ACOUSTIC MODELING AND PARAMETER GENERATION USING RELEVANCE VECTOR MACHINES FOR SPEECH SYNTHESIS

Doo Hwa Hong, Joun Yeop Lee, and Nam Soo Kim

Department of Electrical and Computer Engineering and INMC
Seoul National University, Korea

ABSTRACT

In this paper, we propose a relevance vector machine (RVM) for modeling and generation of a speech feature sequence. In the conventional method, the mean parameter of the hidden Markov model (HMM) state can not consider temporal correlation among corresponding data frames. Since the RVM can be utilized to solve a nonlinear regression problem, we apply it to replace the model parameters of the state output distributions. In the proposed system, RVMs are employed to model the statistically representative process of the state or phone segment which is obtained from normalized training feature sequences by using the semi-parametric nonlinear regression method. We conducted comparative experiments for the proposed RVMs with conventional HMM. It is shown that the proposed state-level RVM-based method performed better than the conventional technique.

Index Terms— HMM, RVM, speech synthesis, acoustic modeling, parameter generation

1. INTRODUCTION

In the hidden Markov model (HMM)-based speech synthesis system, the speech parameter generation algorithm [1] is used to generate spectral and excitation parameters from HMMs to maximize their output probabilities under constraints between static and dynamic features. The statistical averaging in the modeling process improves robustness against data sparseness, and the use of dynamic-feature constraints in the synthesis process enables us to generate smooth feature sequences. However, the synthesized speech signal sounds evidently muffled compared to natural speech. This drawback comes from the reason that the generated feature trajectories are often over-smoothed, i.e., detailed characteristics of speech parameters are eliminated in the modeling stage and cannot be recovered in the synthesis stage.

This research was supported by the Mobile communication division, Samsung Electronics, co. Ltd. and by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP(Institute for Information & communications Technology Promotion)

To obtain better acoustic models, several techniques have been proposed such as trajectory HMM [2], autoregressive HMM [3], and the minimum generation error criterion (MGE) [4]. Although using the advanced acoustic models may increase modeling accuracy, it does not recover the over-smoothing problem in the synthesis algorithm. Post-filtering in the synthesis stage is the simplest way to compensate for over-smoothing. Using this method, the muffled sound of synthesized speech can be reduced. Speech parameter generation considering global variance (GV) [5] is one of the most widely used methods to alleviate the over-smoothing problem. Using GV, the dynamic range of generated trajectories is made close to those of natural ones. This method can be viewed as a statistical post-filtering technique to a certain extent. Although these generation techniques work better than the conventional, the resulting synthetic speech sounds more artificial. Recently, advanced machine learning techniques have been adopted to statistical parametric speech synthesis such as deep neural networks (DNN) [6, 7], and Gaussian process (GP) regression [8]. However, they require a large-size training data set, or spend expensive computational cost and a large amount of resources compared to the conventional method.

In this paper, we propose a feature sequence modeling and generation method using relevance vector machines (RVM) [9] as an alternative to recover the detailed trajectories of speech feature. Since the RVM can be employed to obtain a nonlinear regression, we apply it to replace the model parameters of the state output pdfs. In HMMs, it is assumed that observation sequences corresponding to the same state are quasi-stationary, and the mean parameter of the HMM state can not consider temporal correlation among corresponding data frames. We propose to model the representative trajectory of the state which is obtained from temporally normalized training feature sequences by using the semi-parametric nonlinear regression method. It is shown that the proposed method performs better than the conventional parameter generation algorithm.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. PARAMETER GENERATION USING HMM

In this section, we will give a brief review of the conventional parameter generation algorithm [1]. According to the decision tree, the sentence model is constructed by concatenating the states of corresponding clusters to the given context-dependent labels. After determining state sequence and duration as $\boldsymbol{q} = [q_1, q_2, \cdots, q_T]^\top$, a sequence of speech feature $\boldsymbol{C} = [\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \cdots, \boldsymbol{c}_T^\top]^\top$ is generated by maximizing the HMM likelihood:

$$\hat{\boldsymbol{C}} = \arg\max_{\boldsymbol{C}} P(\boldsymbol{O}|\hat{\boldsymbol{q}}, \boldsymbol{\lambda}) \tag{1}$$

where $\boldsymbol{\lambda}$ is the parameter set of the HMM, and $\boldsymbol{O} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \cdots, \boldsymbol{o}_T^\top]^\top$ is an observation vector sequence. Under a constraint on the relationship between static and dynamic features as given by

