Combining Extreme Learning Machine and Decision Tree for Duration Prediction in HMM based Speech Synthesis

Yang Wang[1], *Minghao Yang*[1], *Zhengqi Wen*[1], *Jianhua Tao*[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

{yangwang,mhyang,zqwen,jhtao}@nlpr.ia.ac.cn

Abstract

Hidden Markov Model (HMM) based speech synthesis using Decision Tree (DT) for duration prediction is known to produce over-averaged rhythm. To alleviate this problem, this paper proposes a two level duration prediction method together with outlier removal. This method takes advantages of accurate regression capability by Extreme Learning Machine (ELM) for phone level duration prediction, and the capability of distributing state durations by DT for state level duration prediction. Experimental results showed that the method decreased RMSE of phone duration, increased the fluctuation of syllable duration, and achieved 63.75% in preference evaluation. Furthermore, this method does not incur laborious manual alignment on training corpus.

Index Terms: speech synthesis, duration prediction, extreme learning machine

1. Introduction

Although Statistical Parametric Speech Synthesis (SPSS) achieved great success in the past twenty years, the naturalness of this type of synthetic speech is still far from that of unit selection system or natural speech [1, 2]. Hidden Markov Model (HMM) based speech synthesis system (HTS) [3, 4] is one typical implementation of SPSS. Among the many factors degrading naturalness, Decision Tree (DT) based context clustering in HTS is a main drawback. DT is the core feature mapping model from linguistic specification (i.e., full context label) to acoustic representation (i.e., the probability density function in each HMM state), thus it is the workhorse to deal with the most difficult problem of *generalization*. Although DT based feature mapping is applied to duration, excitation and spectrum parameters in a similar fashion, this paper focuses only on DT based duration prediction.

Some researchers tried to improve the conventional DT based method for duration prediction within the framework of HTS, e.g., combining duration information in state and phone levels [5], combining the state, phone and longer units [6], using full covariance state durations [7], and modeling duration distribution by Gamma instead of Gaussian [8]. Other approaches tried to employ external duration models independent of HTS, e.g., multivariate regression or ridge regression [9], Bayesian network [10], gradient tree boosting [11], support vector regression [12], multilayer perception (MLP) [13], and fusion of these models [14]. Some researchers tried to embed external duration model directly to HTS framework by forcing the generated speech wave to have the desired duration predicted by external duration models [15, 16]. However, these external models usually assume that a large corpus with manually aligned durations is available, which is not convenient in real applications.

The aim of this work is to improve the naturalness of synthetic speech by improving duration prediction accuracy, while it is not required to manually annotate phone boundaries in a large training corpus. The assumptions behind this work are: 1) the physical meaning of HMM states in HTS is vague and state is a rather low level, thus the need for accurate state duration prediction is less important compared to the need for accurate phone duration prediction; 2) DT based state duration prediction is not ideal, but its capability of duration distribution within a phone is acceptable if reasonable phone duration is given; 3) external duration models should be more accurate if durations are modeled at phone level instead of state level; 4) automatically aligned phone boundaries are useful to train external phone duration model, although they are not ideally accurate.

This paper proposes a combination method for duration prediction in HTS framework, which has the following properties:

- The duration prediction accuracy is improved both objectively and subjectively compared to HTS baseline for the purpose of speech synthesis.
- Laborious manual alignment of phone durations for a training corpus is not required. Instead, the automatically aligned phone durations after outlier removal are directly used to train external phone duration model.
- The method is a cascade of predicting phone durations by Extreme Learning Machine (ELM) [17] and predicting state durations by conventional DT in HTS.

The rest of this paper is organized as follows. Section 2 describes the two level duration prediction method with outlier removal in HTS. Experimental results in objective and subjective evaluations are presented in Section 3. Conclusion remarks are given in Section 4.

2. Proposed two level duration prediction

2.1. Motivation

Current HTS predicts a vector of state durations per phone, as the prediction of excitation and spectrum parameters are implemented at state level. However, state level is a rather low

level, which is not directly related to human perception; on the contrary, durations at higher level, such as phone or syllable, are more meaningful and important to perception. Thus we tried to model durations directly at phone level in this study. Furthermore, to be consistent with state level prediction for excitation and spectrum parameters in current HTS, state level durations are still needed even when phone level durations are given. As a result, a two level duration prediction method by combining prediction at phone and state levels emerges straightforwardly.

2.2. System overview

A two level duration prediction method with outlier removal in HTS is shown in Figure 1. This framework is general, thus it is easy to replace ELM with other learning machines for duration prediction purpose. Some important steps involved are briefly discussed below.

