Recent Development of HMM-Based Expressive Speech Synthesis and Its Applications

Takashi Nose and Takao Kobayashi
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan
E-mail: takashi.nose, takao.kobayashi@ip.titech.ac.jp

***Abstract*—This paper describes the recent development of HMM-based expressive speech synthesis. Although the expressive speech includes a wide variety of expressions such as emotions, speaking styles, intention, attitude, emphasis, focus, and so on, we mainly refer to the speech synthesis techniques for emotions and speaking styles, which would be the most primary expressions in human speech communication. We describe five core techniques, i.e., style modeling, style adaptation, style interpolation, style control, and style estimation. In addition, we also give a brief overview of other applications to expressive speech synthesis and recognition.**

I. Introduction

Recent corpus-based speech synthesis with large-scale speech database dramatically improved the quality of reading-style synthetic speech. However, there is still a great difference between human voice and synthetic one. One of the primary factors is a lack of various expressions such as emotions and speaking styles appearing in the natural speech.

In this paper, we describe the recent development of expressive speech synthesis based on hidden Markov model (HMM) that can provide very flexible speech modeling and generation. The main focus is on the reproduction and control techniques of various emotional expressions and speaking styles, both of them we simply refer to as *styles*. There are five related core techniques, that is, style modeling, style adaptation, style interpolation, style control, and style estimation. Style modeling [1] is a technique for modeling and generating certain styles with a given sufficient amount of training data. Style adaptation [2] reduces the data preparation cost by using the model adaptation from the neutral-style model of the target speaker. Intermediate style expressions can be also generated using style interpolation [3] between two or more representative style models. Style control technique [4] is a more sophisticated technique to intuitively control the intensity of style expressivity appearing in the synthetic speech. Inversely, we can also estimate the style intensity of the actual speech using style estimation [5] by considering the inverse process of the style control.

We also briefly overview some other applications of HMM-based expressive speech synthesis, which include spontaneous and/or conversational speech synthesis [6–10], speaker characteristics emphasis [11], and rapid style adaptation for speech recognition [12].

II. Style Modeling

When a sufficient amount of training data, typically several tens minutes or more, is available for the target style of the target speaker, the most straightforward way for reproducing the style is to use style-dependent modeling. We can easily generate synthetic speech with the same manner of training and synthesis as that for the speaker-dependent modeling. If training data of multiple styles is available, we can also use style-mixed modeling [1] where the style type is explicitly taken into account as one of the contextual factors of the acoustic model. An advantage of the style-mixed modeling is that it needs fewer number of model parameters than the style-dependent modeling, which is desirable in embedded applications with a limited footprint. Experimental results of the subjective classification tests for synthetic speech with four different styles — neutral, rough, joyful, and sad — demonstrated that over 80% of speech samples were correctly perceived as the intended style in both style-dependent and style-mixed models [1].

Recently, the quality of synthetic speech has been improved [13] by introducing the high-quality vocoding system STRAIGHT [14] and the parameter generation algorithm considering global variance (GV) [15]. In the expressive speech synthesis, the GV-based parameter generation would be important for improving the perceptual reproducibility of prosodic features, especially fundamental frequency (F0), as well as the spectral feature. This is because variance of F0 values in each utterance plays a primary role for the expressivity in some emotions such as happiness and hot anger.

III. Style Adaptation

When the amount of the training data is insufficient, e.g., only a few minutes data, the naturalness and style expressivity of the synthetic speech would not be satisfactory. In some practical applications, speech synthesizer is required to express a variety of emotions and speaking styles, and it is not always acceptable to prepare a sufficient amount of training data of these styles. In [2], the neutral-style model trained using a sufficient amount of target speaker's data is used as the prior information of the acoustic model, and the initial model is converted to the target-style model by applying the speaker adaptation technique.