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C} \tag{2}$$

with the weighting matrix $\boldsymbol{W}$ for dynamic feature relation [1], the problem is reformulated as follows:

$$\hat{\boldsymbol{C}} = \arg\max_{\boldsymbol{C}} P(\boldsymbol{W}\boldsymbol{C}|\hat{\boldsymbol{q}}, \boldsymbol{\lambda}). \tag{3}$$

The synthetic speech feature sequence $\hat{\boldsymbol{C}}$ is obtained by solving a linear equation.

Clearly, a state is modeled by time-invariant statistical parameters. Although they involve temporal dynamics, the generated feature trajectories are not enough to represent sophisticated characteristics.

3. SPEECH SYNTHESIS USING RVM

We propose speech synthesis using RVM, in which the normalized segments are regarded as observations of a state, and the mean sequence of the state is obtained by nonlinear regression using RVMs. In this section, the modeling structure, the training procedure, and the parameter generation method of the proposed system are given.

3.1. RVM-based acoustic modeling

Typically using the RVM, the output observation o_n is approximated by

$$\begin{aligned} o_n &= f(\boldsymbol{x}_n) + \epsilon_n & (4)\\ &= \sum_{m=1}^{M} w_m \phi_m(\boldsymbol{x}_n) + \epsilon_n & (5) \end{aligned}$$

where $\boldsymbol{x}_n$ and ϵ_n denote the input vector and additive noise of the index n, respectively, and ϕ_m and w_m indicate the m-th basis function and its weight, respectively. The noise ϵ_n is assumed to be i.i.d. following a Gaussian distribution with zero mean and variance σ^2. Although this model is linear in the parameters, the function f can be highly flexible as the number of basis set may be very large and the basis function ϕ_m is defined by a kernel function:

$$\phi_m(\boldsymbol{x}) = \kappa(\boldsymbol{x}, \boldsymbol{x}_m). \tag{6}$$

Since most of the weights become zero after sparse Bayesian learning, the system takes the non-zero weights and corresponding basis functions only.

In order to apply the RVM on acoustic modeling, it should be extended to a multivariate form [10] in which the output observation is not scalar but in a vector space:

$$\begin{aligned} \boldsymbol{o}_n &= \sum_{m=1}^{M} \phi_m(\boldsymbol{x}_n)\boldsymbol{w}_m + \boldsymbol{\epsilon}_n & (7)\\ &= \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{x}_n) + \boldsymbol{\epsilon}_n & (8) \end{aligned}$$

where $\boldsymbol{o}_n$ and $\boldsymbol{\epsilon}_n$ represent the P-dimensional observation vector and the noise vector of the index n, respectively, $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_M]$ with the weight vector $\boldsymbol{w}_m$ of the m-th basis, and $\boldsymbol{\phi}(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \cdots, \phi_M(\boldsymbol{x})]^\top$. We assume that the probability distribution of the noise vector $\boldsymbol{\epsilon}$ is given by

$$p(\boldsymbol{\epsilon}_n) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{9}$$

with the covariance matrix $\boldsymbol{\Sigma}$. Then, we can define a pdf of observations corresponding a speech segment set by using the multivariate RVM, in which the output vector $\boldsymbol{o}_n$ and the input vector $\boldsymbol{x}_n$ are replaced by an observation $\boldsymbol{o}_t$ at the time index t and a relative time index $\tilde{t}$ associated with t, respectively:

$$\begin{aligned} p(\boldsymbol{o}_t) &= \mathcal{N}(\boldsymbol{\mu}(\tilde{t}), \boldsymbol{\Sigma}) & (10)\\ &= \mathcal{N}(\boldsymbol{W}\boldsymbol{\phi}(\tilde{t}), \boldsymbol{\Sigma}) & (11) \end{aligned}$$

where the mean vector $\boldsymbol{\mu}$ is a function of the relative time index $\tilde{t}$ which can be defined by the relative frame index of the observation $\boldsymbol{o}_t$ in the corresponding segment. Compared to the conventional HMM, this model can take account of continuous temporal characteristics of the given segment set.

