Chapter 14
Prosody Control and Variation Enhancement Techniques for HMM-Based Expressive Speech Synthesis

Takao Kobayashi

Abstract Natural speech has diverse forms of expressiveness including emotions, speaking styles, and voice characteristics. Moreover, the expressivity changes depending on many factors at the phrase level, such as the speaker's temporal emotional state, focus, feelings, and intention. Thus taking into account such variations in modeling of speech synthesis units is crucial to generating natural-sounding expressive speech. In this context, two approaches to HMM-based expressive speech synthesis are described: a technique for intuitively controlling style expressivity appearing in synthetic speech by incorporating subjective intensity scores in the model training and a technique for enhancing prosodic variations of synthetic speech using a newly defined phrase-level context for HMM-based speech synthesis and its unsupervised annotation for training data consisting of expressive speech.

14.1 Introduction

Synthesizing natural-sounding speech with diverse forms of expressiveness including emotions, speaking styles, voice characteristics, focuses, and emphases, is a key to achieving more natural human–computer interactions. In this regard, the promising results of HMM-based speech synthesis (Erickson 2005; Nose and Kobayashi 2011a; Schröder 2009; Zen et al. 2009) have recently led to a number of attempts at applying this approach to expressive speech synthesis using speech corpora recorded under realistic conditions (Eyben et al. 2012; Koriyama et al. 2011; Maeno et al. 2011; Obin et al. 2011a, b).

When the target expressiveness consistently appears in every utterance in a corpus, it can be modeled properly and synthetic speech can be generated with expressiveness similar to that of the corpus (Yamagishi et al. 2003). Moreover, a style interpolation

T. Kobayashi (✉)
Department of Information Processing, Tokyo Institute of Technology,
Yokohama 226-8502, Japan
e-mail: takao.kobayashi@ip.titech.ac.jp

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,*Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_14

technique (Tachibana et al. 2005) or a multiple regression HMM-based speech synthesis approach (Miyanaga et al. 2004; Nose et al. 2006), called the style control technique, can be used to strengthen or weaken the intensity of target expressivity of the synthetic speech. However, the prosodic variation of real expressive speech is generally much larger than that of simulated speech and is not consistent. In other words, expressivity changes depending on many factors, such as the speaker's temporal emotional state, focus, feelings, and intention at the phrase level. In addition, its intensity is not constant within a sentence (Cowie and Cornelius 2003; Doukhan et al. 2011). Thus, incorporating such variations into the modeling of the HMM-based speech synthesis units will be crucial to generating natural-sounding expressive speech.

This paper addresses two issues related to modeling and synthesizing expressive speech in the HMM-based framework. In the following, the expressions of emotions, speaking styles, prominences, etc., which may appear singly or simultaneously, will be referred to simply as *styles* (Miyanaga et al. 2004; Yamagishi et al. 2003).

The first issue is that the original style control technique did not take account of the style intensities of the respective utterances during the model training phase. The style intensity is a subjective score of a certain style expressivity given by listeners. In Miyanaga et al. (2004) and Nose et al. (2006), the style intensity, represented by a style vector, was assumed to be a fixed value regardless of the variations in style intensity appearing in respective training samples. This assumption may result in synthetic speech with consistently lower expressivity than what the user expects, if the average intensity of the target style in the training data is much lower than what the user expects. This problem can be alleviated by using subjective style intensity scores of respective utterances and taking them into account in the model training of the style control technique. Note that this technique is similar to other adaptive training approaches (Anastasakos et al. 1996; Gales 2000; Yu et al. 2001).

The second issue is an inevitable impediment to natural-sounding expressive speech synthesis. In the HMM-based speech synthesis approach, it is possible to reproduce locally appearing styles or prosodic variations only if the corpus for model training has been properly annotated and appropriate context labels have been given (Koriyama et al. 2011; Maeno et al. 2011; Yu et al. 2010). While manually annotating a corpus might work for this purpose, it is time-consuming and tends to be expensive and impractical on a large corpus. In addition, even if the cost is acceptable, another difficulty arises in that consistent annotation of styles, such as emotional expressions, is not always possible. Alternatively, unsupervised clustering of styles for expressive speech synthesis has been examined as a way of avoiding manual annotation and categorization of styles (Eyben et al. 2012; Székely et al. 2011). However, since there is not always an explicit expressiveness in the resultant clusters, users may have difficulty choosing an appropriate cluster to output the desired expressive speech in the speech synthesis phase.

