IDIAP RESEARCH REPORT

SVR VS MLP FOR PHONE DURATION MODELLING IN HMM-BASED SPEECH SYNTHESIS

Alexandros Lazaridis[a] Pierre-Edouard Honnet[a]
Philip N. Garner

Idiap-RR-03-2014

MARCH 2014

[a]Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

SVR vs MLP for Phone Duration Modelling in HMM-based Speech Synthesis

Alexandros Lazaridis, Pierre-Edouard Honnet, Philip N. Garner

Idiap Research Institute, Martigny, Switzerland

{alaza, pierre-edouard.honnet, phil.garner}@idiap.ch

Abstract

In this paper we investigate external phone duration models (PDMs) for improving the quality of synthetic speech in hidden Markov model (HMM)-based speech synthesis. Support Vector Regression (SVR) and Multilayer Perceptron (MLP) were used for this task. SVR and MLP PDMs were compared with the explicit duration modelling of hidden semi-Markov models (HSMMs). Experiments done on an American English database showed the SVR outperforming the MLP and HSMM duration modelling on objective and subjective evaluation. In the objective test, SVR managed to outperform MLP and HSMM models achieving 15.3% and 25.09% relative improvement in terms of root mean square error (RMSE) respectively. Moreover, in the subjective evaluation test, on synthesized speech, the SVR model was preferred over the MLP and HSMM models, achieving a preference score of 35.93% and 56.30%, respectively.

Index Terms: phone duration modelling, Support Vector Regression, Multilayer Perceptron, HSMM explicit duration modelling, HMM-based speech synthesis

1. Introduction

Prosody plays a very important role in verbal communication. Changing prosody can completely change the meaning of the message which is conveyed through speech [1]. There are three main aspects of prosody: duration, pitch and intensity [2]. Duration is a prosodic factor controlling the speaking rate, the rhythm of the speech [3]. Controlling this factor gives the ability to the speaker to emphasize more on some parts of a sentence and less on others, helping the listener to perceive the proper message. In the same way, duration and prosody in general, are essential factors in the field of speech synthesis.

Statistical parametric speech synthesis techniques, and hidden Markov models (HMMs) in particular, provide a framework for the task of speech synthesis, achieving on one hand high quality synthetic speech and on the other hand giving a high degree of flexibility in modelling and transforming various aspects of the speech, such as speaker identity, age, gender, emotions and prosody [4, 5, 6, 7]. Over the last years, many improvements have been introduced in HMM-based speech synthesis, one of them being the use of hidden semi-Markov models (HSMMs) [8]. The advantage of HSMMs is the explicit modelling of state duration using Gaussian distributions instead of the implicit modelling of HMMs by the transition probabilities of the states. In the training phase of HSMM-based speech synthesis, a decision tree is built according to some phonetic and linguistic features and a set of fixed binary (yes/no) questions controlling and even sometimes limiting [9, 10] the structure of the tree. The minimum description length (MDL) is used as a splitting and stopping criterion [11]. In the synthesis phase, the decision tree is traversed for each target unit until reaching a leaf node. The mean value of this leaf node is used as the predicted duration for the target unit. The drawback of this approach is that these trees cannot represent properly all the target units in speech synthesis [12].

To improve synthetic speech and alleviate monotonous prosody and specifically monotonous durations, a lot of research has been done over the last years. Various approaches and techniques, such as modelling duration combining models of multiple levels, e.g. state and phone levels [13], state, phone and syllable levels [14], using full covariance Gaussian distribution [15] or implementing Gamma distribution instead of Gaussian [16], have been introduced for this task. Furthermore, a lot of focus has been given on using external duration models, forcing their predicted durations on HMM-based speech synthesis [17, 18]. A variety of machine learning algorithms have been used for state, phone or syllable duration modelling, such as decision trees [19, 20], Bayesian Networks [21], Linear Regression [22], Instance-based learning [22], Support Vector Regression [23, 24], Multilayer Perceptron [25, 26], or even fusion of these algorithms [27, 28], to improve the accuracy.