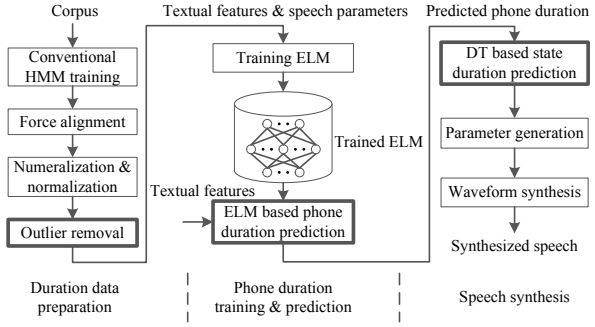

Figure 1: *Two level duration prediction in HTS.*

In duration data preparation part, force alignment is carried out at phone level after conventional HMM training. This step is to segment each utterance into a sequence of shorter and simpler speech units, thus each unit can be modelled independently in subsequent steps. Unlike DT, ELM can only handle numeric features, thus it is necessary to encode all nominal features to be numeric values. Normalization is immediately carried out to transform feature values into a limited interval. Outlier removal is then carried out for all speech units.

In phone duration training and prediction part, ELM is trained with "cleaned" training samples and stored. After that, phone level durations are predicted by ELM for any given textual features of full context label.

In speech synthesis part, state durations are further predicted by DT with phone duration predicted by ELM. Conventional parameter generation and waveform synthesis techniques are used to synthesize speech wave.

2.2.1. Outlier removal

Few previous researchers tried to remove duration outliers in duration prediction task, but outliers have dramatic degradation effect to most learning machines, thus we incorporated outlier removal in duration data preparation step.

2.3. Two level duration prediction

The two level duration prediction is a cascade of ELM based phone duration prediction and DT based state duration prediction.

2.3.1. ELM based phone duration prediction

ELM is single layer feedforward neural network (SLFN) [18]. The weights linking input layer to hidden layer and parameters of hidden neurons are randomly initialized and fixed. The only free parameters to be determined are the weights linking hidden layer to output layer, and the output weights can be solved analytically, thus the training process is typically very fast. An ELM for phone duration prediction is shown in Figure 2, in which the output function $f_L(\cdot)$ of ELM with L hidden nodes can be expressed as

$$f_L(\mathbf{x}) = \sum_{i=1}^{L} \beta_i g(\mathbf{a}_i \cdot \mathbf{x} + b_i), \quad \mathbf{a}_i, \mathbf{x} \in \mathbf{R}^d, \beta_i, b_i \in \mathbf{R} \qquad (1)$$

where β_i is the weight between output node and the i-th hidden node, and $g(\cdot)$ is sigmoid function in this study.

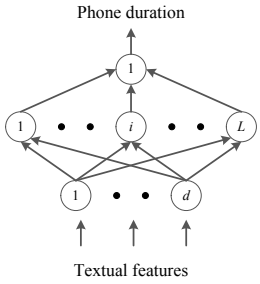

Figure 2: *An ELM for phone duration prediction.*

Despite the seemingly simple architecture, it is quite powerful and accurate for regression and classification tasks, which is assured by its interpolation theorem, universal approximation theorem and generalization bound [17]. Although ELM is similar to conventional MLP with only one hidden layer, the training criteria are completely different: conventional MLP is trained by back propagation and its variants, while ELM is trained by solving a system of linear equations with least square method.

2.3.2. DT based state duration prediction

State durations in HTS are determined by [19]

$$d_k = \mu_k + \rho \cdot \sigma_k^2 \qquad (2)$$

$$\rho = \left(T - \sum_{k=1}^{K} \mu_k\right) \Big/ \sum_{k=1}^{K} \sigma_k^2 \qquad (3)$$

where d_k is the k-th state duration, μ_k and σ_k^2 are the mean and variance of the k-th state duration Gaussian PDF, respectively, K is the number of states in HMM sequence, and T is the speech length of total K states. As speaking rate ρ is associated with T, we can alter speaking rate by setting total speech length T. This property is useful when integrating external duration model into HTS: if a total duration T is specified by some other duration model at a higher level instead of state level, the state durations d_k can be determined by (3) and (2). As Gaussian PDF $\mathcal{N}(\mu_k, \sigma_k^2)$ is determined by duration decision tree and textual features of speech unit, the state duration method given in (2) and (3) is thus called DT based state duration prediction in this paper.