In this context, there is an alternative technique for enhancing the prosodic variations of synthetic expressive speech without requiring the manual annotation of style information in the model training. An additional context, called the phrase-level F0

context, can be introduced, and it is defined by the average difference in prosodic features between the original and synthetic speech of the training sentences. The advantage of using this newly defined context is that proper annotation can be done fully automatically without any heuristics. Moreover, the obtained context has an intuitive prosodic meaning of higher or lower pitch at a certain accent phrase.

14.2 Prosody Control Based on Style Control Technique

14.2.1 Style Control Using Multiple-Regression HMM

Style control is an approach that enables us to intuitively change the style expressivity, i.e., emotional expressions and/or speaking styles and their intensities appearing in synthetic speech (Nose and Kobayashi 2011a). The style control technique is based on the idea of representing the variations of styles by using multiple-regression HMMs (MRHMMs) (Miyanaga et al. 2004) or multiple-regression hidden semi-Markov models (MRHSMMs) (Nose et al. 2006).

In the case of using MRHSMM, probability density functions (pdfs) for the output of the states, i.e., spectral and pitch features, and durations of the states are expressed using Gaussian pdfs with mean parameters that are assumed to be a multiple regression of a low-dimensional vector $\mathbf{s}$, i.e.,

$$\boldsymbol{\mu}_i = \mathbf{H}_{b_i} \boldsymbol{\xi} \tag{14.1}$$

$$m_i = \mathbf{H}_{p_i} \boldsymbol{\xi} \tag{14.2}$$

$$\boldsymbol{\xi} = [1, s_1, s_2, \ldots, s_L]^\top = [1, \mathbf{s}^\top]^\top \tag{14.3}$$

where $\boldsymbol{\mu}_i$ and m_i are the mean parameters of the state-output and state-duration Gaussian pdfs at state i, respectively, and $\mathbf{s} = [1, s_1, s_2, \ldots, s_L]^\top$ is a style vector in a low-dimensional style space. As shown in Fig. 14.1, each axis of the style space represents a certain style, such as joyful, sad, appealing, or storytelling, and each component of the style vector represents the expressivity intensity of a specific style. In addition, $\mathbf{H}_{b_i}$ and $\mathbf{H}_{p_i}$ are respectively $M \times (L+1)$- and $1 \times (L+1)$-dimensional regression matrices, where M is the dimensionality of the mean vector $\boldsymbol{\mu}_i$. These regression matrices are determined with maximum likelihood (ML) estimation.

In the speech synthesis phase, for an arbitrarily given style vector $\mathbf{s}$, the mean parameters of each synthesis unit are determined using (14.1) and (14.2). The speech signal is generated in the same manner as in ordinary HMM-based speech synthesis. Synthetic speech with a corresponding style intensity can be generated by setting the style vector to a desired point in the style space. Moreover, we can continuously and intuitively change the style and expressivity intensity by varying the style vector gradually along the state or phone transition.

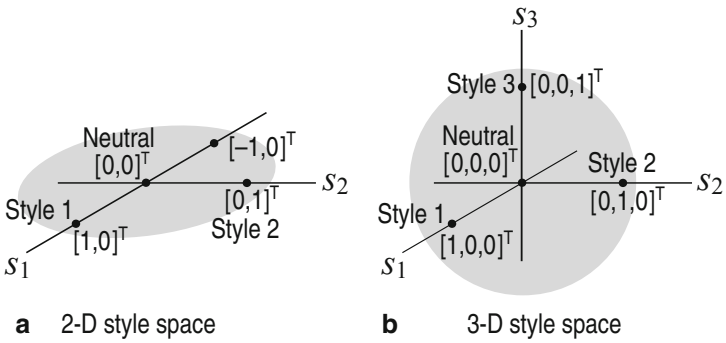

Fig. 14.1 Example of style spaces and style vectors (indicated by *dots*) for training data. **a** 2-D style space. **b** 3-D style space