In the prosody adaptation, explicit modeling of the duration distribution using hidden semi-Markov model (HSMM) [16] is important. Experimental results have shown the significant

advantage of the HSMM-based adaptation using maximum likelihood linear regression (MLLR) against the HMM-based one [2]. It was also reported that decision trees are more appropriate than regression class trees in the adaptation process. This is because the decision trees can take into account supra-segmental features in the construction of parameter tying structures, and such features are important especially for the prosody modeling. Recent research revealed that simultaneous adaptation of speaker and style from average voice model [17] also works well to train representative style models in the style control [4] that will be described in Section V.

IV. Style Interpolation

In human speech communication, there often appear intermediate style expressions, e.g., a little sad, as well as styles with typical expressivity. Although the style adaptation is evidently a promising approach to generate synthetic speech with various types of styles at a low cost, it is unrealistic to record speech for several intermediate expressions of respective styles. This problem is solved by the style interpolation [3] that can generate arbitrary intermediate expressions of multiple styles by only changing the interpolation ratio. A typical application is to mix neutral (non-expressive) and a desired style. The interpolation is done at a model parameter level using the same manner as the speaker interpolation [18]. When representative style models are trained separately with different tying structures, the model parameters of corresponding Gaussian probability distribution functions (pdfs) are interpolated between sentence HMMs of two representative style models for given synthesis labels. In contrast, the parameters can be directly interpolated between two HMM sets in the case where the same model topology and parameter tying structure is used in the model training. For instance, we can achieve this by using shared decision tree context clustering (STC) [19].

V. Style Control

Recently, another approach to changing style expressivity of synthetic speech has been proposed [4, 20], where the degree or intensity of the expressivity is explicitly taken into account not only in the parameter generation process but also in the model training. This technique is based on the multiple-regression HMM [21, 22] or HSMM (MRHSMM) [4, 23], and the concept was named *style control* since the intensity of style expressivity can be controlled intuitively. By using multiple-regression model, we can simultaneously model multiple styles and their intensity of expressivity. Here, we overview the MRHSMM-based style control technique, in which spectral, F0, and duration parameters are simultaneously modeled and controlled.

A. Multiple-Regression Model

In the MRHSMM, mean parameters of each state of output and duration pdfs are expressed by a function of explanatory variables. More specifically, the mean parameter is assumed to be given as the multiple regression of a low-dimensional vector called a style vector. The components of the vector represent the intensity of style expressivity. We assume that mean parameters of output and state-duration pdfs at each state, $\boldsymbol{\mu}_i$ and m_i, respectively, are modeled using multiple regression of the style vector as

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi} \tag{1}$$

$$m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} \tag{2}$$

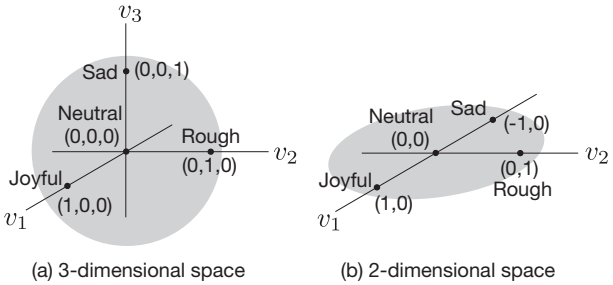
$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{3}$$

where $\boldsymbol{v}$ is the style vector in a low-dimensional style space. Component v_k of the style vector represents the intensity of a specific style of speech. In addition, $\boldsymbol{H}_{b_i}$ and $\boldsymbol{H}_{p_i}$ are $M \times (L+1)$- and $1 \times (L+1)$-dimensional regression matrices, respectively, and M is the dimensionality of $\boldsymbol{\mu}_i$. The pdfs at state i are thus expressed as

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \boldsymbol{\Sigma}_i) \tag{4}$$

$$p_i(d) = \mathcal{N}(d; \boldsymbol{H}_{p_i}\boldsymbol{\xi}, \sigma_i^2) \tag{5}$$

where $\boldsymbol{o}$ and $\boldsymbol{\Sigma}_i$ are the observation vector and covariance matrix of the output pdf, respectively, and d and σ_i^2 are the state duration and variance of state-duration pdf, respectively.



