interspeech 2008 incorporating SST 08

Generating Natural F0 Trajectory with Additive Trees

Yao Qian[1], *Hui Liang**[1,2], *Frank K. Soong*[1]

[1]Microsoft Research Asia, Beijing, P.R.China
[2]School of Information Security Engineering, Shanghai Jiao Tong University, P.R.China
{yaoqian,frankkps}@microsoft.com, hui.ts.liang@gmail.com

10.21437/Interspeech.2008-551

ABSTRACT

In HMM-based TTS, while the segmental quality of synthesized speech is quite acceptable, intonation, especially at the sentence level, tends to be somewhat bland. The maximum likelihood (ML) criterion used in HMM training and parameter trajectory generation is partially responsible for the blandness. Additionally, the F0 trajectory thus generated has a smaller dynamic range than that of natural speech, and the synthesized speech does not sound lively. We propose to use multiple additive regression trees, a gradient-based, tree-boosting algorithm, for producing a more natural F0 trajectory. Multiple additive trees are trained in successive stages to minimize the error squares between natural and predicted F0 values. Additive tree modeling is integrated with MSD-HMM, which is an ideal model for characterizing the partially continuous (voiced/unvoiced) F0 contour. Experimental results in both Mandarin and English TTS trials show that the proposed approach can increase not only the dynamic range of generated F0 trajectory, but improve other objective (RMSE, correlation coefficient, voiced/unvoiced swapping errors) and subjective quality measures.

Index Terms: HMM-based TTS, modeling F0, additive trees

1. INTRODUCTION

HMM-based speech synthesis has been successfully applied to TTS synthesis in many different languages, e.g., Japanese, English and Mandarin [1-3]. In this framework, spectral envelop, fundamental frequency, and duration are modeled simultaneously by the corresponding HMMs. For a given text sequence, speech parameter trajectories can then be generated from trained HMMs in the Maximum Likelihood (ML) sense. Compared with the large corpus based concatenative speech synthesis, HMM-based speech synthesis is statistical model based, instead of waveform concatenation. The speech generated from it is fairly smooth and exhibits no apparent glitches. However, overly-smoothed parameter trajectories tend to make synthesized speech sound less lively than natural.

Many research attempts have been tried to reduce over-smoothing of F0 model and the resultant degraded synthesized speech quality. A straightforward way is to use a mixture of Gaussian components [1] to model F0 but it is not clear to use which mixture component in synthesizing a unique F0 contour. In [4], a parameter generation algorithm is proposed by considering the global variance (GV) of generated parameters. The probability of GV is used as a penalty for the reduced variance of generated trajectory. An extension which applies Gaussian mixture model to the GV term is used to improve the quality of an HMM-based polyglot speech synthesizer [5]. A trajectory model by imposing the explicit relationship between static and dynamic features was also proposed [6]. Minimum generation error is used as alternative criterion in HMM parameter optimization [7]. It tries to adjust HMM parameters trained by the conventional EM algorithm to minimize the generation error between synthesized and original parameter trajectories in training data. In [3], we propose to use the property of clustered line spectrum pairs (LSP) around a spectral peak to augment LSPs with their dynamic counterparts, both in time and frequency, in both HMM modeling and parameter trajectory generation. The formant structures of the generated spectra get clearer and the quality of the synthesized speech is considerably improved.

Generated pitch trajectories are also affected by the over-smoothing in the current HMM-based TTS systems. A naturally varying pitch contour is critical to the perceived synthesized speech quality. Successful approaches in reducing the effect of overly-smoothed spectral trajectory can be, in principle, extended to F0 trajectory. Unfortunately, not too many encouraging results have been reported. Unlike spectral features, no F0 is observed in unvoiced segments. Multi-Space Distribution (MSD) [8] has been proposed to model the partially continuous pitch contours in a statistically compact and rigorous manner. Pitch has limited contributions to phone segmentation in HMM training, thus only voiced segments taken from a corpus with the transcription of phone, syllable and word boundaries by performing a forced alignment are used for training the F0 model by linear regression [9], decision trees [10], nonlinear regression in multiple tiers [11] and additive cubic spline model [12]. However, how to make voiced/unvoiced decision for the states of HMMs in pitch trajectory generation by the above methods [9-12] is yet an unsolved problem. In this paper, we firstly review the current status of F0 modeling in HMM-based TTS. We then propose multiple additive regression trees for F0 modeling, and finally incorporate additive trees into an MSD-HMM framework for F0 trajectory generation.