14.2.2 Model Training with Perceptual Style Expressivity

The model training of MRHSMM requires style vectors for respective training speech samples. A simple choice is using a fixed style vector for each style as is done in Nose et al. (2006). As shown in Fig. 14.1, one specific vector is set as the style vector for each style independently of the intensity of expressivity appearing in respective speech samples, and it is used during the model training. Although it has been shown that this works well for expressivity control, it may cause a problem wherein we cannot always obtain synthetic speech with the desired style expressivity when the perceptual style intensity of the training data is biased, i.e., weaker or stronger than expected.

An alternative way of setting the style vector is to add the subjective style intensities into the model training (Nose and Kobayashi 2011b; Nose and Kobayashi 2013). Specifically, the style intensities perceived for the respective training utterances are quantified in listening tests and the obtained subjective scores are used as the style vectors in the model training. This leads to an additional advantage of requiring only the speech samples of the target style in the modeling of the style and intensity. In contrast, the technique using fixed style vectors (Nose et al. 2006) needs two or more styles for the model training.

14.2.3 Example of Style Expressivity Control

Figure 14.2 shows an evaluation of the controllability of style expressivity using the proposed style control technique (Nose and Kobayashi 2011b). The evaluation used emotional speech data uttered by a professional female narrator. The speech samples consisted of 503 phonetically balanced sentences with joyful and sad styles. Nine participants listened to each utterance in random order and rated the intensity of the

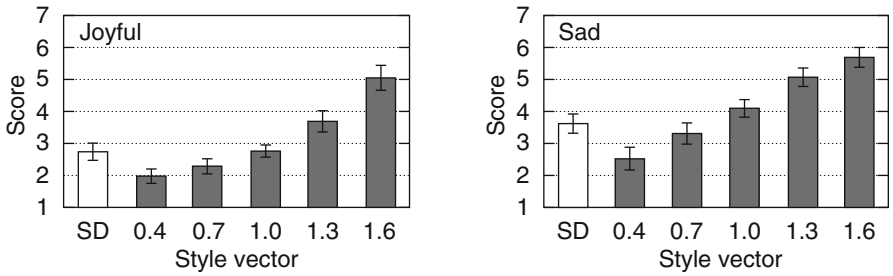

Fig. 14.2 Average style intensities of synthetic joyful and sad style speech for different style vectors. *SD* represents the result for the case of *style-dependent* HSMM trained without using style intensity information

style expressivity as "1.5" for strong, "1.0" for standard, "0.5" for weak, and "0" for not perceivable. The average score of the nine participants was taken to be the style intensity score. After the subjective scoring, 40 sentences with a perceptual score of 1.0 were chosen as the test data, since this score was expected to represent the standard expressivity of the target styles. The other 463 sentences were used as training data. Speech signals were sampled at a rate of 16 kHz, and STRAIGHT analysis (Kawahara et al. 1999) with a 5-ms shift was used to extract the spectral features. A five-state left-to-right model topology was used for the MRHSMM. The other conditions are detailed in Nose and Kobayashi (2011b).

In the parameter generation, the style vector was changed from 0.4 to 1.6 with an increment of 0.3. Seven participants listened to a speech sample chosen randomly from test sentences and rated its style expressivity by comparing it to that of a reference speech sample whose perceptual scores were 1.0. The reference samples were vocoded speech in the target style. The rating was a seven-point scale with "1" for very weak, "4" for similar to that of the reference, and "7" for very strong. The figure plots average subjective scores against given style vectors with a confidence interval of 95 %. It can be seen from the figure that the style control technique enables us to control the style intensity in accordance with the value of the style vector. Further evaluation results and detailed discussions can be found in Nose and Kobayashi (2011b, 2013).