(a) 3-dimensional space (b) 2-dimensional space

Fig. 1. Example of style vectors for training data.

B. Model Training and Speech Synthesis

In the model training, first we train respective style-dependent models and construct decision trees with a common structure for parameter tying using STC. Consequently, mean parameters are included in the corresponding leaf nodes of all styles, and the initial mean of MRHSMM is calculated from these means and given style vectors using a least squares criterion. We label each training utterance with a corresponding style vector. An example of style vectors is shown in Fig. 1. As for the variance parameters, those of the style-independent model are used as initial values. Then, the parameters are re-estimated based on a maximum likelihood (ML) criterion using a similar manner to the case of standard HSMM training.

In the speech synthesis, for a given style vector, the mean parameters of each synthesis unit are modified using Eqs. (1) and (2). Speech signal is then generated in the same parameter generation algorithm as that for HSMM. By setting the style vector to a desired point in the style space, synthetic speech with a corresponding style intensity can be generated. Moreover, we can continuously change the style and expressivity by varying the style vector gradually along the state or phone transition.

The speaker adaptation has also been proposed for MRHSMM in the case of a small amount of training data [24, 25]. There are two different techniques. The first is based on MRHSMM-based MLLR that is applied to the initial MRHSMM trained using a sufficient amount of multiple styles data. The second is average-voice-based technique. The initial MRHSMM is obtained using style-dependent models trained using simultaneous adaptation of speaker and style, and is refined using MAP-like estimation. Experimental results have shown that both techniques can effectively reduce the required amount of speech data of the target speaker, and the average-voice-based technique is slightly better than the MLLR-based technique in terms of the speech naturalness [25].

C. Perceptual Expressivity Modeling

One of the problems of the conventional expressive speech synthesis with a style-dependent model is that the intensity of style expressivity appearing in synthetic speech completely depends on the training data. Although the style intensity can be changed by using MRHSMM-based style control, the dependency still remains unsolved when fixed style vectors, e.g., those in Fig. 1 are used during the model training. We can alleviate this problem by introducing the perceptual scoring of style intensity for each training utterances [26]. Experimental results have shown that we can control the style intensity more intuitively when the perceptual style intensity of training data is biased, i.e., weaker or stronger than expected. Another merit of introducing the perceptual expressivity into the model training is that we can train a single-style MRHSMM without any extra style speech. In the conventional MRHSMM training, we assumed that training data is available for at least two styles. This means that we always need to prepare additional neutral-style speech even if we want to control the style intensity of a single expressive style. The perceptual expressivity modeling relax this restriction, and we can choose a more suitable technique depending on the application.

VI. Style Estimation

In the MRHSMM-based style control, we estimate the speech parameters, i.e., spectral and F0 sequences, when the trained MRHSMM and a style vector are given. By considering the inverse process, we can also estimate the style vector for given expressive speech. This process is called *style estimation* [5, 27]. A similar approach has been studied in the acoustic-to-articulatory mapping [28, 29].

We consider a problem of estimating the style vector, $\boldsymbol{v}$, for an input observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ when a trained MRHSMM set is given. First, we conduct phoneme recognition for an input utterance because phonetic information is needed to create a sentence MRHSMM corresponding to the input speech by concatenating the given MRHSMMs. Then, the style vector is estimated for every input utterance using the sentence MRHSMM in ML sense. To derive the re-estimation formula of the EM algorithm, we rewrite Eqs. (1) and (2) in the following form.