The motivation behind this work is to investigate how Support Vector Regression (SVR) and Multilayer Perceptron (MLP) algorithms, which have been used successfully in various tasks, could improve, as external phone duration models, the prediction accuracy in order to improve the quality of synthesized speech. To our best knowledge, these two algorithms have never been compared in the same experimental conditions. We believe that the limitations caused by the use of HSMMs for the duration modelling, e.g. the use of a specific set of questions for the decision tree, or the difficulty of the decision trees to model complex context dependencies [9], could be overcome with the use of the external duration models. Furthermore, we consider that the ability of SVR in coping well with high-dimensional space in respect to the training data will result in a more robust duration model in comparison to a model build using the MLP. An American English male database is used (CMU-ARCTIC-RMS) for these experiments [29].

The rest of the paper is organized as follows. The HSMM explicit duration model and the two external phone duration modelling approaches, MLP, SVR are presented in Section 2 and 3 respectively. The experimental setup and results are presented in Section 4. In Section 5, the conclusions are given.

2. HSMM-based Duration Modelling

In HMM-based speech synthesis duration modelling is done at the state level, through the state sequence modelling. Consequently, the phone duration modelling is performed through the state modelling. The reasoning behind this approach is the fact that the state sequence modelling is not only responsible for the duration of the phones, but also is the basic structural element of the HMMs for spectrum and f0 modelling and generation.

In HMM-based speech synthesis duration modelling is done implicitly through the transition probabilities of the HMM

states i.e. an exponential distribution, making this structure unsuitable for modelling properly the timing in synthetic speech. The advantage of HSMMs in respect to HMMs, is the explicit modelling of state duration using Gaussian distributions instead of the implicit modelling by the transition probabilities of the states. Although the Gaussian distribution is clearly wrong (it implies negative durations are possible), it is a suitable approximation to the true distribution.

In the training phase, using the state durations and the phonetic and prosodic context-dependent features of the training data, a decision tree is built. This decision tree is constructed based on some predetermined binary questions concerning the content of the features (e.g. is the current syllable accented, is the previous phone fricative, etc.). For controlling the growth of the tree and the splitting of the nodes, the minimum description length (MDL) criterion is used [11]. Finally, the leaf nodes of the tree correspond to different clusters of the training data, sharing the same distributions (i.e. mean and variances). In the synthesis phase, according to the unseen data, the tree is traversed from the root node until a leaf node is reached. In this way, the Gaussian distributions of the leaf nodes are used to determine the duration of the synthesized speech.

Using HMMs/HSMMs for modelling duration in speech synthesis leads to some drawbacks. First, it is inefficient to express complex context dependencies such as XOR, parity or multiplex problems by decision trees [9]. In order to be able to cope with such cases, decision trees must become very large. Furthermore, the mean values of the Gaussian distributions of the leaf nodes are inadequate, due to over-generalization, to deal properly, in respect to duration modelling, with all the cases of the unseen data during the synthesis phase.

For overcoming these problems and improving the duration modelling accuracy, two external models are implemented and evaluated in this work, using the SVR and MLP algorithms.

3. External Phone Duration Modelling

In this section the two external phone duration models, the MLP and SVR, are described. When an external phone duration model is used in HMM-based speech synthesis, the predicted duration of the phone during the synthesis phase, has to be forced upon the HMMs [30]. Consequently the HMMs for each phone, having the predefined phone duration, are used only to distribute the predicted phone duration to the states.

3.1. Multilayer Perceptron

The Multilayer Perceptron is a feed-forward neural network having one or more hidden layers between the input and output layers [31]. Having a feed-forward architecture means that the connections between all the units and layers follow only one direction, from input units to the output units. Apart from the input units, each unit is modelled using a non-linear activation function. Furthermore, each unit of a layer is connected with a specific weight to every unit of the next layer. Consequently, the input layer is connected to the output layer through a weighted linear combination of non-linear functions. In this way the input data are transformed into another space, where can be linearly separable. In our experiments the MLP was implemented using one hidden layer.

3.2. Support Vector Regression

A Support Vector Machine (SVM) constructs a hyperplane in a high-dimensional space, which can be used for classification (SVM) and regression (SVR) tasks [32]. The basic idea governing the SVR is the production of a model that can be expressed through support vectors which define the hyperplane. A linear regression function is used to approximate the training instances by minimising the prediction error. A parameter ε defines a tube around the regression function. In this tube the errors are ignored. The parameter ε controls how closely the function will fit the training data. The parameter C is the penalty for exceeding the allowed deviation defined by ε. The larger the C, the closer the linear regression function can fit the data [33].