3. Experiments

A Chinese speech corpus of 3,800 utterances (about 4.4 hours) recorded by native female speaker is used for training, and a separate set of 200 utterances is reserved for testing. The speech corpus consists of high quality, clean speech data under controlled recording conditions. Speech signal is sampled at 16 KHz frequency, windowed by 25-ms Blackman window for each frame with 5-ms shift, then 40th order Linear Spectral Pair (LSP) coefficients and fundamental frequency F0 in log scale are extracted as static features. The first and second order dynamics are appended to static features to form observation vector. Multi-space Probability Distribution Hidden Semi-Markov Model (MSD-HSMM) [20, 21] of 7 states, left-to-right with no skip topology are used to represent basic speech units. Single Gaussian with diagonal covariance matrix is used in each HMM state. Speech waves are forced aligned with its text transcription by HTS tool `HSMMAlign` [4]. The case 1 algorithm in [22] is used throughout our experiments for its simplicity.

Concerning the textual features used for training the baseline HTS and ELM, a variety of linguistic and phonetic features are used, such as the Initial/Final (IF) identity, identities of the two previous and next IFs, the number of IFs in a syllable, the tone of current syllable and its neighbors, resulting in a total of 75 textual features in full context label. All features in full label for the baseline HTS is encoded to numeric values and normalized, to be exact, nominal feature such as IF identity in full label is encoded with one hot method, and numeric feature is divided by its maximum value. All encoded values are then concatenated as predictive vectors of 604 dimensions to train ELM. It is noted that *initials* and *finals* are more natural than phonemes when studying Chinese Mandarin speech, and the term "phone" actually means initial or final hereafter.

Two duration prediction methods are compared:

1) Baseline: DT based duration prediction in HTS;

2) Proposed: (ELM+DT) based two level duration prediction with outlier removal.

3.1. Outlier removal

Careful manual checking shows that the aligned IF duration boundaries are roughly acceptable, but not very accurate. Some alignment errors may also occur if the speech wave does not strictly correspond to its text transcription. These errors usually cause the durations of some speech units in the problematic utterance too long, i.e., longer than a maximum duration based on our prior knowledge. Figure 3 shows the box plot of durations for each initial in Mandarin, where the jittered red dots are beyond the 99.3% coverage if duration data are normally distributed, and the blue plus sign shows the upper 1 percentile of all durations for each initial. It is clear that some initials are unreasonably long, and mismatch between speech wave and its text transcription is verified by visual checking.

Therefore, a simple outlier removal method is designed: given any initial, the upper 1% of durations of this initial is regarded as the maximum duration, and any utterance containing an automatically aligned speech unit longer than the maximum duration is removed; given any final, the upper 0.3% is used analogously. Such removal method shows its effectiveness in preliminary experiment.

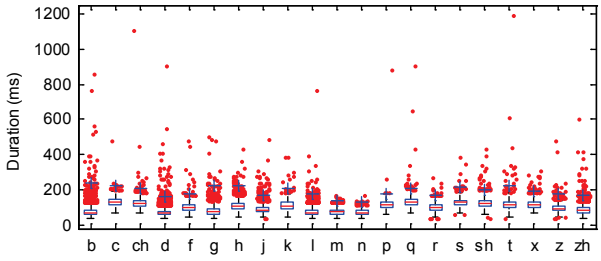

Figure 3: *Box plot of initial durations.*

3.2. Ground truth of initial/final durations

Before moving to the detailed experimental comparison, we must decide to use manual or automatically aligned durations as ground truth for objective evaluation. As the seemingly "true" durations need laborious human alignment, and the benefit of manual alignment vs. that of automatic alignment is not clear, we firstly compared these two alignment types. 10 utterances are randomly selected from the corpus, and segmented by automatic alignment and carefully manual alignment, respectively. These utterances are then synthesized using baseline HTS, where the only difference is how to set the phone level durations. We then compare which type of synthetic speech is more close to the original recorded speech in the aspect of perceptual duration assignment. The preference test result is show in Table 1. To our surprise, automatic alignment is better than manual alignment. This is probably because: 1) automatic alignment is inherently consistent with the reestimation procedure of HSMM [21], i.e., the duration predicted by duration model matches the spectrum and excitation models; 2) although manually aligned durations may be more accurate for some IFs, it is essentially very difficult to be consistent over a corpus; 3) some consecutive IFs are really difficult to be aligned with "true" boundaries by annotator perceptually and visually, especially when a syllable is consisted of only final. Therefore, the automatically aligned durations are used as rough ground truth.