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi} = \boldsymbol{h}_0^{(b_i)} + \boldsymbol{A}_{b_i}\boldsymbol{v} \tag{6}$$

$$m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} = \boldsymbol{h}_0^{(p_i)} + \boldsymbol{A}_{p_i}\boldsymbol{v} \tag{7}$$

where

$$\boldsymbol{H}_{b_i} = \left[\boldsymbol{h}_0^{(b_i)}, \cdots, \boldsymbol{h}_L^{(b_i)}\right] \tag{8}$$

$$\boldsymbol{A}_{b_i} = \left[\boldsymbol{h}_1^{(b_i)}, \cdots, \boldsymbol{h}_L^{(b_i)}\right] \tag{9}$$

$$\boldsymbol{H}_{p_i} = \left[h_0^{(p_i)}, \cdots, h_L^{(p_i)}\right] \tag{10}$$

$$\boldsymbol{A}_{p_i} = \left[h_1^{(p_i)}, \cdots, h_L^{(p_i)}\right] \tag{11}$$

$$\boldsymbol{v} = \left[v_1, \cdots, v_L\right]^\top. \tag{12}$$

The optimal style vector, $\boldsymbol{v}^*$, for the input observation sequence $\boldsymbol{O}$ is defined in the ML sense as

$$\boldsymbol{v}^* = \operatorname*{argmax}_{\boldsymbol{v}} P(\boldsymbol{O}|\lambda, \boldsymbol{v}). \tag{13}$$

The obtained values of style components give quantities as to how much each style affects the acoustic features of speech, including spectral and prosodic information compared to those of the training data in the ML sense. As a result, we can expect that the estimated values of the style components can be used to detect emotions and speaking styles expressed in speech.

Experimental results for speech with acted joyful and sad emotions have shown that the MRHSMM-based estimation gave higher correlation between perceptual and estimated scores than a linear regression analysis. The style estimation can be also applied to classification of emotions and speaking styles. In [5], the performance of speaking style classification was evaluated for spontaneous speech of reading and academic presentation styles, and the MRHSMM-based method consistently outperformed the SVM-based classification.

VII. Other Applications

In this section, we overview some ongoing studies for other applications of HMM-based expressive speech synthesis. An ultimate goal of the speech synthesis is to develop a system that can generate spontaneous conversational speech and to make the dialog more similar to human-human interaction. Although quality of the synthetic speech might be still not satisfactory, there have been several attempts: prosody modeling based on quantification theory type I [6], reading to spontaneous speech conversion [7], analysis by blending read and spontaneous speech [8], average-voice-based two-stage model adaptation [9], and context extension for spontaneous speech [10].

Since the basis of speech modeling by HMM is statistical expectation operation, it is well known that the generated speech suffers from the degradation of the speaker individuality and expressivity compared to the natural speech. To mitigate the degradation, a technique of controlling and emphasizing speaker characteristics was proposed [11]. This technique is based on MRHSMM with average voice model. The key idea came from the way of imitating voice by

professional impersonators since impersonators effectively utilize exaggeration of a target speaker's voice characteristics. Although this approach is fundamentally different from the conventional techniques of postfiltering [30] and GV-based parameter generation [15], we confirmed that the reproducibility of the speaker individuality was improved from several experimental results.

Recently, the idea of style control and style estimation was also integrated for rapid style adaptation in emotional speech recognition [12]. In the technique, the style vector is estimated for the input emotional utterance using MRHSMM. The acoustic models are then adapted to the input emotion using the estimated style vector. The intensity of its expressiveness is estimated utterance-by-utterance, The advantage of the technique against similar rapid model adaptation, e.g., eigenvoice-based one [31], is that we can obtain not only the linguistic information but also the paralinguistic information in the recognition process, which would be very important in human-computer interaction.

VIII. Conclusions

In this paper, we described the recent development of HMM-based expressive speech synthesis including the reproduction, adaptation, interpolation, control, and estimation of emotions and speaking styles. The HMM-based approach shows the capability of synthesizing speech that has more similar expressions to human than conventional concatenative synthesis with a realistic amount of speech data. We believe that the importance of synthesizing expressive speech would increase more and more in the near future.

References

[1] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 3, pp. 503–509, 2005.

[2] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 1092–1099, Mar. 2006.

[3] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[4] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, Sept. 2007.

[5] T. Nose and T. Kobayashi, "A technique for estimating intensity of emotional expressions and speaking styles in speech based on multiple-regression HSMM," *IEICE Trans. Inf. & Syst.*, vol. 93, no. 1, pp. 116–124, 2010.