For our experiments the support vector regression (SVR) model [34], which employs the sequential minimal optimization (SMO) algorithm for training a support vector classifier [32], was used. Many kernel functions have been used in SVR such as the polynomial, the radial basis function (RBF) and the Gaussian functions [35], etc. In this paper, after some preliminary experiments, the RBF kernel was selected [35].

4. Experiments

As mention earlier, there are two hypotheses we are interested in verifying with the following experiments. Firstly, whether external models could build more robust phone duration models than the explicit modelling of HSMMs. Secondly, whether the SVR model, since SVMs cope in a better way with the high-dimensionality of the feature space than MLP, would outperform the MLP external PDM.

4.1. Experimental Setup

In this section the database along with the feature set used in the experiments are presented. Furthermore, the setup of the HSMM explicit duration model and the external SVR and MLP phone duration models are described. The same HSMM framework, used for the explicit duration modelling, was also used for the HMM-based speech synthesis models used for synthesizing speech (using the predicted by each model durations) for the subjective evaluation test. The implementation of the external SVR and MLP PDMs was done with the WEKA software [36].

4.1.1. Database and Feature Set

For the experiments, the RMS voice of the CMU-ARCTIC database was used [29] which is a database of standard size for speaker-dependent HMM-based speech synthesis. The RMS voice is an American English male containing 1320 sentences of reading style speech. The data were divided into three sets, a training set containing 900 sentences for training the three models, a development set of 100 sentences for fine tuning the external phone duration models and a test set of 132 sentences for evaluating the three models with objective and subjective evaluation tests. Throughout the entire database, all starting and ending silences in each sentence were removed. Only the internal silences (silences between words in each sentence - phone "pau") were kept and modelled as the rest of the phones. This concluded to a phone set of 41 phones.

Concerning the features used for training the HSMM system, a standard in HMM-based speech synthesis set of features was used, composed of phonetic and prosodic features such as phone identity, identity of the two previous and next phones, number of syllables in a word, accented/stressed syllable, etc., concluding to a 53 feature set. For the external phone duration models, the same feature set was used expanded with some additional articulatory features. These articulatory features correspond to information such as the category of the phone (e.g. vowel, approximant, nasal, etc.), vowel length (e.g. short, long, etc.), height (e.g. high, middle, low), frontness (e.g. front, mid-

