Prediction of voice aperiodicity based on spectral representations in HMM speech synthesis

Hanna Silén, Elina Helander, Moncef Gabbouj

Department of Signal Processing, Tampere University of Technology, Finland

hanna.silen@tut.fi, elina.helander@tut.fi, moncef.gabbouj@tut.fi

Abstract

In hidden Markov model-based speech synthesis, speech is typically parameterized using source-filter decomposition. A widely used analysis/synthesis framework, STRAIGHT, decomposes the speech waveform into a framewise spectral envelope and a mixed mode excitation signal. Inclusion of an aperiodicity measure in the model enables synthesis also for signals that are not purely voiced or unvoiced. In the traditional approach employing hidden Markov modeling and decision tree-based clustering, the connection between speech spectrum and aperiodicities is not taken into account. In this paper, we take advantage of this dependency and predict voice aperiodicities afterwards based on synthetic spectral representations. The evaluations carried out for English data confirm that the proposed approach is able to provide prediction accuracy that is comparable to the traditional approach.

Index Terms: aperiodicity prediction, hidden Markov model, speech synthesis

1. Introduction

Hidden Markov model (HMM) based speech synthesis [1] provides a flexible framework for statistical parametric speech synthesis. It enables simultaneous modeling of all speech features of the parameterization scheme and easy modification of individual features. A typical speech parameterization scheme employs source-filter decomposition that provides representations for the speech spectrum and excitation signal. STRAIGHT vocoder [2] is a high-quality speech analysis/synthesis tool that is widely used in HMM-based speech synthesis.

Modeling of speech aperiodicities is essential for high-quality waveform synthesis. A binary voicing decision describes whether the signal is voiced or not i.e. whether there is a fundamental frequency (F_0) associated with the signal or not. However, even for the speech segments defined voiced the vocal-cord vibration is not perfectly periodic. The amount of devoicing, occurring especially in the high frequencies, is described by the voice aperiodicity [3] and including it in the parameterization improves the vocoding quality. In HMM-based speech synthesis, average band aperiodicity (BAP) [4] is typically used for modeling mixed excitations. An alternative two-band mixed excitation parameterization for HMM-based speech synthesis has been proposed e.g. in [5].

In the conventional HMM-based speech synthesis, speech features such as spectral parameters, F_0, and voice aperiodicity are modeled simultaneously within the same HMM but clustered separately to provide prediction for unseen contexts and to cope with the data sparseness. For the prediction, typically statewise decision trees are built for each speech feature and the possible correlation between the spectral and aperiodicity parameters is thus not taken into account. In the training, the spectrum part is needed to create reliable labeling for the training data and to provide segmental intelligibility for synthesis. Intonation is modeled segmentally or supra-segmentally based on the training data F_0 values. Voicing decisions are typically derived from the weights of the voiced and unvoiced distributions of the multi-space probability distribution HMMs (MSD-HMMs) [6] used for F_0 modeling. Even though the aperiodicity measure is needed in synthesis, its role in HMM training is rather limited. However, increasing the number of model parameters also increases the computational load of the training.

Asynchronous speech feature modeling for HMM-based speech synthesis has been proposed e.g. in [7], where the asynchronous HMM parameter estimation of spectral parameters and F_0 was found to increase the F_0 prediction accuracy. In this paper, we investigate the modeling of voice aperiodicity and propose an alternative, asynchronous modeling scheme for the bandwise aperiodicity and voicing decisions. Instead of the traditional synchronous HMM training of STRAIGHT speech parameters combined with the asynchronous model clustering, we propose to predict signal aperiodicity and voicing decisions afterwards based on synthetic spectral parameters. Prediction based on the spectral representation instead of the context-dependent labels also takes into account the possible correlation between spectral and aperiodicity parameters.

The proposed approach enables more efficient HMM training by decreasing the number of HMM model parameters and the use of synthetic spectral parameters as a basis for the prediction ensures that the voice aperiodicity and voicing decisions are aligned with the spectral representation. The proposed prediction scheme employs local multivariate regression-based modeling with Gaussian mixture models (GMM) and dynamic modeling, an approach similar to [8], where the approach was used for spectral transformation in the framework of voice conversion. The objective evaluation shows that the proposed approach is able to provide a prediction accuracy comparable to the traditional HMM-based approach.