For sparsity, the prior distribution with the precision vector $\boldsymbol{\alpha}_m$ of the weight vector $\boldsymbol{w}_m$ is assumed to be

$$\begin{aligned} p(\boldsymbol{w}_m|\boldsymbol{\alpha}_m) &= \mathcal{N}(\boldsymbol{0}, \mathrm{diag}(\boldsymbol{\alpha}_m)^{-1}) & (12)\\ &= \prod_{p=1}^{P} \mathcal{N}(0, \alpha_{mp}^{-1}) & (13) \end{aligned}$$

where α_{mp} denotes the precision of the prior with respect to the p-th element of the weight $\boldsymbol{w}_m$. When the noise covariance matrix $\boldsymbol{\Sigma}$ is diagonal, the elements of the weight $\boldsymbol{W}$ are independent to each others, and the model can be considered as augmentation of multiple RVMs. To select the relevant basis functions and estimate the corresponding weights, the sparse Bayesian learning technique [9, 11] is applied.

3.2. Training RVMs

In order to obtain RVMs, the outcomes of the conventional HMM training procedure are used. The HMM state tying results by decision tree-based clustering are employed to RVM training. The state segmentation labels of speech data signals are obtained by performing forced alignment with pre-trained HMMs. For each cluster, the length and the frame rate of the corresponding speech segments are normalized: the frame rate should be rapid enough to overcome the local overfitting problem. Then, we can train as many state-level multivariate RVMs as the number of clusters by the sparse Bayesian learning algorithm. If the clustering algorithm is applied not to state units but to phone models which consist of corresponding states, we can obtain phone-level RVMs by using the phone segmentation labels.

One of the alternatives in defining the basis functions is using the Gaussian RBF kernel, which can represent the local characteristics at the centered point by appropriate weighting. Due to normalization, unlike a typical regression cases, there are as many observations corresponding to a certain time index as the number of segments in their cluster at most. Thus, a kernel function set made by centering at each training observation could be redundant, and kernel functions centered respectively at all the possible time indices are enough without redundancy or missing of a basis.

Table 1. Comparison between HMM and RVM with respect to the number of parameters

HMM	RVM
$2PS$	$(2\tilde{M}+1)PS$

Table 1 shows the number of parameters of the conventional HMM system and the proposed RVM system, where $\tilde{M}$ and S indicate the average number of non-zero elements in a row of $\boldsymbol{W}$ and the number of clusters, respectively. The number of parameters to represent the mean sequence of an RVM is required to be $\tilde{M}$ times more than the conventional.

3.3. Parameter generation using RVMs

Parameter generation using RVMs is based on the conventional algorithm for HMMs by which the most probable speech parameter sequence is generated under a constrain of dynamic feature relation as (2). The difference is that, when using RVMs, the mean vector sequence of HMMs corresponding to state sequence is replaced by that of RVMs which is conducted by concatenating non-stationary, time-varying mean segments. The discontinuity between adjacent segments can be resolved by this algorithm.

Computational complexity of the conventional parameter generation algorithm is $O(P^3T^3)$ and additional computational complexity of parameter generation using RVMs is $O(PT)$. Therefore, the proposed algorithm does not increase computational load of parameter generation significantly compared to the conventional one.

4. EXPERIMENTS

In order to evaluate the performance of the proposed technique when applied to speech synthesis, we conducted several experiments on objective measurement and subjective listening test. All the speech data collected for speech synthesis were Korean spoken language.

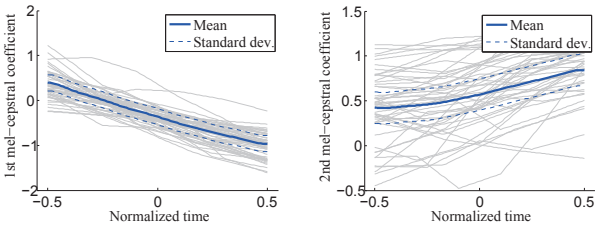
For the construction of the baseline speech synthesizer, a Korean speech database spoken by a male (HNC) and a female (YMK) speakers was applied. Each speaker provided 1,050 utterances of narrative speech data amounting to 131 and 128 minutes, respectively. A baseline narrative speech synthesizer was trained for each gender separately. Among the 1,050 utterances, we used 1,000 utterances for training the regression matrices and the remaining 50 utterances for evaluating the performance for each gender.