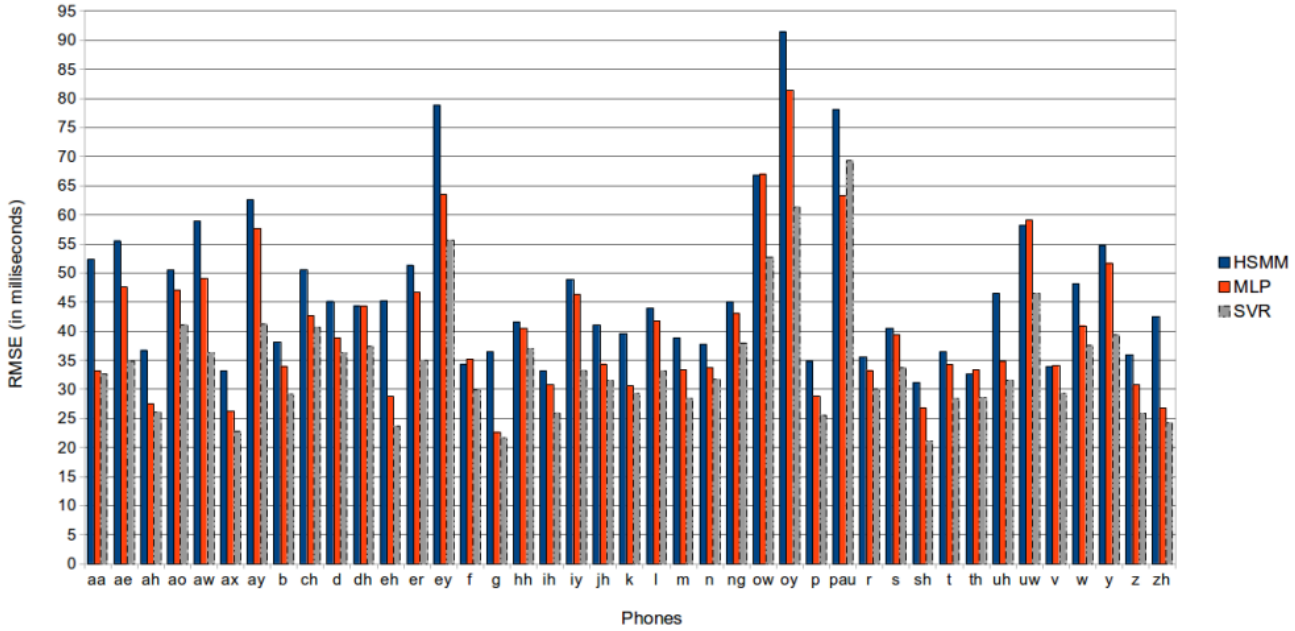

Figure 1: *This figure shows the RMSE for the three PDMs (HSMM, MLP and SVR) per phone*

dle, back), place of articulation (e.g. labial, alveolar, palatal, etc.), etc. The Relief [37] feature selection algorithm was used in some preliminary experiments for selecting among 37 binary articulatory features and their temporal (one previous and one following phones) information, the most appropriate ones. The final feature set consisted of 100 features.

4.1.2. HSMM model

For the implementation of the HSMM model, the version 2.2 of the HTS toolkit [38] was used. The speech data which were used had 16kHz sampling frequency. Five-state, left-to-right, no-skip HSMMs were used. The speech parameters which were used for training the HSMMs were 24th order mel-cepstral coefficients [39], log-f0 and 21-band aperiodicities [38], along with their delta and delta-delta features, extracted every 5 milliseconds (ms). The number of the used questions and the number of the leaf nodes of the decision tree were 304 and 547 respectively. STRAIGHT [40] was used for the analysis and synthesis phase of the HSMM-based speech synthesis.

4.1.3. MLP model

For the MLP model, a backpropagation based approach was used. The 100 input units (features) were converted to 570 units since all the nominal (categorical) features were converted to binary ones - an attribute with k nominal values is transformed into k binary attributes. The MLP model consisted of the input layer with 570 units, a hidden layer (H) with 10 units and an output layer with one unit (phone duration). The learning rate (L) of MLP, to ensure that the weights converge to a response fast enough without producing oscillations [41], was set equal to 0.05. The momentum term (M), which determines the degree to which each weight change will depend on the previous weight change, was set equal to 0.05. The epoch of the MLP (N), which determines the maximum number of iterations in which the full training set is presented in the model, was set equal to 500. These values were selected after a grid search (H={5:1:90}, L={0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 1.0, 1.5, 2.0}, M={0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 1.0, 1.5, 2.0}, N={50, 500, 1000, 5000, 10000, 50000}) of the model on the development set in respect to the RMSE of the model.

4.1.4. SVR model

For training the SVR model, as in the case of MLP, 570 features binary features were used. In our experiments the RBF kernel was used as mapping function for the SVR. 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 0.005 and 0.5 respectively. The gamma (G) parameter of the RBF function, determining the RBF width, was set equal to 0.05. These values were selected after a grid search fine tuning (ε={0.001, 0.003, 0.005}, C={0.5, 1.0, 1.5, 10, 100}, G={0.01, 0.03, 0.05}) of the model on the development set in respect to the RMSE of the model.

4.2. Experimental Results

For the evaluation of the models both objective and subjective tests were done for evaluating the accuracy of the models and the overall quality of the synthesized speech respectively.