The paper is organized as follows. Section 2 gives an overview of the HMM-based speech synthesis and the widely used STRAIGHT parameterization scheme. Section 3 describes the proposed prediction approach using multivariate regression and GMM modeling. Analysis of the prediction accuracy is given in Section 4. Section 5 concludes the paper.

2. Overview of HMM-based synthesis

2.1. Speech parameterization using STRAIGHT

In parametric speech synthesis, such as HMM-based synthesis, speech is parameterized into a form that allows control on the perceptually important features of speech. Typi-

cal parameterization schemes use the familiar source-filter decomposition to decompose speech into spectral and excitation parts. STRAIGHT analysis/synthesis tool [2] provides a flexible framework for this decomposition. It parameterizes speech waveform into a spectral envelope without periodic interferences in time or frequency domain [2] and a mixed mode excitation signal [9].

The mixed mode excitation signal of STRAIGHT consists of F_0 and the level of voice aperiodicity. Frequency domain voice aperiodicity is defined as the relative energy of aperiodic components [9] and it is estimated as a ratio between the inharmonic component energy and the total energy of the warped spectrum with a constant F_0 and regular harmonic structure [9]. For HMM-modeling, BAP values are typically used instead of aperiodicity of every frequency bin. Binary voicing decisions determine whether there is F_0 related to the signal segment or not.

Estimation of the spectral envelope of STRAIGHT uses pitch-adaptive time windows and complementary windows to reduce the time-domain periodic interferences. This is followed by inverse filtering removing frequency domain interference while preserving harmonic structure [2]. For HMM-modeling, spectral envelope is typically encoded as perceptually better motivated mel-cepstral coefficients (MCCs) [10] or line spectral frequencies (LSFs).

2.2. HMM modeling of speech

In HMM-based speech synthesis, speech is modeled using context dependent HMMs [1]. Modeling typically involves 5-state left-to-right HMMs with no skips allowed. Statewise observations are modeled using Gaussian densities or mixtures of them. Duration densities can be explicitly included in the modeling by using hidden semi-Markov models (HSMMs) [11].

The training phase HMM parameter estimation aims at finding a parameter set λ^* that maximizes the probability $P(\mathbf{O}|\lambda)$:

$$\lambda^* = \arg\max_{\lambda} P(\mathbf{O}|\lambda) = \arg\max_{\lambda} \sum_{all\ \mathbf{q}} P(\mathbf{O}, \mathbf{q}|\lambda), \quad (1)$$

where $\mathbf{O}$ denotes a matrix of training data observations of delta-augmented speech parameters and $\mathbf{q}$ a hidden state sequence. $P(\mathbf{O}, \mathbf{q}|\lambda)$ refers to the conditional probability of $\mathbf{O}$ and $\mathbf{q}$ given the model parameters λ. A local optimum is found by the expectation maximization algorithm.

The HMM models are further used in synthesis to generate synthetic parameter trajectories. A sentence HMM is formed by concatenating the required context-dependent models and one of the speech parameter generation algorithms [12] is used for finding the maximum-likelihood matrix $\mathbf{O}^*$ of T observations:

$$\mathbf{O}^* = \arg\max_{\mathbf{O}} P(\mathbf{O}|\lambda^*, T). \quad (2)$$

2.3. Prediction of unseen contexts

Including all context-dependent phones of a language in the training database is practically impossible. To cope with the data sparseness and to enable synthesis for unseen contexts, minimum description length (MDL) based decision tree clustering is typically used [13]. In the training phase, statewise decision trees are formed for each speech feature and state and these trees are then used in synthesis to predict model parameters for the labels unseen in the training data. Construction of a MDL-based decision tree takes into account both the acoustic similarity of the cluster models and the tree complexity. Data is clustered iteratively using binary decisions based on MDL criterion.

3. Local prediction models using multivariate regression and GMMs

In this paper, we propose to use local prediction for aperiodicity features (BAP and voicing decisions) based on synthetic spectral features in HMM-based speech synthesis. The training phase aims at finding a prediction function that provides a mapping from predictors into responses and in the synthesis phase, the formed prediction function is then further used for mapping data unseen in the training phase.

