Analysis of Duration Prediction Accuracy in HMM-Based Speech Synthesis

Hanna Silén[1], *Elina Helander*[1], *Jani Nurminen*[2], *Moncef Gabbouj*[1]

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Nokia Devices R&D, Tampere, Finland

hanna.silen@tut.fi, elina.helander@tut.fi, Jani.K.Nurminen@nokia.com

Abstract

Appropriate phoneme durations are essential for high quality speech synthesis. In hidden Markov model-based text-to-speech (HMM-TTS), durations are typically modeled statistically using state duration probability distributions and duration prediction for unseen contexts. Use of rich context features enables synthesis without high-level linguistic knowledge. In this paper we analyze the accuracy of state duration modeling against phone duration modeling using simple prediction techniques. In addition to the decision tree-based techniques, regression techniques for rich context features with high collinearity are discussed and evaluated.

1. Introduction

Accurate prediction of phone durations is essential for high-quality text-to-speech (TTS). The use of unsuitable phoneme durations can deteriorate synthesis quality by decreasing the perceived speech naturalness. In some languages, such as in Finnish, phoneme quantity has a distinctive phonological role denoting that short and long phoneme quantities convey differing meanings. Hence, using an unacceptable phone duration can even change the meaning of a word.

Hidden Markov models (HMMs) provide a flexible framework for statistical parametric speech synthesis. HMM-based TTS (HMM-TTS) [1] uses HMMs for modeling a given speech database, and further, for generating speech representations corresponding to the input phoneme sequences. Two separate phases are included. In the training phase, an HMM representation of the speech database is constructed and during the synthesis phase, required models are concatenated and a parameterization for the synthetic speech is generated using one of the speech parameter generation algorithms [2].

In addition to spectrum modeling, in HMM-TTS, HMMs are used to model prosodic features. In this paper we focus on duration modeling. Explicit models are not typically used in HMM-TTS, although it would be possible to utilize for example CART (classification and regression tree) [3], multivariate regression, Bayesian networks [4] or artificial neural networks [5] for phone duration modeling. Instead, prosody is created through the use of context dependent training and a rich context feature representation. This representation involves not only context phone identities but also a large number of other phonetic and linguistic features. One major difference to the conventional duration modeling is that in HMM-TTS phoneme durations become modeled through combining HMM state durations. Another major difference is that the conventional duration models usually employ a relatively small set of carefully selected features while HMM-TTS relies on the use of a very large set of less sophisticated context features, say 1000-1500 binary features. Not all the features are important or effective, and the selection of the important ones is done automatically during prediction model training.

It is impossible to include all the rich context feature combinations of a language in the training data. For example in the case of 1000 features, in theory, there are 2^{1000} different feature combinations. Therefore, a mechanism for the prediction of unseen combinations must be provided. This is typically done by employing clustering of trained models. In HMM-TTS, a decision tree using minimum description length (MDL) difference as a cluster splitting criterion [6] is employed. The question set determining the possible splits contains a large amount of binary yes/no questions about the model representation. In addition to the prediction of unseen models, the use of model clustering decreases the TTS system footprint and also alleviates the effect of outliers present in the training data.

In this paper we evaluate the accuracy of different simple duration prediction techniques for an English and a Finnish speech database. In addition to the MDL-based clustering, regression techniques for rich context features with high collinearity are discussed and evaluated. Furthermore, we analyze the structure and the characteristics of the binary rich context feature matrix.

The paper is organized as follows. In Section 2 and 3, the duration modeling and prediction in HMM-TTS are described. The analysis of the duration prediction accuracy is presented in Section 4. Section 5 concludes the paper.

2. Duration modeling in HMM-based TTS

The duration of a phoneme is typically modeled through HMM state durations: each context-dependent phoneme is modeled as a sequence of states and the duration of the states is modeled. A state transition probability denoting a probability of moving from one state to another is determined. Typically, left-to-right models with no state skips are used, hence the transition probability for the transitions to other states except for the following state and the state itself are set to zero.