2. F0 MODELING IN HMM-BASED TTS SYSTEM

In HMM-based speech synthesis, MSD-HMM [8] was proposed to model two, discrete and continuous, probability spaces: discrete for unvoiced regions and continuous for voiced F0 contours. It also uses a stream separated from the spectral features for F0 modeling. Richer prosodic contexts are used to capture F0 co-articulation effects in HMM modeling. However, in practice, limited by insufficient training data, we usually have to cluster models of long contexts into generalized ones to predict unseen contexts in test robustly. State tying via a clustered decision tree is commonly used. In synthesis, contextual MSD-HMM parameters are retrieved by traversing the trained decision trees, along with duration and the voiced/unvoiced decisions, a maximum likelihood F0 trajectory is generated with the dynamic feature constraints.

*Work performed as an intern in the Speech Group, Microsoft Research Asia

Figure 1 shows the sentence-based standard deviations of the natural and generated F0 trajectories of 20 English (Eng.) and 20 Mandarin (Man.) sentences selected randomly from our training data set. It illustrates that the variation range of synthesized F0 values is smaller than that of natural speech. On the other hand, the log likelihoods of synthesized F0 trajectories, computed in forced alignment, are higher than those of natural speech and they are tabulated in Table 1. It is interesting to rethink whether ML, which is used universally for training HMMs and generating speech trajectories, should be adopted as the ultimate criterion. It is also enlightening to observe that in the same figure F0 variances of Mandarin sentences are larger than those of English ones. The larger variance of Mandarin is partially due to the lexical nature of Mandarin tones where the variations of four (or five) lexical tones increase the dynamic range of F0 intrinsically.

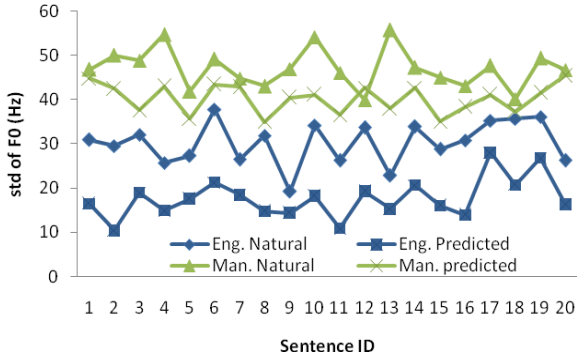

Fig. 1. The standard deviations of the natural and generated F0 trajectories of 20 English (Eng.) and 20 Mandarin (Man.) sentences.

Table 1. The average log likelihoods per frame of natural and F0 trajectories in training data.

	Natural	Predicted
Mandarin	2.65	3.35
English	3.09	3.77

3. ADDITIVE TREE MODELING OF F0

Classification and regression trees (CART) is an effective data mining tool which can efficiently handle messy data, missing values, or predictor variables measured in different scales. However, it also has some limitations e.g., lack of smoothness and difficulty in capturing underlying additive structure of the data. The lack of smoothness is due to the piecewise constant assumption. Using dynamic features to constrain parameter generation (in HMM-based speech synthesis) compensates this shortcoming to certain degree. The additive structure of F0 is commonly observed across different languages. For Mandarin, the theory of superposition of lexical tone and phrase or sentence intonation or the so-called "small ripples on top of big waves" was first proposed by Chao [14] and empirically confirmed. For English, multi-layer models have also been used to model the hierarchical structure of intonation components [11,12]. Multiple additive regression trees (MART) [13], an extension of CART model, is a generalization of gradient based tree-boosting algorithm that can capture additive structure of features.

We propose to use additive regression trees in conjunction with MSD-HMM for modeling F0 trajectory. The squared error, whose negative gradient is just the ordinary residual, is used as objective function in training. The algorithm of additive tree modeling of F0 in HMM-based speech synthesis is shown in Figure 2.

1.	Initialize; calculate global F0 mean of all voiced frames, $\hat{\alpha} = E(y_n), \quad \hat{y}_n^l \equiv 0, \quad l = 0$
2.	For l = 1 to L; grow L trees successively, given state alignment S for training sentences
	2.1 calculate residuals; $\varepsilon_n^l = y_n - \hat{\alpha} - \sum_{k=0}^{l-1} \hat{y}_n^k, \quad \Delta\varepsilon_n^l, \quad \Delta\Delta\varepsilon_n^l$
	2.2 update untied full-context model means, variances and weights of voiced and unvoiced subspaces; $\varepsilon_n^l, \quad \Delta\varepsilon_n^l, \quad \Delta\Delta\varepsilon_n^l \Rightarrow \Theta_l, \ \Theta = \{\mu, \Delta\mu, \Delta\Delta\mu; \quad \Sigma, \Delta\Sigma, \Delta\Delta\Sigma; \quad w\}$
	2.3 grow a tree for state tying which captures the l-th layer co-articulation and output tied full-context model, $\Theta_l \Rightarrow T_l(\Theta_l) \Rightarrow \hat{\Theta}_l$
	2.4 generate F0 trajectory $\hat{y}_n^l$, given $\hat{\Theta}_l$ and S in the ML sense with the dynamic feature constraints
3.	Stop when modeling error is less than a threshold ρ $E[(y_n - \hat{\alpha} - \sum_{l=1}^{L} \hat{y}_n^l)^2] \le \rho$

Fig. 2. The algorithm of additive tree modeling of F0 in HMM-based speech synthesis.