14.3 Prosodic Variation Enhancement Using Phrase-Level F0 Context

14.3.1 Phrase-Level Prosodic Contexts

The global and local characteristics of the prosodic features of expressive speech often differ from those of neutral or reading style speech. Global prosodic features are generally well modeled using conventional statistical approaches by simply adding

a global context (Yamagishi et al. 2003). In contrast, it is not easy to model local variations, typically phrase-level ones, because they are rather diverse, depending on a variety of factors such as speakers, styles, and other paralinguistic contexts. Although we can reflect such variations in synthetic speech by using manually annotated speech samples for model training, manual annotation is time-consuming and impractical for large speech corpora. Moreover, consistent annotation is especially difficult for expressive speech. As a way of solving this problem, additional contexts for HMM-based speech synthesis, called phrase-level prosodic contexts, are defined, and they enables us to annotate training data automatically and enrich the prosodic variations of synthetic speech (Maeno et al. 2013, 2014). While the phrase-level prosodic contexts are defined for F0, duration, and power features, the following deals with only the phrase-level F0 context.

Consider a context labeling process for given training data. Let us assume that the ordinary context labels including accent phrase boundary information are available for the HMM-based speech synthesis and that conventional context-dependent HMMs using those labels are trained in advance. By using the F0s extracted from the training speech sample and the synthetic speech sample generated from the obtained HMMs, the average log F0 difference at each accent phrase is expressed as

$$d = f_o - f_s \tag{14.4}$$

where f_o and f_s are average log F0 values of the original and synthetic speech within each accent phrase. Then, for a prescribed positive value α, the phrase-level F0 context is defined as

- "Low" for $d < -\alpha$
- "Neutral" for $-\alpha \leq d < \alpha$
- "High" for $d \geq \alpha$.

Finally, every accent phrase is labeled with the above phrase-level F0 context associated with the value of d.

14.3.2 Automatic Labeling of Phrase-Level F0 Context

In the phrase-level F0 context labeling, an appropriate classification threshold α should be determined before labeling. For this purpose, an optimal α that minimizes the F0 error can be chosen using a simple grid search approach. The algorithm for obtaining the optimal threshold for the phrase-level F0 context is summarized as follows:

1. Specify a value of α between the possible lower and upper bounds, α_s and α_e.
2. Perform phrase-level F0 context labeling with the specified α for all training samples.
3. Train new HMMs using the context labels including the obtained phrase-level F0 context, and generate F0 sequences for all training samples using the newly trained HMMs and the context labels.

4. Calculate the root mean square (RMS) error E_α of log F0s between all the original and newly synthesized speech samples.
5. Specify a new value of α ($\alpha_s \leq \alpha \leq \alpha_e$) different from the value used in the previous iteration and repeat steps 2 to 4.
6. Finally, choose the optimal threshold α^* that minimizes E_α as

$$\alpha^* = \arg\min_{\alpha} E_\alpha. \tag{14.5}$$

A simple way of performing a grid search is specifying $\alpha = \alpha_s$ in the first iteration and in the nth iteration

$$\alpha = \alpha_s + (n-1)\Delta\alpha \quad (\alpha \leq \alpha_e) \tag{14.6}$$

where α is the increment in each iteration. In addition, the algorithm may use another F0 error, e.g., the maximum log F0 error, instead of the RMS error.

Note that there is a similar approach to prosodic tagging that uses the difference between the prosodic features of the generated and original speech (Suni et al. 2012; Vainio et al. 2005). However, this approach requires manually labeled training data or empirically determined weights. In contrast, the algorithm described here has the advantage that the classification threshold for the phrase-level F0 context is automatically optimized depending on the target expressive corpus without using any heuristics.