The prediction employs spectral parameters modeled using MCCs and GMM-based local modeling originating from the GMM-based voice conversion introduced in [14]. We use local mappings from dynamics-augmented spectral representations into aperiodicity parameters. The approach was successfully applied to spectral conversion in [8], where GMM-based local mappings and dynamic information combined with the use of partial least squares regression were used for transforming the spectral parameters of one speaker into the spectral parameters of another specific speaker. In this paper, standard multivariate regression using pseudoinverse is used instead of partial least squares regression.

3.1. GMM modeling

In GMM modeling, the distribution of an input vector $\mathbf{x}_t$ is modeled as a sum of N Gaussian components:

$$p(\mathbf{x}_t) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n), \quad (3)$$

where α_n is the prior probability of the nth Gaussian and $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_n$ and covariance $\boldsymbol{\Sigma}_n$. Parameters of the Gaussian model can be estimated using expectation maximization algorithm.

The posterior probability $\omega_{n,t}$ of the observation $\mathbf{x}_t$ belonging to the nth cluster is defined as:

$$\omega_{n,t} = \frac{\alpha_n \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)}{\sum_{m=1}^{N} \alpha_m \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}. \quad (4)$$

In the following section, the posterior probabilities are used to enable the forming of local mappings from spectral parameters into aperiodicity parameters. Compression of the posterior probability dynamics can be used to avoid the dominance of single Gaussian components.

3.2. Multivariate regression with GMMs

The mapping from spectral representations into aperiodicity representations can be found using multivariate regression. It aims at modeling the relation between predictors $\mathbf{x}_t$ and responses $\mathbf{y}_t$:

$$\mathbf{y}_t = \boldsymbol{\beta}\mathbf{x}_t + \mathbf{e}_t, \quad (5)$$

where $\boldsymbol{\beta}$ denotes a regression matrix providing a mapping from a column vector $\mathbf{x}_t$ into a column vector $\mathbf{y}_t$ ($t = 1, \ldots, T$) and $\mathbf{e}_t$ denotes modeling error.

Instead of one global mapping, we employ a set of local mappings enabled by the use of GMM-based modeling as proposed in [8]. The spectral parameter vectors are expanded to

form new predictors $\tilde{\mathbf{x}}_t$:

$$\tilde{\mathbf{x}}_t = \begin{bmatrix} \omega_{1,t}\mathbf{x}_t \\ \omega_{2,t}\mathbf{x}_t \\ \cdots \\ \omega_{N,t}\mathbf{x}_t \end{bmatrix}, \quad (6)$$

where the weights $\omega_{n,t}$ are taken from the posterior probabilities of (4). The new regression model is:

$$\mathbf{y}_t = \boldsymbol{\beta}\tilde{\mathbf{x}}_t + \mathbf{e}_t. \quad (7)$$

The standard multivariate regression using pseudoinverse can be used to find the least-squares solution for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left[(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}\right]^T, \quad (8)$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_T]^T$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T]^T$, both centered to zero-mean.

3.3. Dynamic modeling

In speech signals, adjacent frames tend to exhibit rather strong correlation. To exploit this correlation to improve the modeling accuracy, we augment the predictor $\tilde{\mathbf{x}}_t$ of time t with the corresponding representations of the neighboring frames.

The dynamics-augmented input vector of multivariate regression at time t is denoted by $\tilde{\mathbf{x}}_t^d$ and it can be used instead of $\tilde{\mathbf{x}}_t$ in (7). It is formed by augmenting the regression input vector $\tilde{\mathbf{x}}_t$ with the representations of the preceding frame $\tilde{\mathbf{x}}_{t-1}$ and the following frame $\tilde{\mathbf{x}}_{t+1}$:

$$\tilde{\mathbf{x}}_t^d = \begin{bmatrix} \tilde{\mathbf{x}}_{t-1} \\ \tilde{\mathbf{x}}_t \\ \tilde{\mathbf{x}}_{t+1} \end{bmatrix}. \quad (9)$$

Dynamic modeling is likely to increase the correlation of the adjacent predictors since after augmentation, vector $\tilde{\mathbf{x}}_t$ can be included in the regression input vectors $\tilde{\mathbf{x}}_{t-1}^d$, $\tilde{\mathbf{x}}_t^d$, and $\tilde{\mathbf{x}}_{t+1}^d$.