4.2.1. Objective Evaluation

In the objective evaluation, the root mean square error (RMSE) in terms of milliseconds (ms) between the predicted and the reference (the original phone boundaries of the database) phone durations was used. To determine the phone duration prediction in ms using the HSMM model, the sum of frames of each of the five states on each phone was calculated and multiplied by the frame shift of the model. In Table 1, the overall performance accuracy of the three models (HSMM, MLP, SVR) on the development and test sets is presented. The MLP and the SVR models managed to outperform the HSMM one with a relative improvement in terms of RMSE of 11.56% and 25.09%

Table 1: *This table reports the accuracy in terms of RMSE (ms) for the three PDMs for the developement and test sets.*

Set	Phones	HSMM	MLP	SVR
Dev	All	43.89	37.11	**33.07**
Test	All	43.97	38.89	**32.94**
Test	Vowels	48.96	42.13	**33.95**
Test	Cons	39.91	36.31	**31.87**

Table 2: *This table shows the subjective evaluation (ABX test) for the three pairs (HSMM vs MLP, HSMM vs SVR and MLP vs SVR).*

ABX test	***HSMM vs MLP***			***HSMM vs SVR***			***MLP vs SVR***		
Set	**HSMM**	**Eq.**	**MLP**	**HSMM**	**Eq.**	**SVR**	**MLP**	**Eq.**	**SVR**
Set1	26.43%	28.57%	45.00%	15.00%	26.43%	58.57%	16.43%	51.43%	32.14%
Set2	24.62%	32.31%	43.08%	11.54%	34.62%	53.85%	18.46%	41.54%	40.00%
Both	25.56%	30.37%	44.07%	13.33%	30.37%	56.30%	17.41%	46.67%	35.93%

respectively, verifying our first hypothesis, i.e. these external models are able to build more robust models than the HSMM explicit duration modelling. Furthermore, the SVR model in comparison to the MLP model achieved a relative improvement of 15.3%, showing the superiority of the SVR over the MLP model, verifying our second hypothesis, i.e. the SVR could model better the phone durations in comparison to MLP. As it was expected, the SVR model managed to cope in a better way with the high-dimensionality of the feature space in comparison to the MLP model.

Moreover, as it was expected, since the development and test sets are not involved in the training procedure of the HSMM model, the RMSE for these sets are almost identical. In the case of the MLP model on the test set, a 4.79% relative decrease in terms of RMSE in respect to the development set (used for the fine tuning of the model) is achieved, showing some degree of overfitting of the model to the development set. On the other hand, in the case of the SVR model, even though the model is fine tuned using the development set, the RMSE for the development and test sets are almost identical, showing an additional advantage of the SVR over the MLP, i.e. the ability of SVR to make robust model without overfitting to the development set.

In Table 1, the overall accuracy in terms of RMSE on the test set separately for vowels and consonants is presented. It can be seen that these results follow the overall results described above. For all models, the RMSE calculated on the vowels is higher than the one on the consonants, which can be attributed to the fact that the mean of the standard deviation of the vowels (on the reference-original durations of the phones) is significantly higher than the respective one for the consonants.

In Figure 1, the RMSE in milliseconds for each phone for the three models on the test set is presented. It is shown that the SVR model managed to outperform the MLP and the HSMM models for all phones apart from the silence ("pau"), where MLP model achieved the best performance followed by the SVR model. The biggest difference between the SVR and the HSMM models is shown in phone "ey", where SVR model achieved a 37.74% relative improvement over the HSMM model in terms of RMSE. On the other hand the smallest difference for the respective models is shown in phone "hh", where SVR model achieved a 10.9% relative improvement over the HSMM model in terms of RMSE. In the comparison between SVR and MLP models, the biggest difference is shown in phone "ay", where SVR model achieved a 28.36% relative improvement over the MLP model in terms of RMSE. The smallest difference is shown in phone "aa", where SVR model achieved a 1.64% relative improvement over the MLP model in terms of RMSE. Comparing the MLP and HSMM models, it can be noticed that in several cases (e.g. "f", "th", "uw" phones), the MLP model was outperformed by the HSMM model. On the other hand the biggest difference for the respective models is shown in phone "g", where MLP model achieved a 37.76% relative improvement over the HSMM model in terms of RMSE.

4.2.2. Subjective Evaluation

In order to investigate whether the objective performance among the three models is reflected to the overall quality of the synthesized speech, a subjective evaluation test was done. Using the HSMM-based speech synthesis framework described earlier, the phone durations predicted by the SVR and MLP models were forced on the speech synthesis system, in order to synthesize speech using the predicted durations and determining internally the state sequence of the model.