Table 1. *Manual vs. automatic alignment.*

	Manual alignment	Automatic alignment	No preference
Preference (%)	19.1	32.7	48.2

Although there is no manual alignment in training corpus for ELM, i.e., no very accurate and reliable duration information are supplied to ELM, ELM can still perform better for phone duration prediction than DT. That is partly because the automatic alignment procedure takes three sources of information: speech, text, and model parameters, thus the alignment result is quite good, and further refinement by outlier removal enhances the accuracy and reliability of training corpus for phone duration prediction purpose.

3.3. Objective evaluation

Reserved 200 utterances are utilized to carry out the following objective evaluations. As silences before the start and after the end of an utterance are not very meaningful, they are excluded.

Figure 4 shows ELM performance vs. its hidden node number, in which the horizontal axis is the number of hidden nodes L in ELM, and the vertical axis is the Root Mean Squared Error (RMSE). Obviously, ELM is very unlikely to be over-fitting in a wide range of parameter values, which is a

merit. When the hidden node number is greater than 2500, ELM outperforms the baseline HTS in the sense of RMSE of phone duration prediction. The hidden node number 3500 is chosen as the optimal value in the following experiment.

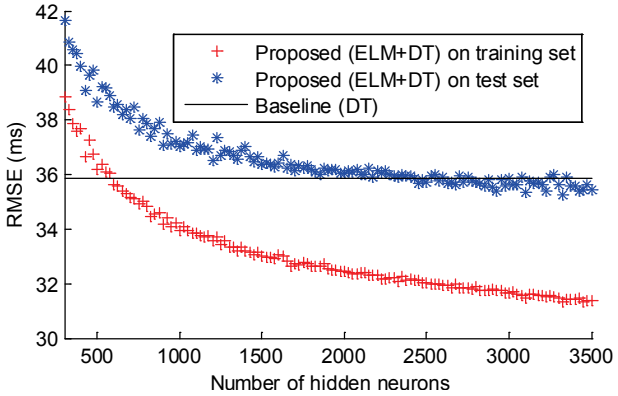

Figure 4: *ELM performance vs. its hidden node number.*

As the parameters of hidden nodes of ELM are randomly generated from uniform distribution over interval $(0,1)$, the performance may vary at different trials. We thus fixed hidden node number to be 3500, and train ELM for arbitrarily 13 times. The fluctuation of performance is small, as shown in Figure 5.

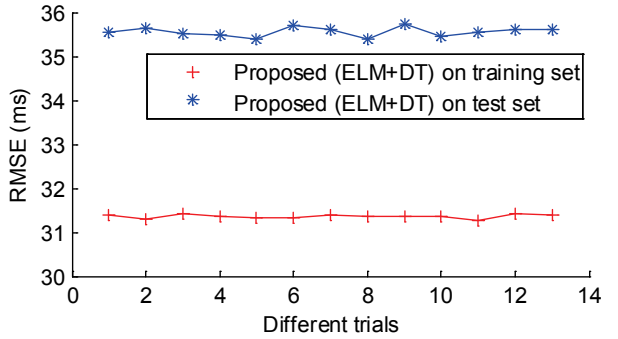

Figure 5: *The performance fluctuation of ELM.*

We then demonstrate the improvement of proposed method by two objective measures. The first measure is RMSE between the predicted durations and automatically aligned durations, where phone duration prediction is based on ELM and conventional DT in HTS, respectively. The result is presented in Table 2. The error decreased by 1.3%.

Table 2. *RMSE on phone duration prediction.*

	Baseline (DT)	Proposed (ELM+DT)
RMSE (ms)	35.86	35.38

The second objective measure is on the fluctuation of syllable durations in utterances. As predicted durations by HTS are normally over averaged, the standard deviation (SD) of all syllable durations in an utterance tends to be smaller compared to natural speech, thus it is an indicator for naturalness, and reasonably larger SD means improved naturalness. We calculate SD in syllable level instead of IF level, because initials are normally much shorter than finals, and if SD is calculated at IF levels, the contribution to SD are much affected by the interleaving of IF durations, not the more reasonable syllable durations. The result is presented in Table 3, which shows successful improvement: the fluctuation, measured by SD, increased roughly 4.5%.

Table 3. *Syllable duration fluctuation in utterances.*

	Baseline (DT)	Proposed (ELM+DT)
Fluctuation (ms)	59.50	62.19

3.4. Subjective Evaluation

10 out of the reserved 200 utterances are randomly selected to carry out subjective evaluation. 8 speech experts are asked to give their preferences where no preference is permitted. The result is presented in Table 4.