Each utterance was sampled at 16 kHz and a 20 ms Hamming window was applied with 5 ms frame shift for speech feature extraction. The acoustic features were obtained by STRAIGHT analysis [12]. As for the spectrum feature, a 25th-order mel-scaled cepstrum vector was extracted at each frame. By attaching the Δ- and $\Delta\Delta$-cepstra derived from the extracted mel-scaled cepstrum sequence, the spectrum feature could be represented by a 75-dimensional vector at each frame. We also extracted the fundamental frequency and 5-dimensional band aperiodicity from each frame as for the excitation feature. As the basic unit of speech synthesis, we applied quinphones followed by context-dependent reading-style text analysis. Each quinphone was modeled by a 5 state left-to-right structured HMM where the observation distribution at each state was given by a single Gaussian pdf with diagonal covariance matrix. All systems were implemented by modification of the HMM-based Speech Synthesis System (HTS) version 2.3 beta [13].

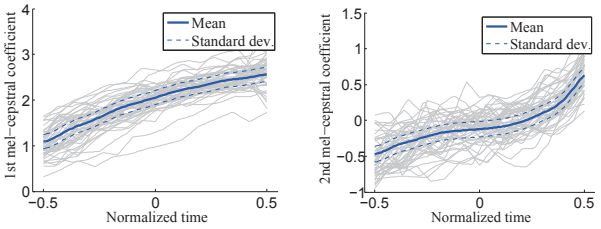
Fig. 1 shows examples of state-level and phone-level RVMs, respectively, for the first and second mel-cepstral coefficients. From the figures, we can see how the RVMs capture the mean trajectory by the proposed techniques from the given data.

4.1. Objective performance evaluation

We compared the outputs of three different algorithms: using conventional HMMs, state-level RVMs, and phone-level RVMs. For each speaker, three acoustic models using the comparative target algorithms were trained. To obtain phone-level clustering, a typical decision tree-based clustering method is modified to compute the likelihood of a phone-level HMM rather than state. Using RVMs, the number of clusters were set to be the same as using HMMs, which



(a) State-level RVM



b) Phone-level RVM

Fig. 1. Modeling examples of mel-cepstral coefficients by state-level RVMs and phone-level RVMs. The mean sequence and one standard deviation of the segment are shown by the bold line and the region between the dotted lines, respectively. Also, real speech segments are represented as shaded lines.

means that clustering for state-level RVMs is identical with one for HMMs and splitting on phone-level clustering was performed until the number of cluster became the same to the summation of the number of leaf nodes of each decision tree. In RVM training, the length of a segment was normalized to 1.0, frame rate for normalization was set to be 100 by simple resampling based on interpolation, and the kernel width of RBF kernel was determined by 0.1. To make continuous f0 sequence, the f0 values of the unvoiced region were filled during the normalization process.

Table 2. Objective measurement of comparative models (male)

method	Mel-cepstral distance	RMSE of lf0
HMM	4.552	0.1758
state-level RVM	**4.393**	**0.1587**
phone-level RVM	4.579	0.1775

Table 3. Objective measurement of comparative models (female)

method	Mel-cepstral distance	RMSE of lf0
HMM	4.899	**0.1106**
state-level RVM	**4.629**	0.1317
phone-level RVM	5.025	0.1188

Tables 2 and 3 show objective measurement of comparative models using different data set. In the overall, the mel-cepstral distance of state-level RVM was lower than the conventional, but phone-level RVM did not improve the performance. The root mean square error (RMSE) of log f0 became worse using the proposed method with the female voice. We consider that the reason of f0 degradation was due to the unobserved frame, i.e., unvoiced part. The performance of the proposed system related to f0 could be better if an accurate continuous pitch contour is available by any prediction algorithm.

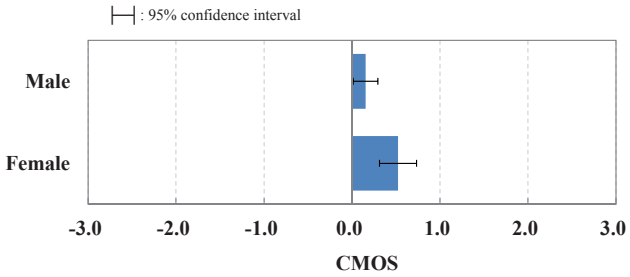


Fig. 2. Results of CMOS test: HMM vs. state-level RVM.