To model the state durations for synthesis, duration probability distribution for each state is determined. In HMM-TTS duration modeling the distributions are formed based on the statistics from HMM parameter re-estimation [7]. Each state duration probability distribution is regarded as a single Gaussian with a certain mean and variance. The mean and variance are extracted based on the average of all possible durations, each of them weighted with the corresponding state occupancy probability (i.e. probability of occupying the given state during the given time interval).

In HMM-based modeling, duration distributions are used in synthesis but they are not explicitly present in HMM parameter estimation. The state transition probabilities control the state durations instead. A more accurate modeling can be achieved using hidden semi-Markov model (HSMM)-based techniques [8]. In HSMM-based modeling, the duration distributions are explicitly present already in HMM training.

3. Methods for duration prediction

3.1. Decision tree-based clustering and MDL

Decision trees are popularly used for many reasons. First of all, they are non-parametric and suitable for modeling different kinds of input data. Secondly, they can give rather accurate predictions despite outliers and redundant input data. Furthermore, they are intuitive and thus easily understandable. Decision tree can be a regression or classification tree. Since phone duration is a continuous target variable, it is modeled by a regression tree. One problem related to the trees is that they do not have interpolation properties and thus the modeling of rare instances is prevented. Another problem for duration modeling is discontinuity, a pruned tree with a small amount of data can give rather distinct values for phone durations.

One issue that needs to be considered in tree building is the termination criterion for data splitting. Usually the tree is first grown into its maximal depth and then pruning is carried out using the optimal configuration given by cross-validation. Nevertheless, cross-validation is computationally demanding. Thus some alternative splitting criteria have been developed such as splitting based on MDL criterion.

In MDL-based decision tree clustering [6], both the acoustic similarity of a data cluster and an observation as well as the complexity of the resulting tree are taken into account during the tree construction. Contrary to the traditional decision tree clustering techniques, MDL-based tree construction can terminate unassisted when the optimal model for the data has been found. The tree construction algorithm starts with a one-node tree with all the data in one tree node and progresses in stages by selecting one node at a time to be split using a question selected from a given question set. In HMM-TTS, the data to be clustered consists of model distributions and the given question set of binary rich context feature questions. To provide reliable phone-level duration models, the state duration distributions of a phoneme are clustered together unlike in clustering of other model features, where the distributions are clustered statewise.

Each node-question pair is selected to minimize the description length of the resulting decision tree. Description length $l(U)$ of a model U is defined as [6]

$$\begin{aligned} l(U) = &\sum_{m=1}^{M} \frac{1}{2}\Gamma_m \left(K + K\log(2\pi) + log\,|\Sigma_m|\right) \\ &+ KM\log W + C \end{aligned} \tag{1}$$

where Γ_m denotes the total occupancy count at node S_m, K the dimensionality of the data vector, Σ_m the covariance of the Gaussian distribution at node S_m, and C a fixed constant value. $W = \sum_{m=1}^{M} \Gamma_m$, where M is the total number of leaf nodes in U. Splitting node S_m into two nodes S_{mqy} and S_{mqn} results in a new model U' with $M + 1$ leaf nodes. The best split is the one that minimizes the difference $\Delta_m(q)$

$$\begin{aligned} \Delta_m(q) =& l(U') - l(U) \\ =& \frac{1}{2}(\Gamma_{mqy} \log|\Sigma_{mqy}| + \Gamma_{mqn} \log|\Sigma_{mqn}| \\ &- \Gamma_m \log|\Sigma_m|) + K\log W \end{aligned} \tag{2}$$

subscript indices mqy and mqn indicating the *yes* and *no* branches of the split model starting from the node S_m. For each leaf-node, all possible questions are considered and the node-question pair minimizing the description length difference is selected, assuming the difference to be less than zero. Splitting is repeated until there are no possible nodes left for splitting.

3.2. Multivariate regression

Regression analyzes the relationship between two variables, X and Y, and constructs a model to describe it. The least squares solutions for the linear multivariate regression problem is

$$\beta = (X^T X)^{-1} X^T Y \tag{3}$$

where X is nxp matrix of predictors and Y is nxk matrix of responses. They are both centered to zero-mean.