Table 4. *Preference evaluation.*

	Baseline (DT)	Proposed (ELM+DT)	No Preference
Preference (%)	6.25	63.75	30.00

Careful listening of synthetic samples shows that the improvement of duration prediction mostly lies in relatively long phrases (typically more than three syllables), in which the last syllable (if not unstressed) usually gets longer and one or more intermediate syllables gets shorter. Such finding is consistent with our prior knowledge. We also note that there are 30% of utterances getting "no preference" which are related to the fact that most of these utterances have no relatively long phrases, or original recordings are flat in duration assignment, i.e., the fluctuation of syllable durations is inherently small in these utterances.

The samples used to carry out subjective evaluation are supplied as attachment to interested reader.

3.5. Discussion

The proposed duration prediction method may be improved in the following aspects: 1) it should be advantageous to predict durations of initials and finals with different ELMs, as initials are typically much shorter than finals; 2) current outlier removal method is simple and "rude", and careful design of outlier detection and removal may be helpful.

4. Conclusions

We demonstrated the success of two level duration prediction method together with outlier removal in HTS for more accurate duration prediction. The proposed method does not require manually aligned corpus, instead, it takes automatically aligned durations with outlier removed to train ELM. Objective and subjective evaluations showed that the proposed method is more accurate than the conventional DT based state duration prediction method in HTS.

Future work includes combining ELM and DT for mapping excitation and spectrum parameters in HTS.

5. Acknowledgements

Thank three anonymous reviewers for their helpful comments. This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61403386, No.61305003, No.61332017, No.61375027, No.61273288, No.61233009, No.61203258), and the Major Program for the National Social Science Fund of China (13&ZD189).

6. References

[1] S. King, "Measuring a decade of progress in Text-to-Speech," in *Loquens*, 2014, pp. 1-12.

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech Synthesis Based on Hidden Markov Models," *Proceedings of the IEEE,* vol. 101, pp. 1234-1252, 2013.

[3] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," 2007, pp. 294-299.

[4] *HTS* [Online]. Available: http://hts.sp.nitech.ac.jp/

[5] Y. Wu and R. Wang, "HMM-Based Trainable Speech Synthesis for Chinese," *Journal of Chinese Information Processing,* vol. 4, pp. 75-81, 2006.

[6] Y. Qian, Z. Wu, B. Gao, and F. K. Soong, "Improved Prosody Generation by Maximizing Joint Probability of State and Longer Units," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1702-1710, 2011.

[7] 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 *ICASSP*, 2009, pp. 4033-4036.

[8] 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 Technical Report, SP2001-812001.

[9] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Analysis of duration prediction accuracy in HMM-based speech synthesis," in *proceedings of Speech Prosody*, 2010.

[10] O. Goubanova and S. King, "Bayesian networks for phone duration prediction," *Speech Communication,* vol. 50, pp. 301-311, 2008.

[11] J. Yamagishi, H. Kawai, and T. Kobayashi, "Phone duration modeling using gradient tree boosting," *Speech Communication,* vol. 50, pp. 405-415, 2008.

[12] A. Lazaridis, I. Mporas, T. Ganchev, G. Kokkinakis, and N. Fakotakis, "Improving phone duration modelling using support vector regression fusion," *Speech Communication,* vol. 53, pp. 85-97, 2011.

[13] U. Ogbureke, J. Cabral, and J. Berndsen, "Explicit duration modelling in HMM-based speech synthesis using continuous hidden Markov model," in *Information Science, Signal Processing and their Applications (ISSPA)*, 2012, pp. 700-705.

[14] 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, pp. 274-292, 2012.

[15] J. Latorre, S. Buchholz, and M. Akamine, "Usages of an external duration model for HMM-based speech synthesis," in *Proc. 5th Speech Prosody Workshop, Chicago*, 2010.

[16] A. Lazaridis, P.-E. Honnet, and P. N. Garner, "SVR vs MLP for Phone Duration Modelling in HMM-based Speech Synthesis," 2014.

[17] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Networks,* vol. 61, pp. 32-48, 2015.

[18] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," *International Journal of Machine Learning and Cybernetics,* vol. 2, pp. 107-122, 2011.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *ICSLP-98*, 1998, pp. 29-32.

[20] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-Space Probability Distribution HMM," *IEICE Transactions on Information and Systems,* vol. 85, pp. 455-464, 2002.

[21] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A Hidden Semi-Markov Model-Based Speech Synthesis System," *IEICE Transactions on Information and Systems E series D,* vol. 90, pp. 825-834, 2007.

[22] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*, 2000, pp. 1315-1318.