as a base classifier for the cases of AR and BG algorithms (AR.R.Tr., BG.R.Tr.).

As reported earlier, the SVR model outperformed all the other models throughout all the categories of emotional speech. This advantage of SVR over all the other algorithms evaluated in this work can be explained by the ability of SVMs to cope better with high-dimensional feature space [37,38]. Due to the curse of dimensionality, the other machine learning techniques are unable to build robust models from the available training data.

V. Conclusions

In this work, we investigated of the applicability of ten phone duration modeling algorithms (among which are model and regression trees, linear regression, lazy learning and meta-learning based methods) in the context of emotional speech. In addition, the support vector regression algorithm, which to the extent of our knowledge has not been used so far for phone duration modeling in the context of emotional speech was introduced and evaluated. All experiments were performed on a Greek database of emotional speech, which consists of five archetypal emotion categories: anger, fear, joy, neutral and sadness. The results showed that all the machine learning algorithms managed to build robust phone duration models; however, the support vector regression model presented by far the best accuracy. It achieved a relative reduction ranging from 6.8% to 9.2%, in terms of RMSE, over all the emotion categories compared to the second-best model.

References

[1] Dutoit T.. An Introduction to Text-To-Speech Synthesis [B]. Dordrecht: Kluwer Academic Publishers. 1997.

[2] Möbius B, Santen P H J. Modeling Segmental duration in German Text-to-Speech Synthesis [C]. Proceedings of ICSLP'96, Philadelphia, USA, 1996, 2395–2398.

[3] Barbosa P A, Bailly G.. Characterisation of rhythmic patterns for text-to-speech synthesis [J]. Speech Communication, 1994, 15: 127–137.

[4] Bell A, Jurafsky D, Fosler-Lussier E, Girand C, Gregory M, Gildea D. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation [J]. Journal of the Acoustical Society of America, 2003, 113(2): 1001–1024.

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

[6] Gregory M, Bell A, Jurafsky D, Raymond W. Frequency and predictability effects on the duration of content words in conversation [J]. Journal of the Acoustical Society of America, 2001, 110(5): 27–38.

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

[8] van Santen J P H. Contextual effects on vowel durations [J]. Speech Communication, 1992, 11: 513–546.

[9] Klatt D H. Synthesis by rule of segmental durations in English sentences [B]. In B. Lindlom, and S. Ohman (Eds.), Frontiers of Speech Communication Research. New York: Academic Press, 1979, 287–300.

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

[11] Chien J T, Huang C H. Bayesian Learning of Speech Duration Models [J]. IEEE Trans. on Speech and Audio Processing, 2003, 11(6): 558–567.

[12] Lazaridis A, Zervas P, Kokkinakis G. Segmental Duration Modeling for Greek Speech Synthesis [C]. In Proceedings of ICTAI'07, Patras, Greece, 2007, 518–521.

[13] Carlson R, Granstrom B. A search for durational rules in real speech database [J]. Phonetica, 1988, 43: 140-154.

[14] Bartkova K, Sorin C. A model of segmental duration for speech synthesis in French [J]. Speech Communication, 1987, 6: 245–260.

[15] Epitropakis G, Tambakas D, Fakotakis N, Kokkinakis G. Duration modelling for the Greek language [C]. In Proceedings of EUROSPEECH'93, Berlin, Germany, 1993, 1995–1998.

[16] Rao K S, Yegnanarayana B. Modeling durations of syllables using neural networks [J]. Computer Speech & Language, 2007, 21(2): 282–295.

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

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

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

[20] Murray I R, Arnott J L. Implementation and testing of a system for producing emotion-by-rule in synthetic speech [J]. Speech Communication, 1995, 16: 369–390.

[21] Burkhardt F,. Sendlmeier W F. Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis [C]. Proceedings of the ISCA Workshop on Speech & Emotion, Northern Ireland, 2000, 151–156.

[22] Heuft B, Portele T, Rauth M. Emotions in Time Domain Synthesis [C]. Proceedings of ICSLP'96, Philadelphia, USA, 1996, 1974–1977.

[23] Rank E, Pirker H. Generating Emotional Speech with a Concatenative Synthesizer [C]. Proceedings of ICSLP'98, Sydney, Australia, 1998, 671–674.

[24] Black A. Unit Selection and Emotional Speech [C]. Proceedings of EUROSPEECH’03, Geneva, Switzerland, 2003, 1649–1652.

[25] Iida A, Campbell N, Iga S, Higuchi F, Yasumura M. A Speech Synthesis System for Assisting Communication [C]. Proceedings of the ISCA Workshop on Speech & Emotion, Northern Ireland, 2000, 167–172.

[26] Inanoglu Z, Young S. Data-driven emotion conversion in spoken English [J]. Speech Communication, 2009, 51: 268–283.

[27] Jiang D N, Zhang W, Shen L, Cai L H. Prosody Analysis and Modeling for Emotional Speech Synthesis [C]. Proceedings of ICASSP’05, Philadelphia, USA, 2005, 281–284.

[28] Tesser F, Cosi P, Drioli C, Tisato G. Emotional Festival-Mbrola TTS Synthesis [C]. Proceedings of INTERSPEECH’05, Lisboa, Portugal, 2005, 505–508.