Multicollinearity is a term used to describe a situation where two or more variables in X correlate with each other. This can affect the matrix rank. The matrix rank means a maximum number of linearly independent rows of X. Rich-context features in HMM-TTS exhibit strong multicollinearity, for instance the feature *Index of the syllable in the sentence* $\leq$ *12* is likely to correlate with the feature *Index of the syllable in the sentence* $\leq$ *11*. Although there may be instances that fulfill only the first condition, they may be rare if the sentences are short. Multicollinearity can result in a singular $X^T X$ term and further leading to a model that is still unbiased but results in high standard deviation of the prediction error. In Section 4 we analyze the feature matrix ranks for an English and a Finnish database.

Multicollinearity can be tackled by taking a pseudoinverse of $X^T X$ in Eq. 3. It is based on singular value decomposition; all the singular values that are below a certain threshold are omitted. Furthermore, *ridge regression* is another alternative technique coping with multicollinearity. It adds a constant term to the covariance matrix as

$$\beta = (X^T X + \lambda I)^{-1} X^T Y \tag{4}$$

where λ is a biasing or ridge parameter and I is an identity matrix. Ridge regression attenuates the effect of lower-variance components and results in a biased estimate but can give lower prediction error. The ridge parameter can be chosen by cross-validation. Several methods for determining λ have been proposed e.g. [9].

4. Experiments

4.1. Speech databases and context features

The aim of this paper is to analyze the effect of different duration prediction methods in the framework of HMM-TTS. The data used for the evaluations consisted of two manually annotated male speech databases. For English, a publicly available database CSTR US KED Timit (available at `http://festvox.org/dbs/dbs_kdt.html`) containing 450 utterances was used. For Finnish, a prosodically rich database of 650 utterances was used. Separate prediction question sets for English and Finnish were employed. The question set for English is presented in [1]. The Finnish question set was slightly modified from the English set; part-of-speech and accent were left out while some additional phoneme identity-related features were included.

For Finnish, long phoneme quantities were found problematic in terms of duration. For the speech parameter modeling, it is justified to use the same model for short and long phoneme quantities. However, in duration modeling the situation is more complicated since duration of a long phoneme cannot be defined simply as a combination of the duration of two short ones. For Finnish, long vowels were modeled as one phoneme, separate from the corresponding short phonemes. Models of long vowels were still initiated using initial models of the corresponding short phonemes. The high number (18) of diphthongs occurring in Finnish prevents their modeling as a single phoneme; Finnish diphthongs were modeled as two separate phonemes while the English diphthongs were treated as a single phoneme.

The statistics for the English and Finnish database are given in Table 1 and Table 2, respectively. In the tables, All refers to consonants, vowels, and pauses, i.e. all phoneme instances in the database. Further, separate analysis for consonants and vowels was carried out, the results being presented in the tables as well. The number of non-zero features refers to the number of those features that do not have all the instances either true or false. For example, when considering only consonants, the question *Is the current feature vowel* is false for every instance and thus it is removed. In addition, we calculated the matrix ranks for each case. As can be seen, the rank is much lower than the number of non-zero features, indicating high collinearity. Extraction of principal components for the Finnish data revealed that the first 250 components were able to explain 97% of the variance.

Table 1: *Database and feature matrix statistics for the English database. See the text for a more detailed description.*

	Instances	Features	Non-zeros	Rank
All	14 627	1 339	1 212	566
Consonants	8 304	1 339	1 102	496
Vowels	5 423	1 339	1 090	483

Table 2: *Database and feature matrix statistics for the Finnish database. See the text for a more detailed description.*

	Instances	Features	Non-zeros	Rank
All	45 065	1 378	1 140	492
Consonants	23 182	1 378	1 035	463
Vowels	19 994	1 378	1 016	451

4.2. Model training and state alignment

For duration prediction, explicit state durations were determined. To get the state alignment for each database phoneme, left-to-right HSMM models with no state skips were trained using the HMM-based speech synthesis system (HTS) [10], version 2.1. For speech parameterization, Mel-cepstral coefficients of order 39 with the dynamic features were used.

State durations of the speech data were estimated by Viterbi state alignment using the trained models. HSMM models were converted into HMMs by estimating the probabilities of state transitions. For each phone, a manually labeled quinphone context was taken into account and the alignment was allowed to move the context phone boundaries.