In the algorithm, L additive regression trees are built successively. The negative gradient, an unconstrained steepest descent direction, is used to minimize the mean square error between natural F0 trajectories of original speaker and generated F0 trajectories after the last iteration of standard HMM training. The HMM state sequences S of all training sentences are obtained by forced alignment with the trained spectral models. $\hat{\alpha}$ is the global F0 mean of all voiced frames in the training data; y_n is the F0 value of the n-th frame; $\hat{y}_n^l$ is the predicted F0 value by the l-th regression tree; and ε_n^l is the residual of the sum of the predictions from all $\{l\text{-}1\}$ trees. In building the l-th tree, we first update the untied full-context model parameters, Θ_l, including: means, variances and weights in the voiced and unvoiced subspaces. Then we grow a tree T_l for state tying which captures the l-th layer co-articulation and output tied full-context model, $\hat{\Theta}_l$. Finally, a predicted F0 trajectory, $\hat{y}_n^l$, is generated and the residuals are used for building the tree in the next stage, given $\hat{\Theta}_l$ and S. The additive tree growing process stops when the squared error is less than a pre-specified threshold ρ.

Voiced/unvoiced decisions made by L trees for a frame can be different in terms of the weight values of voiced and unvoiced subspaces. The motivation to do boosting is that by combining the outputs of many "weak" classifiers we can have "strong" classification results. The voiced/unvoiced predictions from all trees are combined through a weighted majority vote to produce a final decision.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

Two phonetically and prosodically rich broadcast news style corpora in American English and Chinese Mandarin are used in our experiments. The English corpus consists of 900 training

sentences and 100 testing sentences. The Mandarin corpus consists of 1,000 training sentences and 50 testing sentences. Both corpora were recorded by female speakers. Speech signals are sampled at 16 kHz, windowed by a 25-ms window shifted every 5-ms. The LPC of 40th order is transformed into static LSPs and their dynamic counterparts. Five-state, left-to-right HMM phone models, where each state is modeled by a single Gaussian, diagonal covariance output distribution, are adopted.

The phonetic and prosodic contexts are used as a question set in growing decision trees. They include tones and breaks for Mandarin; stress, accent and POS for English; quin-phone, the position of phone, syllable and word in phrase and sentence, and the length of word and phrase for both Mandarin and English. Minimum description length (MDL) criterion [15] for balancing model complexity and training data size is used as a stopping criterion for state clustering in decision tree growing.

4.2. Evaluation Results and Analysis

Three objective measures and one subjective measure are used to evaluate the performance of the additive trees based F0 model. Synthesis quality is measured objectively in terms of distortions between natural (speech of the original speaker) and synthesized speech frame-synchronously where oracle state durations (obtained by forced alignment) of natural speech are used. The three objective measures are F0 distortion in the root mean squared error (RMSE), correlation coefficient between the natural and synthesized F0 trajectories over aligned voiced frames, and voiced/unvoiced (v/uv) swapping errors. The subjective measure is an AB preference test between speech sentence pairs synthesized with the same segmental information but two F0 trajectories synthesized by the baseline (one tree) and additive tree models.

Although in theory MDL does not need any extra parameters to control the model size, it is still useful to control the model complexity more explicitly. Table 2 shows RMSE between natural and generated F0 trajectories with different values of the MDL control parameters. A smaller parameter value yields a larger sized model. It can then produce a lower RMSE in training. However, a compromised value which yields the best testing performance is 1 and 2 for Mandarin and English, respectively. It is not surprising that for English a larger parameter is found since English has a smaller F0 variance than Mandarin, as discussed in Section 2. We use these optimized parameters in all additive tree modeling experiments.

Table 2. RMSE between the natural and the generated F0 trajectories with different MDL control parameters.