14.3.3 Model Training Example with Optimum Prosodic Labeling

The proposed phrase-level F0 context labeling was applied to two types of Japanese expressive speech data: appealing speech in sales talk and fairytale speech in storytelling. Speech samples were recorded under realistic circumstances in which no speech styles were specified to the speakers and only the target domain (situation) was made known to them (Nakajima et al. 2010). Appealing style samples were uttered by a female professional narrator, whereas fairytale style samples were uttered by a male professional narrator. The amounts of speech data of appealing and fairytale speech were approximately 33 and 52 min, respectively. Speech signals were sampled at a rate of 16 kHz, and STRAIGHT analysis with a frame shift of 5 ms was applied. A five-state left-to-right HSMM with a no-skip topology was used for the modeling of acoustic features including F0. The optimal threshold for the phrase-level F0 context was determined on the basis of fourfold crossvalidation for each style and by setting $\alpha_s = 0$, $\alpha_e = 0.3$ (519 cent), which is the maximum value of d for all accent phrases of the training data, and $\Delta\alpha = 0.01$ (17 cent). In addition, since the prosodic variation could affect the current phrase as well as the adjacent phrases, the phrase-level F0 context labels for the preceding and succeeding accent phrases as well as the current one were taken into account in the context clustering process. The other conditions are described in Maeno et al. (2014).

Table 14.1 RMS log F0 error (cent) between original and synthetic speech for test data

Style	BASELINE	PROPOSED
Appealing	254	**201**
Fairy tale	359	**273**

Table 14.1 compares RMS log F0 errors between the original and synthetic speech for test samples that were not included in the training samples. The entries for BASELINE represent the results for the model without using the phrase-level F0 context, whereas those for PROPOSED represent the case of using the phrase-level F0 context with the optimum labeling. It can be seen that the F0 distortion significantly decreased as a result of using the proposed phrase-level F0 context. Fig. 14.3 shows the F0 contours of the original and synthetic speech for a fairytale style utterance. The figure shows that the F0 reproducibility is much improved by using phrase-level F0 context labels in the HMM-based speech synthesis.

14.3.4 Prosodic Variation Enhancement for Expressive Speech Synthesis

Although the proposed phrase-level F0 context labeling solves the annotation problems for the training data and improves the F0 reproducibility of expressive synthetic speech, the technique faces a problem when it is to be applied to text-to-speech (TTS) systems. That is, the proposed phrase-level F0 context labels are not obtained from the input text automatically, because they are determined using real utterances of the target expressive speech, which are not available for arbitrarily input text.

Instead, let us consider the usage of the proposed technique when the phrase-level F0 context information for the input text is unknown. A typical example of such usage is when users want to create synthetic speech samples of voice actresses/actors with higher prosodic variability for audiobook and movie content. In this case, users first synthesize speech for the target sentence using F0 context labels whose values are all

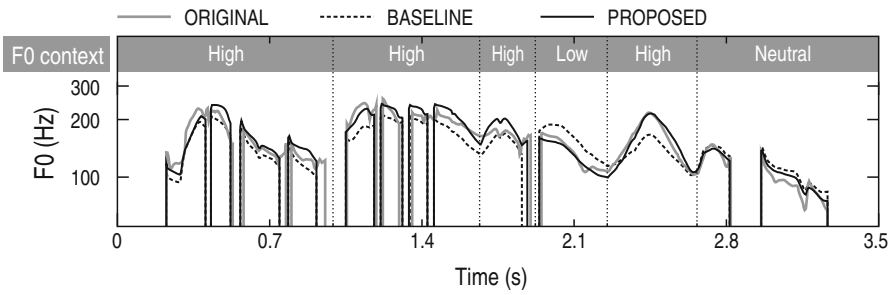

Fig. 14.3 Example of F0 contours generated with the phrase-level F0 context for fairly tale style speech sample. *BASELINE* shows the result without using the phrase-level F0 context