4.3. Evaluation of the duration prediction accuracy

The methods of Section 2 were compared in terms of duration prediction accuracy. The comparisons were carried out in three different categories: 1) Viterbi phone alignment 2) Viterbi state alignment and 3) manual phone durations. Both phone and state durations from Viterbi alignment were considered in the evaluations at phone level. The accuracy was evaluated by comparing predicted values to the reference durations given by the Viterbi alignment. In addition, prediction accuracy for manually labeled durations was evaluated separately from Viterbi alignment results.

Consonants and vowels were evaluated separately. For both, the training data was divided randomly into five subsets and the sets were kept fixed for all three categories. One subset was used as test data and the rest as training data. This was repeated for all of five subset divisions in a way that each phone was included exactly once in a test data set. In order to prevent the distortion caused by annotation outliers, a rough approximation for detecting them was used and they were removed before dividing the data into subsets. The phones with a manually labeled duration differing from the duration mean value by more than five standard deviations were removed.

The evaluation results are presented in tables 3-5. In the results, CART refers to a decision tree-based prediction using 10-fold cross-validation for tree pruning. Regression refers to multivariate regression of Eq. 3 using a pseudoinverse of $X^T X$. For ridge regression of Eq. 4, ridge parameter $\lambda = \left(p\sum_{i=1}^{n} e_i^2\right) / \left((n-p)\sum_{i=1}^{p}\beta_i^2\right)$ [9] is used. Here n, p, β_i, and e_i denote the number of samples, number of features, regression-based prediction, and resulting prediction error, respectively. MDL refers to MDL-based decision tree clustering of Eq. 1-2. For the regression-based techniques, each feature was required to have at least 1% of the instances to be true/false, otherwise the feature was ignored.

The prediction accuracy for Viterbi-based phone durations is presented in Table 3. For both English and Finnish vowels, ridge regression was able to provide the lowest root mean square error (RMSE) value and the highest (pseudo) R^2 value indicating the level of correlation between predicted and reference durations for the test data. For English consonants, difference of the prediction accuracy of MDL and regression-based techniques was negligible, CART-based approach producing the lowest accuracy. For Finnish consonants, the best accuracy was provided by MDL and CART-based techniques.

Table 3: *Phone duration prediction error for the speech databases using phone durations from Viterbi alignment. The average number of leaf nodes is given in parenthesis for both languages.*

	English		Finnish	
	RMSE (ms)	R^2	RMSE (ms)	R^2
Consonants				
CART (45,63)	24.9	0.50	**24.3**	0.40
Regression	24.1	**0.53**	25.0	0.37
Ridge regression	**24.0**	**0.53**	25.0	0.37
MDL (145,231)	24.1	**0.53**	**24.3**	**0.41**
Vowels				
CART (20,40)	30.3	0.50	25.6	0.51
Regression	28.4	0.53	24.6	0.54
Ridge regression	**28.0**	**0.55**	**24.5**	**0.55**
MDL (116,205)	29.5	0.50	25.1	0.52

For the state-level Viterbi durations, three of the prediction methods were considered: regression, ridge regression, and MDL-based tree clustering. The evaluation was done at phone-level, and the phone durations were obtained by summing over the five predicted state durations. As in Viterbi phone duration prediction, regression-based techniques outperformed MDL-based tree clustering for the vowels of both languages, ridge regression providing the lowest RMSE and highest R^2

value. For consonants, RMSE differences were smaller, ridge regression resulting in the lowest RMSE for English consonants and MDL-based clustering for Finnish consonants.

Table 4: *Phone duration prediction error for the speech databases using state durations from Viterbi alignment. The average number of leaf nodes is given in parenthesis for both languages.*

	English		Finnish	
	RMSE (ms)	R^2	RMSE (ms)	R^2
Consonants				
Regression	24.2	0.52	24.9	0.37
Ridge regression	**23.9**	**0.53**	24.9	0.38
MDL (148,262)	24.7	0.50	**24.4**	**0.40**
Vowels				
Regression	28.5	0.53	24.6	0.54
Ridge regression	**28.0**	**0.55**	**24.5**	**0.55**
MDL (96,258)	30.6	0.46	26.4	0.47

In addition to the state and phone durations from Viterbi alignment, the prediction techniques were also applied to the manually assigned phone durations. Prediction errors for both databases and each method are displayed in Table 5. For English and Finnish vowels and Finnish consonants, regression-based techniques produced the most accurate predictions. For English consonants, the accuracy differences between CART-based and regression-based prediction techniques were very small. The results for the English database are rather well in line with the results presented in [3].