MDL factor		0.5	1	2	4
Training	Man.	18.93	22.54	24.48	26.06
	Eng.	8.04	15.51	21.22	23.32
Testing	Man.	22.64	**22.46**	23.45	24.46
	Eng.	25.05	23.31	**22.73**	23.25

Additive regression trees, as presented in Section 3, are used for modeling F0 in both Mandarin and English. The RMSE and correlation coefficient measured for both training and testing data are shown in Figures 3 and 4. For training data, the RMSE of the natural and the generated F0 trajectories by additive trees is improved by 2.6Hz and 2.9Hz for Mandarin and English, over the baseline system with just one tree. The tree growing converges at four and six trees. For testing sentences, RMSE improvements of 0.90Hz and 1.01Hz, are obtained in Mandarin and English sentences, respectively. The correlation coefficient is improved from 0.91 to 0.93 for Mandarin and 0.76 to 0.82 for English in training, and from 0.89 to 0.90 for Mandarin and 0.67 to 0.70 for English in testing. A high correlation coefficient of 0.89 achieved by the baseline Mandarin TTS prevents it from being further improved significantly. The percentage of unvoiced/voiced swapping is measured on the testing data, as shown in Figure 5. There is a minimal change of unvoiced/voiced decisions when more additive trees are used, compared with the baseline. The percentage of such swapping for Mandarin is smaller than that of English. The above results are obtained with modeling F0 in log frequency. In linear frequency, the performance is only slightly higher.

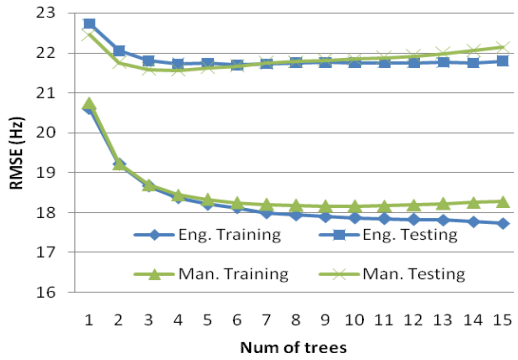

Fig. 3. RMSE between the natural and the generated F0 trajectories by the additive regression trees.

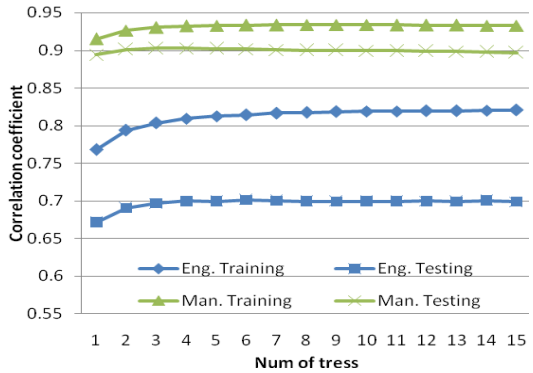

Fig. 4. Correlation coefficient between the natural and the generated F0 trajectories by the additive regression trees.

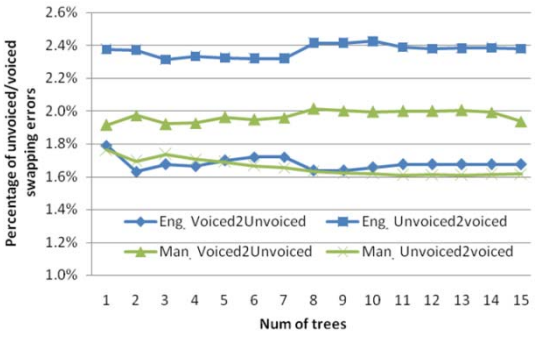

Fig. 5. The percentage of unvoiced/voiced swapping errors in F0 trajectories generated by the additive regression trees for testing sentences, compared with the natural F0 trajectories.

50 Mandarin and 50 English sentences, which are randomly selected from the testing set and synthesized by our baseline and additive trees with same state durations and spectral models, are evaluated in an AB preference test by 14 subjects composed of three English and three Mandarin language experts (LE) and eight bilingual graduate students (STU). The preference score between the baseline and the additive trees is shown in Table 3. The scores from both LE and STU subjects show additive tree model (66% and 56%) outperforms the baseline (34% and 44%).

Table 3. The preference score of the baseline and the additive trees

	LE	STU
Baseline (one tree)	34%	44%
Additive trees	66%	56%

We compared the standard deviation of F0 in all voiced frames in natural and synthesized speech generated with the baseline and additive tree models. The results of Mandarin and English in testing data are listed in Table 4, where it illustrates that F0 std is increased from 42.82Hz to 45.23Hz, and from 18.63Hz to 22.09Hz, or relative improvements of 35.0% and 26.4%, for Mandarin and English, respectively. The breakdown of RMSE improvement for the sentences in the AB test shows that about 78% of sentences predicted by the additive trees have lower RMSE than those of the baseline. An example of English sentence in its natural and generated F0 trajectories predicted by the baseline and additive tree model is shown in Figure 6. The predicted F0 by the additive trees fits the F0 trajectory of natural speech more closely, especially around the peak and valley regions. We also compared the footprint of the system using additive tree based F0 model with that of the baseline. Although the size of F0 model is tripled by using four additive trees, the footprint of the system is only 5% larger than that of the baseline since the footprint is dominated by the high dimensional segmental spectrum model.