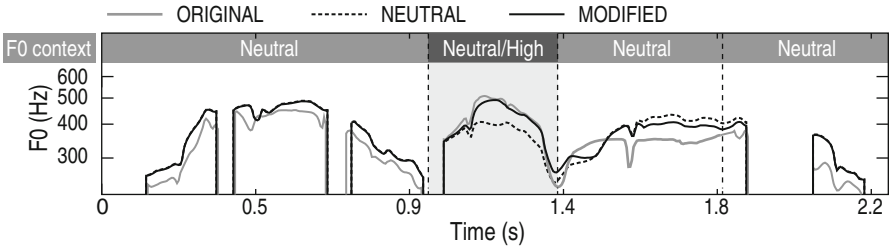

Fig. 14.4 Example of F0 contours before and after changing the F0 context for an appealing style speech sample. The F0 context in the accent phrase indicated by the shaded region was modified from "Neutral" to "High"

set to "Neutral." Synthetic speech obtained under this condition would sometimes result in poor prosodic variability compared with the real expressive speech. Next, the users listen to the synthetic speech sample and modify the F0 context of a certain accent phrase to "High" or "Low" if they want to enhance the F0 variation in that phrase. Then they synthesize speech again with the modified context labels and check the resultant synthetic speech. By repeating this procedure, users can obtain better synthetic speech in terms of F0 variations.

Figure 14.4 illustrates an example of this F0 variation enhancement process. Conditions for the phrase-level F0 context labeling and model training are the same as described in Sect. 14.3.3. A user first listened to an appealing style synthetic speech sample generated with the phrase-level F0 context labels being set all "Neutral" (the F0 contour is denoted as NEUTRAL in the figure). Since the user felt a lack of prominence in the second accent phrase, its phrase-level context label was changed to "High" and the speech sample was resynthesized. The figure shows that the resultant F0 contour denoted by MODIFIED became closer to that of the real speech denoted by ORIGINAL. Further experimental results and detailed discussions are provided in Maeno et al. (2013, 2014).

14.4 Conclusions

Techniques of prosody control and prosodic variation enhancement were discussed for HMM-based expressed speech synthesis. First, a brief review of the prosody control technique based on multiple regression HSMMs (MRHSMM) was given. Then subjective style intensities was incorporated into the technique to achieve more intuitive control of the styles. The use of subjective style intensities in the training of MRHSMMs normalizes the variation of style intensities appearing in the training data, and this results in an intensity that users would deem normal for the style of speech (Nose and Kobayashi 2013). Moreover, the training of the MRHSMMs can be done using only data for a single style by introducing style intensity scores for the respective training samples.

Next, an unsupervised labeling technique for phrase-level prosodic variations was described. The technique can be used to enhance the prosodic variation of synthetic expressive speech. A new prosodic context, the phrase-level F0 context, for HMM-based speech synthesis was defined and a fully automatic labeling algorithm for the newly defined context was described. Experiments on the prosodic context labeling revealed that the variations in the F0 feature appearing in the training samples were effectively captured with the proposed technique. Although phrase-level F0 context labels are unknown for arbitrary input text in practical TTS situations, the technique enables users to intuitively enhance the prosodic characteristics of a target accent phrase by manually changing the proposed context label from "Neutral" to "High" or "Low."

Acknowledgement The author would like to thank T. Nose, Y. Maeno, and T. Koriyama for their contributions to this study at Tokyo Tech. He would also like to thank O. Yoshioka, H. Mizuno, H. Nakajima, and Y. Ijima for their helpful discussions and providing expressive speech materials.

References

Anastasakos, T., J. McDonough, R. Schwartz, and J. Makhoul. 1996. A compact model for speaker adaptive training. Proceedings of ICSLP, 1137–1140.

Cowie, R., and R. R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication* 40 (1–2): 5–32.

Doukhan, D., A. Rilliard, S. Rosset, M. Adda-Decker, and C. d'Alessandro. 2011. Prosodic analysis of a corpus of tales. Proceedings of INTERSPEECH, 3129–3132.

Erickson, D. 2005. Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology* 26 (4): 317–325.

Eyben, F., S. Buchholz, N. Braunschweiler, J. Latore, V. Wan, M. J. F. Gales, and K. Knill. 2012. Unsupervised clustering of emotion and voice styles for expressive TTS. Proceedings of ICASSP, pp. 4009–4012.

Gales, M. J. F. 2000. Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing* 8 (4): 417–428.