4. Evaluation

The experiments carried out consisted of objective evaluation of the prediction accuracy of the traditional aperiodicity modeling described in Section 2 and the proposed prediction scheme of Section 3 predicting aperiodicities based on synthetic spectral parameters.

The objective evaluation shows that the accuracy difference between the traditional and proposed method is rather small. In BAP prediction, the traditional approach slightly outperformed the proposed approach whereas for the prediction of voicing decisions, the proposed approach with GMMs resulted in slightly higher accuracy. The small difference suggests that comparable accuracy can be achieved by the traditional and proposed approaches. The small differences are, however, extremely difficult to detect in the synthesized waveforms due to the vocoding and no full-scale listening tests were carried out. Instead, the reader is encouraged to listen to the randomly chosen synthesis samples available at *http://www.cs.tut.fi/sgn/arg/silen/is2011/AperiodicityPrediction.html*.

4.1. Speech databases

The evaluation data consisted of English speech data from CMU ARCTIC databases available at *http://festvox.org/cmu_arctic/*, a female voice database `slt` and a male voice database `rms`. For both speakers, half of the data (Set A with 593 sentences) was used for training and the remaining half (Set B with 539 sentences) for testing. All the models were trained speaker-dependently using all the sentences of the training data.

Speech waveforms were parameterized using STRAIGHT into a spectral envelope, F_0, and relative voice aperiodicity that were further modeled as MCCs of order 24, logarithmic F_0, and BAP values of the five frequency bands (0-1kHz, 1-2kHz, 2-4kHz, 4-6kHz, 6-8kHz), respectively. For HMM training described in Section 2, speech parameters were augmented with the 1st and 2nd order deltas. In the BAP and voicing decision prediction of Section 3, the systems with dynamic modeling employed the source representation augmentation of (9).

4.2. System description

Five systems were considered in the evaluation:

- *Proposed I*: prediction based on spectral parameters (MCCs) using GMMs and multivariate regression with dynamic modeling,
- *Proposed II*: as *Proposed I* but without dynamic modeling,
- *Proposed III*: prediction based on spectral parameters (MCCs) using standard multivariate regression with dynamic modeling (no GMMs),
- *Proposed IV*: as *Proposed III* but without dynamic modeling, and
- *Baseline*: traditional HMM-modeling with HSMMs and decision tree-based context clustering.

In the systems *Proposed I-II*, BAP values and voicing decisions were predicted from synthetic spectral parameters modeled as MCCs using the approach of Section 3. In the training, GMMs with $N = 8$ Gaussian components with diagonal covariance matrices were trained based on synthetic versions of the training data MCCs. Mappings from MCCs into BAP values or voicing decisions were obtained using the regression model of (7) with compressed posterior probabilities. In the synthesis phase, the models were used to predict aperiodicities based on MCCs unseen in the training. As a reference, systems *Proposed III-IV* were trained using direct prediction based on MCCs without using GMMs. In each case, the prediction models for BAP and voicing decisions were trained separately using synthetic MCCs (omitting the zeroth coefficient) and the BAP values or voicing decisions of the recorded data of Set A. For the systems *Proposed I* and *III*, dynamic modeling of (9) was used.

The system *Baseline* refers to the standard HMM-based approach with the context clustering resulting in approximately 100 nodes for each of the five states in BAP modeling for the voice `slt`. The voicing decisions of the system were derived from the probabilities of the voiced and unvoiced distributions (decision threshold 0.5) of the trained F_0 MSD-HMMs with approximately 300 clusters for each state for the voice `slt`. For the voice `rms`, the number of nodes in both BAP and F_0 modeling was somewhat higher.

For all the systems, models for the spectral MCC values and logarithmic F_0 as well as the BAP models in *Baseline* were trained using the standard HMM-based approach with HSMMs (continuous-density or MSD) and context clustering provided by the Hidden Markov model-based speech synthesis system (HTS) [15]. For the systems *Proposed I-IV*, only the center-most frame of each non-pause state was considered in the BAP and voicing decision model training whereas for the *Baseline*,

Table 1: *RMSE values for the average voice aperiodicity of the five frequency bands and voicing decision error percentages for the CMU ARCTIC databases* `slt` *and* `rms`.