[6] T. Akagawa, K. Iwano, and S. Furui, "Toward hidden Markov model-based spontaneous speech synthesis," *J. Acoust. Soc. America*, vol. 120, pp. 3037–3038, 2006.

[7] C.H. Lee, C.H. Wu, and J.C. Guo, "Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation," in *Proc. ICASSP 2010*, 2010, pp. 4826–4829.

[8] S. Andersson, J. Yamagishi, and R. Clark, "Utilising spontaneous conversational speech in HMM-Based speech synthesis," in *Proc. 7th ISCA workshop on speech synthesis (SSW7)*, 2010.

[9] T. Koriyama, T. Nose, and T. Kobayashi, "Conversational spontaneous speech synthesis using average voice model," in *Proc. INTERSPEECH 2010*, 2010, pp. 853–856.

[10] T. Koriyama, T. Nose, and T. Kobayashi, "On the use of extended context for HMM-based spontaneous conversational speech synthesis," in *Proc. INTERSPEECH 2010*, 2010, (to appear).

[11] T. Nose, J. Asada, and T. Kobayashi, "HMM-based speaker characteristics emphasis using average voice model," in *Proc. INTERSPEECH 2009*, 2009.

[12] Y. Ijima, T. Nose, M. Tachibana, and T. Kobayashi, "A rapid model adaptation technique for emotional speech recognition with style estimation based on multiple-regression HMM," *IEICE Trans. Inf. & Syst.*, vol. 93, no. 1, pp. 107–115, 2010.

[13] X. Gonzalvo, P. Taylor, C. Monzo, I. Iriondo, and J. Socoró, "High quality emotional HMM-based synthesis in Spanish," *Advances in Nonlinear Speech Processing*, pp. 26–34, 2010.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "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, pp. 187–207, Sept. 1999.

[15] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[16] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[17] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[18] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *J. Acoust. Soc. Jpn. (E)*, vol. 21, pp. 199–206, Apr. 2000.

[19] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Inf. & Syst.*, vol. E86-D, no. 3, pp. 534–542, Mar. 2003.

[20] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for speech synthesis using multiple regression HSMM," in *Proc. INTERSPEECH 2006-ICSLP*, Sept. 2006, pp. 1324–1327.

[21] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," in *Proc. ICASSP 2001*, May 2001, pp. 513–516.

[22] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," in *Proc. INTERSPEECH 2004-ICSLP*, Oct. 2004, pp. 1437–1440.

[23] N. Niwase, J. Yamagishi, and T. Kobayashi, "Human walking motion synthesis with desired pace and stride length based on HSMM," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2492–2499, Nov. 2005.

[24] T. Nose, Y. Kato, and T. Kobayashi, "A speaker adaptation technique for MRHSMM-based style control of synthetic speech," in *Proc. ICASSP 2007*, Apr. 2007, pp. 833–836.

[25] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," *IEICE Trans. Inf. & Syst.*, vol. E92-D, no. 3, pp. 489–497, Mar. 2009.

[26] T. Nose and T. Kobayashi, "A perceptual expressivity modeling technique for speech synthesis based on multiple-regression HSMM," in *Proc. INTERSPEECH 2011*, 2011, (to appear).

[27] T. Nose, Y. Kato, and T. Kobayashi, "Style estimation of speech based on multiple regression hidden semi-Markov model," in *Proc. INTERSPEECH 2007*, Aug. 2007, pp. 2285–2288.

[28] S. Hiroya and T. Mochida, "Multi-speaker articulatory trajectory formation based on speaker-independent articulatory hmms," *Speech communication*, vol. 48, no. 12, pp. 1677–1690, 2006.

[29] Z.H. Ling, K. Richmond, and J. Yamagishi, "HMM-based text-to-articulatory-movement prediction and analysis of critical articulators," *Speech communication*, vol. 52, no. 10, pp. 834–846, 2010.

[30] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. EUROSPEECH 2001*, 2001, vol. 3, pp. 2263–2266.

[31] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, Sept. 2000.