The subjective evaluation was composed by three ABX tests, comparing each of the model to the other two. Two sets of ten sentences where randomly chosen from the test set and were evaluated by 14 and 13 subjects respectively. For every sentence, the subjects were presented with three pairs of samples (HSMM vs MLP, HSMM vs SVR and MLP vs SVR) in random order, without any knowledge about the three systems and a reference sample. In each case the subjects had to choose between the two samples of the pair in terms of sounding closer to the reference one (synthesized with forced alignment using the reference durations) or they could choose "equal" if they had no preference over them.

In Table 2, the results of the ABX tests are presented. As can be seen the SVR model was preferred over the MLP and HSMM models with a score of 35.93% over 17.41% and 56.30% over 13.33% respectively. Furthermore, the MLP model was preferred over the HSMM model with a score of 44.07% over 25.56%. These results follow the trend of the objective evaluation, showing that the SVR managed to build a robust model capable of outperforming the explicit HSMM model and moreover the other external model using MLP, in the overall quality of the synthesized speech.

5. Conclusions

In this paper we compared the HSMM-based explicit duration modelling with two external phone duration models using Support Vector Regression (SVR) and Multilayer Perceptron (MLP). The goal was to investigate whether and in which degree the external duration models outperform the HSMM model and if this is perceived in a subjective evaluation test on the quality of the synthesized speech. The experiments were done on an American English male speaker database. In both objective and subjective tests, it was shown clearly that the external duration models are more appropriate in the task of phone duration modelling in comparison to the explicit duration modelling of HSMMs, verifying our initial hypothesis.

Additionally, between the two external models, the SVR one outperformed the MLP model verifying our second hypothesis that the SVR would managed to tackle this task more efficiently. In the objective test, SVR model managed to outperform the MLP and HSMM ones showing a relative improvement in terms of root mean square error of 15.3% and 25.09% respectively. Finally, the subjective evaluation test showed that the superiority of the SVR model over the MLP and HSMM models is reflected also on the quality of synthetic speech, achieving a preference score of 35.93% and 56.30% over them, respectively. As future work it would be interesting to investigate how these three models perform when larger databases.

6. Acknowledgements

This work has received funding from the Swiss National Science Foundation under the SIWIS project.

7. References

[1] J. Laver, *Principles of Phonetics*. Cambridge: Cambridge University Press, 1994.

[2] T. Dutoit, *An Introduction to Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers, 1997.

[3] J. Yamagishi, H. Kawai, and T. Kobayashi, "Phone duration modeling using gradient tree boosting," *Speech Communication*, vol. 50, no. 5, pp. 405 – 415, 2008.

[4] T. Yoshimura, K. Tokuda, T. Kobayashi, T. Masuko, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *EUROSPEECH*, 1999.

[5] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *IEEE ICASSP*, vol. 2, 2001, pp. 805–808.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *EUROSPEECH*, 1997, pp. 2523 – 2526.

[7] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE - Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406 – 1413, 2007.

[8] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. of ICSLP*, 2004.

[9] S. Esmeir, S. Markovitch, and C. Sammut, "Anytime learning of decision trees," *Journal of Machine Learning Research*, vol. 8, pp. 891 – 933, 2007.

[10] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using Deep Neural Networks," in *IEEE ICASSP*, 2013, pp. 7962–7966.

[11] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *The Journal of The Acoustical Society of Japan (e)*, vol. 21, pp. 79 – 86, 2000.

[12] Y. Q. Zhi-Jie Yan and F. K. Soong, "Rich context modeling for high quality HMM-based TTS," in *INTERSPEECH*, 2009, pp. 1755–1758.

[13] Y.-J. Wu and R.-H. Wang, "HMM-based trainable speech synthesis for chinese," *J. Chinese Inf. Process.*, vol. 20, pp. 75 – 81, 2006.

[14] B. Gao, Y. Qian, Z. Wu, and F. K. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood." in *INTERSPEECH*, 2008, pp. 2266–2269.

[15] H. Lu, Y.-J. Wu, K. Tokuda, L.-R. Dai, and R.-H. Wang, "Full covariance state duration modeling for HMM-based speech synthesis," in *IEEE ICASSP*, 2009, pp. 4033–4036.

[16] Y. Ishimatsu, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Investigation of state duration model based on gamma distribution for HMM-based speech synthesis," IEICE Tech. Rep., Tech. Rep. 352, 2001.