	Band aperiodicity					Voicing
	1	2	3	4	5	
`slt`						
Baseline	**4.53**	**4.32**	**2.98**	**2.37**	**2.01**	**7.2 %**
Proposed I	**4.66**	**4.52**	**3.06**	**2.40**	**2.03**	5.5 %
Proposed II	4.94	4.72	3.12	2.41	**2.03**	**5.4 %**
Proposed III	5.24	5.05	3.28	2.47	2.05	5.7 %
Proposed IV	5.57	5.28	3.36	2.49	2.06	5.7 %
`rms`						
Baseline	**3.97**	**3.70**	**3.03**	**1.99**	**1.84**	**6.5 %**
Proposed I	**4.38**	**3.93**	**3.15**	**2.05**	**1.88**	**6.1 %**
Proposed II	4.66	4.10	3.25	2.08	1.89	6.4 %
Proposed III	5.27	4.53	3.49	2.17	1.93	8.5 %
Proposed IV	5.52	4.64	3.54	2.19	1.94	8.9 %

all the data in Set A was used. For a straightforward prediction accuracy comparison, Viterbi-aligned state durations of the corresponding recorded test sentences were used in synthesis.

4.3. Analysis of the prediction accuracy

The results of the objective evaluations for BAP and voicing decision prediction are given in Table 1. For the systems *Proposed 1* and *Baseline*, the differences in root mean square error (RMSE) in BAP prediction and the percentage of erroneous voicing decisions are rather small for both datasets `slt` and `rms`.

In BAP prediction, the differences in RMSE values for the systems *Proposed 1* and *Baseline* are small, suggesting that the proposed prediction scheme of Section 3 is able to provide prediction accuracy comparable to the traditional HMM-based approach of Section 2. The comparison of the systems *Proposed I-IV* shows that both the use of GMM-based modeling and dynamics can increase the prediction accuracy compared to the direct mapping from spectral parameters into bandwise aperiodicities.

The relative amount of errors in voicing decision prediction is shown in the rightmost column of the table. For both speakers, the systems *Proposed I-II* have provided a somewhat smaller prediction error compared to the system *Baseline*. For the female speaker `slt`, the differences between the systems *Proposed I-IV* are small whereas for the male speaker `rms` there is a larger difference depending on whether the GMM-based local mapping is used.

5. Conclusions

In this paper, we proposed a method for the prediction of voice aperiodicities in the framework of HMM-based speech synthesis. Instead of the traditional approach using context dependent HMM modeling and context clustering for all speech features, we proposed to use a prediction scheme with spectral parameter-based prediction. The proposed approach employs GMM modeling and multivariate regression to form local mappings from synthetic spectral features into bandwise aperiodicities and voicing decisions. The role of band aperiodicity in HMM parameter estimation is limited and it can therefore be left out from the training. The voicing decision modeling is typically embedded in the F_0 modeling. Analysis of the prediction accuracy on English data reveals that there is only a small accuracy difference between the proposed and traditional approaches, for the preference of the traditional approach in bandwise aperiodicity prediction and for the preference of the proposed approach in the voicing decision prediction. The results suggest that in the framework of HMM-based speech synthesis, voice aperiodicities can be predicted based on synthetic spectral features.

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] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, 1999.

[3] O. Fujimura, "An approximation to voice aperiodicity," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 1, 1968.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Eurospeech*, 2001.

[5] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Parameterization of vocal fry in HMM-based speech synthesis," in *Interspeech*, 2009.

[6] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *ICASSP*, 1999.

[7] C.-C. Wang, Z.-H. Ling, and L.-R. Dai, "Asynchronous F0 and spectrum modeling for HMM-based speech synthesis," in *Interspeech*, 2009.

[8] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 18, no. 5, 2010.

[9] 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.

[10] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *ICASSP*, 1992.

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

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

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

[14] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech Audio Process.*, vol. 6, no. 2, 1998.

[15] 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.