Table 4. The standard deviation of F0s in all voiced frames of natural and generated testing data.

std of F0 (Hz)	Natural	Predicted	
		One tree	Additive trees
Man.	49.71	42.82	45.23
Eng.	31.68	18.63	22.09

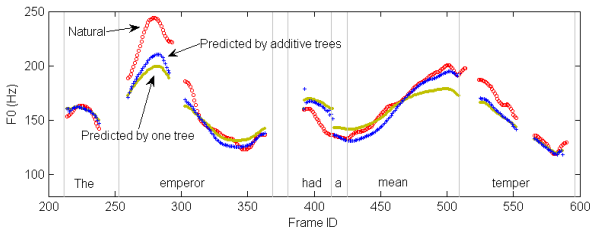

Fig. 6. An English sentence example of the natural and generated F0 trajectories predicted by the baseline and additive tree models.

4.3. Discussions

Two parameters need to be tuned in additive tree training: the size of each tree, J_{T_l}, and the number of trees, L. A smaller J usually results in a larger L. To choose the right-size tree, we have tried different values of J by adjusting the MDL control parameter. The best performance is obtained by setting the MDL control parameter to 1 for Mandarin and 2 for English, as shown in Table 2. An attempt in growing trees at various levels with specific question sets, e.g., global level: sentence or phrase, and local level: word, syllable and phone, and trials in a backfitting algorithm [12,13] was not successful.

5. CONCLUSIONS

Additive regression tree modeling is proposed to improve modeling F0 trajectory in HMM-based speech synthesis. The trees are generated in successive stages to minimize the error squares between the F0 trajectories of training sentences and predicted by additive regression trees. The experimental results also show that the proposed approach improves both three objective measures (RMSE, correlations and voiced/unvoiced swapping errors) and one subjective measure of AB preference. The synthesized sentences sound livelier than the ones generated by the baseline systems.

6. REFERENCES

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis", *Proc. of ICASSP*, 2000.

[2] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based Speech Synthesis System Applied to English", *Proc. of IEEE Workshop on Speech Synthesis,* 2002.

[3] Y. Qian, F. K. Soong, Y. N. Chen, and M. Chu, "An HMM-based Mandarin Chinese Text-To-Speech System", *Proc. of ISCSLP* , Springer LNAI Vol. 4274, pp.223-232, 2006.

[4] T. Toda, and K. Tokuda, "Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis", *Proc. of Eurospeech*, 2005.

[5] J. Latorre, K. Iwano, and S. Furui, "Combining Gaussian Mixture Model with Global Variance Term to Improve the Quality of An HMM-based Polyglot Speech Synthesizer", *Proc. of ICASSP*, 2007.

[6] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as A Trajectory Model by Imposing Explicit Relationships between Static and Dynamic Feature Vector Sequences", *Computer Speech & Language*, vol.21, no.1, pp.153-173, 2007.

[7] Y. J. Wu, R. H. Wang, and F. K. Soong, "Full HMM Training for Minimizing Generation Error in Synthesis", *Proc. of ICASSP*, 2007.

[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space Probability Distribution HMM", *IEICE Trans. Inf. & Syst.*, E85-D(3), pp.455-464, 2002.

[9] A. W. Black, and A. J. Hunt, "Generating F0 Contours from ToBI Labels Using Linear Regression", *Proc. of ICSLP*, 1996.

[10] K.E. Dusterhoff, A. W. Black, and P. Taylor, "Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours", *Proc. of Eurospeech*, 1999.

[11] X. Sun, "F0 Generation for Speech Synthesis Using A Multi-tier Approach", *Proc. of ICSLP*, 2002.

[12] S. Sakai, "Additive Modeling of English F0 Contour for Speech Synthesis", *Proc. of ICASSP*, 2005.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer, 2001.

[14] Y. Chao, *A Grammar of Spoken Chinese*, Univ. of California Press, 1968.

[15] K. Shinoda, and T. Watanable, "MDL-based Context-Dependent Sub-word Modeling for Speech Recognition", *J. Acoust. Soc. Jpn(E)*, vol.21, no.2, pp.79-86, 2000.