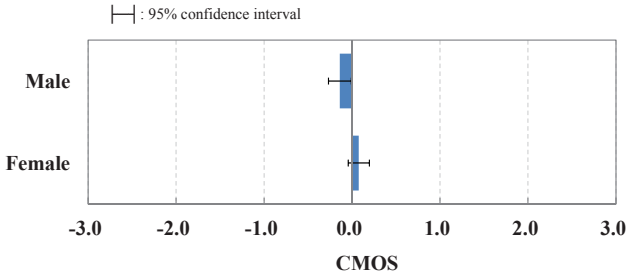


Fig. 3. Results of CMOS test: HMM vs. phone-level RVM.

4.2. Subjective performance evaluation

We also conducted a subjective listening test to compare the proposed algorithms to the conventional modeling technique, for which 14 listeners participated and 10 sentences were used. In the test, each listener was provided with speeches synthesized through different methods, and the speech quality was measured in terms of the comparative mean opinion score (CMOS) [14], where for each test a pair of two speech files were given and each subject provided his/her preference in speech quality in the range from -3 to 3 with a positive value indicating that the former shows a better quality than the latter, and vice versa.

Figs. 2 and 3 show the results of CMOS test evaluating by subjective scores of state-level and phone-level RVMs respectively compared to the conventional HMMs. We can find that, although phone-level RVMs did not outperform the conventional, the system of state-level RVMs generated better qual-

ity of synthetic speech than the conventional one.

4.3. Discussion

From the experimental results, the performance of phone-level RVMs was not better than conventional method. Main differences of phone-level RVM from state-level RVM are segmentation and clustering. We used the conventional clustering for state-level RVMs which was obtained by the widely used maximum description length criterion. Clustering for phone-level RVMs was done by using phone-level HMMs, but performance can be improved by applying proper clustering technique for time varying feature segments.

Advanced normalization is required to decrease the modeling error of both state-level and phone-level RVMs. The feature dynamics could be different according to the length of the segment. Linearly normalized frame indices do not consider this characteristics, therefore the modeling error could occur due to them. Since phone-level segments are definitely longer then state-level segments, this kind of normalization error affects phone-level structure more.

5. CONCLUSIONS

In this paper, we propose as a feature sequence modeling and generation method using RVMs as an alternative to recover the detailed trajectories of speech feature. Since the RVM can be employed to obtain a nonlinear regression, we apply it to replace the model parameters of the state output pdfs. We propose RVMs to model the representative trajectory of the state or phone segment which is obtained from temporally normalized training feature sequences by using the semi-parametric nonlinear regression method. From the experimental results, it is shown that the proposed state-level RVM-based method performs better than the conventional technique.

REFERENCES

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3, 2000, pp. 1315–1318.

[2] 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," *Comput. Speech Lang.*, vol. 21, no. 1, pp. 153–173, Jan. 2007.

[3] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 3, pp. 587–597, Mar. 2013.

[4] L. Qin, Y.-J. Wu, Z.-H. Ling, R.-H. Wang, and L.-R. Dai, "Minumum generation error linear regression based model adaptation for HMM-based speech synthesis," in *Proc. ICASSP*, 2008, pp. 3953–3956.

[5] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. 90, no. 5, pp. 816–824, May 2007.

[6] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2129–2139, Oct. 2013.

[7] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[8] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on gaussian process regression," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 173–183, Apr. 2014.

[9] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, Jun. 2001.

[10] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. Torr, and R. Cipolla, "Multivariate relevance vector machines for tracking," in *Proc. European Conference on Computer Vision (ECCV)*, 2006, pp. 383–390.

[11] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse bayesian models," in *Proc. AISTATS*, vol. 1, no. 3, 2003.

[12] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. ICASSP*, vol. 2, 1997, pp. 1303–1306.

[13] K. Tokuda, K. Oura, K. Hashimoto, S. Shiota, S. Takaki, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, and A. W. Black. (2014) The HMM-based speech synthesis system (HTS). [Online]. Available: http://hts.sp.nitech.ac.jp/

[14] V. Grancharov and W. B. Kleijn, "Speech quality assessment," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 83–100.