Table 5: *Phone duration prediction error for the speech databases using phone durations from manual alignment. The average number of leaf nodes is given in parenthesis for both languages.*

	English		Finnish	
	RMSE (ms)	R^2	RMSE (ms)	R^2
Consonants				
CART (37,58)	**20.6**	**0.54**	16.9	0.64
Regression	20.8	0.53	**16.4**	**0.66**
Ridge regression	20.7	**0.54**	**16.4**	**0.66**
MDL (132,216)	21.7	0.48	16.6	0.65
Vowels				
CART (22,55)	25.1	0.55	18.0	0.66
Regression	23.8	0.60	**17.3**	**0.68**
Ridge regression	**23.2**	**0.61**	**17.3**	**0.68**
MDL (77,237)	27.1	0.47	17.6	0.67

In general, the regression-based prediction methods were able to provide more accurate duration prediction compared to the tree-based prediction methods. For consonants, accuracy differences were rather small and the prediction technique providing the best accuracy varied. This can indicate that for vowels, a global prediction model might be formed, while consonants are more challenging in terms of duration prediction. To further increase the accuracy of consonant duration prediction, consonant subsets could be formed in order to build better prediction models from the data.

Phone-level durations are prosodically better motivated units than state-level durations. This could be exploited by using proportional state durations to predict the state-level durations, but predict the phone durations directly using phone-level models. Currently in HMM-TTS, only state-level durations are employed. In the evaluations, MDL-based clustering turned out to produce less accurate predictions for the the Viterbi aligned state durations compared to the Viterbi aligned phone durations. For regression-based techniques, RMSE differences of state and phone-level predictions were negligible.

5. Conclusions

In this paper, we have studied the duration modeling accuracy in HMM-TTS. The evaluations included tests on English and Finnish speech data using simple and straightforward techniques (CART, MDL-tree, pseudoinverse regression, and ridge regression). Phone duration prediction was evaluated using both phone and state durations received from the Viterbi alignment of the training data. In addition, manually labeled database phone durations were considered. The results indicate, that for vowels, the regression-based techniques outperform the MDL-based tree clustering traditionally used in HMM-TTS while for consonants, the method providing the best accuracy varies.

This paper has studied duration modeling using objective measurements only. Although the preference for objective tests has been clear in the literature studying duration modeling, the authors acknowledge the fact that objective metrics may not reveal all the perceptual effects. Nevertheless, in the case of HMM-TTS, it is not straightforward to arrange meaningful listening tests because it is not possible to isolate the perceptual effects caused by duration changes. Thus, we considered objective tests to be more meaningful for the purpose of this paper.

6. Acknowledgements

This work was supported by the Academy of Finland, (application number 129657, Finnish Programme for Centres of Excellence in Research 2006-2011).

7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Eurospeech*, 1999.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *ICASSP*, 2000.

[3] J. Yamagishi, H. Kawai, and T. Kobayashi, “Phone duration modeling using gradient tree boosting,” *Speech Commun.*, vol. 50, no. 5, 2008.

[4] O. Goubanova and S. King, “Bayesian networks for phone duration prediction,” *Speech Commun.*, vol. 50, no. 4, 2008.

[5] M. Vainio, “Artificial neural network based prosody models for Finnish text-to-speech synthesis,” Ph.D. dissertation, University of Helsinki, Department of Phonetics, Faculty of Arts, 2001.

[6] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” *Acoustical Science and Technology*, vol. 21, no. 2, 2000.

[7] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, “Duration modeling for HMM-based speech synthesis,” in *ICSLP*, 1998.

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

[9] A. Hoerl, R. Kennard, and K. Baldwin, “Ridge regression: some simulations,” *Communications in Statistics*, vol. 4, 1975.

[10] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, “The HMM-based speech synthesis system (HTS) version 2.0,” in *6th ISCA Workshop on Speech Synthesis*, 2007.