[17] H. Siln, E. Hel, J. Nurminen, and M. Gabbouj, "Analysis of duration prediction accuracy in HMM-based speech synthesis," in *Proc. of Speech Prosody*, 2010.

[18] J. Latorre, S. Buchholz, and M. Akamine, "Usages of an external duration model for HMM-based speech synthesis," in *Proc. of Speech Prosody*, 2010.

[19] M. D. Riley, "Tree-based modeling for speech synthesis," in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoît, and T. R. Sawallis, Eds. Amsterdam: Elsevier, 1992, pp. 265 – 273.

[20] Q. Guo, N. Katae, H. Yu, and H. Iwamida, "Decision tree based duration prediction in Mandarin TTS system." *Journal of Chinese Language and Computing*, vol. 17, no. 1, pp. 97–106, 2007.

[21] O. Goubanova and S. King, "Bayesian Networks for phone duration prediction," *Speech Communication*, vol. 50, no. 4, pp. 301 – 311, 2008.

[22] A. Lazaridis, T. Ganchev, T. Kostoulas, I. Mporas, and N. Fakotakis, "Phone duration modeling: overview of techniques and performance optimization via feature selection in the context of emotional speech," *International Journal of Speech Technology*, vol. 13, no. 3, pp. 175–188, 2010.

[23] A. Lazaridis, I. Mporas, T. Ganchev, and N. Fakotakis, "Support Vector Regression fusion scheme in phone duration modeling," in *IEEE ICASSP*, 2011, pp. 4732–4735.

[24] K. K. Sreenivasa Rao and B. Yegnanarayana, "Modeling syllable duration in indian languages using Support Vector Machines," in *Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on*, 2005, pp. 258–263.

[25] U. Ogbureke, J. Cabral, and J. Berndsen, "Explicit duration modelling in HMM-based speech synthesis using a hybrid hidden Markov model-Multilayer Perceptron," in *SAPA - SCALE Conference*, 2012.

[26] K. S. Rao and B. Yegnanarayana, "Modeling duration of syllables using Neural Networks," *Computer Speech & Language*, vol. 21, no. 2, pp. 282 – 295, 2007.

[27] A. Lazaridis, I. Mporas, T. Ganchev, G. Kokkinakis, and N. Fakotakis, "Improving phone duration modelling using Support Vector Regression fusion," *Speech Communication*, vol. 53, no. 1, pp. 85 – 97, 2011.

[28] A. Lazaridis, T. Ganchev, I. Mporas, E. Dermatas, and N. Fakotakis, "Two-stage phone duration modelling with feature construction and feature vector extension for the needs of speech synthesis," *Computer Speech & Language*, vol. 26, no. 4, pp. 274 – 292, 2012.

[29] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223 – 224.

[30] T. Masuko, "HMM-based speech synthesis and its applications," Ph.D. dissertation, Tokyo Institute of Technology, 2002.

[31] S. Haykin, *Neural Networks: A Comprehensive Foundation, 2nd ed.* New York: Macmillan College Publishing, 1998.

[32] A. Smola and B. Scholkopf, "A tutorial on Support Vector Regression," Royal Holloway College, London, U.K., Tech. Rep. NeuroCOLT Tech. Rep. TR 1998-030, 1998.

[33] H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, second ed.* Morgan Kauffman Publishing, 2005.

[34] J. Platt, "Fast training of Support Vector Machines using sequential minimal optimization," in *Advances in Kernel Methods*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 185 – 208.

[35] B. Scholkopfand and A. Smola, *Learning with Kernels*. MIT Press, Cambridge, 2002.

[36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[37] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proceedings of the Tenth National Conference on Artificial Intelligence*, ser. AAAI'92. AAAI Press, 1992, pp. 129–134.

[38] "HMM-based speech synthesis system version 2.2," 2011. [Online]. Available: http://hts.sp.nitech.ac.jp

[39] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis — a unified approach to speech spectral estimation," in *Proceedings of the International Conference on Spoken Language Processing*, Yokohama, Japan, 1994, pp. 1043–1046.

[40] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *MAVEBA*, 2001.

[41] "Multilayer Perceptron in Wikipedia." [Online]. Available: http://en.wikipedia.org/wiki/Multilayer_perceptron