[29] Kääriäinen M, Malinen T. Selective Rademacher Penalization and Reduced Error Pruning of Decision Trees [J]. Journal of Machine Learning Research, 2004, 5: 1107–1126.

[30] Quinlan R J. Learning with continuous classes [C]. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Tasmania, 1992, 343–348.

[31] Wang Y, Witten I H. Induction of model trees for predicting continuous classes [C]. Proceedings of the 9th European Conference. on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague, Czech, 1997, 128–137.

[32] Aha D, Kibler D, Albert M. Instance-based learning algorithms [J]. Journal of Machine Learning, 1991, 6: 37–66.

[33] Atkeson C G, Moorey A W, Schaal S. Locally Weighted Learning [J]. Artificial Intelligence Review, 1996, 11: 11–73.

[34] Breiman L. Bagging Predictors [J]. Journal of Machine Learning, 1996, 24(2): 123–140.

[35] Friedman J H. Stochastic gradient boosting [J]. Computational Statistics and Data Analysis, 2002, 38(4): 367–378.

[36] Witten H I, Frank E. Data Mining: Practical Machine Learning Tools and Techniques [B], second ed. San Francisco: Morgan Kaufmann Publishing, 2005.

[37] Vapnik V. The Nature of Statistical Learning Theory [B]. Springer, New York, 1995.

[38] Vapnik V. Statistical Learning Theory [B]. Wiley, New York, 1998.

[39] Scholkopf B, Smola A J. Learning with Kernels [R]. MIT Press, Cambridge, 2002

[40] Oatley K, Johnson-Laird P. The communicative theory of emotions [B]. In J. Jenkins, K. Oatley, and N. Stein (Eds), Human Emotions: A Readr. Oxford: Blackwell, 1998, 84–87.

[41] Febrer A, Padrell J, Bonafonte A. Modeling Phone Duration: Application to Catalan TTS [C]. Workshop of Speech Synthesis, Australia, 1998, 43–46.

[42] Krishna N S, Talukdar P P, Bali K, Ramakrishnan A G. Duration Modeling for Hindi Text-to-Speech Synthesis System [C]. Proceedings of ICSLP’04, Jeju Island, Korea, 2004, 789–792.

[43] Chung H. Duration models and the perceptual evaluation of spoken Korean [C]. Proceedings of Speech Prosody, France, 2002, 219–222.

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

[45] Goubanova O, King S. Bayesian network for phone duration prediction [J]. Speech Communication, 2008, 50: 301–311.

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

[47] Krishna N S, Murthy H A. Duration modeling of Indian languages Hindi and Telugu [C]. Proceedings of the 5th ISCA Speech Synthesis Workshop, Pittsburgh, USA, 2004, 197–202.

[48] Lee S, Oh Y H. CART-based modelling of Korean segmental duration [C]. Proceedings of the Oriental COCOSDA’99, Taipei, Taiwan, 1999, 109–112.

[49] Teixeira, J P, Freitas D. Segmental durations predicted with a neural network [C]. Proceedings of EUROSPEECH’03, Geneva, Switzerland, September, 2003, 169–172.

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

[51] Akaike H. A new look at the statistical model identification [J]. IEEE Trans. on Automatic Control, 1974, 19: 716-723.

Alexandros Lazaridis was born in Thessaloniki, in 1981. He graduated in September of 2005 from the Department of Electrical & Computer Engineering at Aristotle University of Thessaloniki, in Greece. He received his PhD at the Department of Electrical and Computer Engineering at the University of Patras, in February of 2011. Currently he is post-doctoral researcher at the University of Patras and non-tenured Lecturer at the University of Western Macedonia and non-tenured Assistant Professor at the Technological Educational Institute of Serres. He is author and co-author in more than 20 papers. His fields of research include Speech Processing, Voice Conversion, Speech Synthesis and Speech Prosody.

Iosif Mporas was born in Athens, Greece, in 1981. He graduated in 2004 (Diploma) from the Department of Electrical and Computer Engineering of the University of Patras, Greece. He received his PhD degree in July 2009. Currently he is post-doctoral researcher at the University of Patras and non-tenured Assistant Professor at the Technological Educational Institute of Patras. He is author and co-author in more than 50 publications in scientific journals and international conferences. His research interests include speech and audio signal processing, pattern recognition, automatic speech recognition, automatic speech segmentation and spoken language/dialect identification.

Todor Ganchev received the Diploma Engineer degree in Electrical Engineering from the Technical University of Varna, Bulgaria, in 1993 and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Patras, Greece, in 2005. From February 1994 to August 2000, he consequently held Engineering, Research, and Teaching Staff positions at the Technical University of Varna. Since September 2000, he has been a Researcher at the Wire Communications Laboratory, University of Patras, where he currently holds a Senior Researcher position. His research interests are in the areas of pattern recognition, signal processing, and their applications.

How to cite this paper: Alexandros Lazaridis, Iosif Mporas, Todor Ganchev,"Phone Duration Modeling of Affective Speech Using Support Vector Regression", International Journal of Intelligent Systems and Applications(IJISA), vol.4, no.8, pp.1-9, 2012. DOI: 10.5815/ijisa.2012.08.01