Kawahara, H., I. Masuda-Katsuse, and A. de Cheveigne. 1999. 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* 27 (3–4): 187–207.

Koriyama, T., T. Nose, and T. Kobayashi. 2011. On the use of extended context for HMM-based spontaneous conversational speech synthesis. Proceedings of INTERSPEECH, 2657–2660.

Maeno, Y., T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka. 2011. HMM-based emphatic speech synthesis using unsupervised context labeling. Proceedings of INTERSPEECH, 1849–1852.

Maeno, Y., T. Nose, T. Kobayashi, T. Koriyama, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka. 2013. HMM-based expressive speech synthesis based on phrase-level F0 context labeling. Proceedings of ICASSP, pp. 7859–7863.

Maeno, Y., T. Nose, T. Kobayashi, T. Koriyama, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka. 2014. Prosodic variation enhancement using unsupervised context labeling for HMM-based expressive speech synthesis. *Speech Communication* 57:144–154.

Miyanaga, K., T. Masuko, and T. Kobayashi. 2004. A style control technique for HMM-based speech synthesis. Proceedings of INTERSPEECH-ICSLP, 1437–1440.

Nakajima, H., N. Miyazaki, A. Yoshida, T. Nakamura, and H. Mizuno. 2010. Creation and analysis of a Japanese speaking style parallel database for expressive speech synthesis. http://desceco.org/O-COCOSDA2010/proceedings/paper_30.pdf. Accessed 6 Dec 2014.

Nose, T., and T. Kobayashi. 2011a. Recent development of HMM-based expressive speech synthesis and its applications. Proceedings of APSIPA ASC. http://www.apsipa.org/proceedings_2011/pdf/APSIPA189.pdf. Accessed 6 Dec 2014.

Nose, T., and T. Kobayashi. 2011b. A perceptual expressivity modeling technique for speech synthesis based on multiple-regression HSMM. Proceedings of INTERSPEECH, 109–112.

Nose, T., and T. Kobayashi. 2013. An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. *Speech Communication* 55 (2): 347–357.

Nose, T., J. Yamagishi, and T. Kobayashi. 2006. A style control technique for speech synthesis using multiple-regression HSMM. Proceedings of INTERSPEECH-ICSLP, 1324–1327.

Obin, N., A. Lacheret, and X. Rodet. 2011a. Stylization and trajectory modelling of short and long term speech prosody variations. Proceedings of INTERSPEECH, 2029–2032.

Obin, N., P. Lanchantin, A. Lacheret, and X. Rodet. 2011b. Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation. Proceedings of INTERSPEECH, 2785–2788.

Schröder, M. 2009. Expressive speech synthesis: Past, present, and possible futures. In: *Affective information processing*, ed. J. H. Tao and T. N. Tan, 111–126. London: Springer.

Suni, A., T. Raitio, M. Vainio, and P. Alku. 2012. The GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach. Proceedings of Blizzard Challenge Workshop. http://festvox.org/blizzard/bc2012/HELSINKI_Blizzard2012.pdf. Accessed 6 Dec 2014.

Székely, E., J. Cabral, P. Cahill, and J. Carson-Berndsen. 2011. Clustering expressive speech styles in audiobooks using glottal source parameters. Proceedings of INTERSPEECH, 2409–2412.

Tachibana, M., J. Yamagishi, T. Masuko, and T. Kobayashi. 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transactions on Information and Systems* E88-D (11): 2484–2491.

Vainio, M., A. Suni, and P. Sirjola. 2005. Accent and prominence in Finnish speech synthesis. Proceedings of International Conference on Speech and Computer (SPECOM), 309–312.

Yamagishi, J., K. Onishi, T. Masuko, and T. Kobayashi. 2003. Modeling of various speaking styles and emotions for HMM-based speech synthesis. Proceedings of INTERSPEECH, 2461–2464.

Yu, K., H. Zen, F. Mairesse, and S. Young. 2001. Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis. *Speech Communication* 53 (6): 914–923.

Yu, K., F. Mairesse, and S. Young. 2010. Word-level emphasis modelling in HMM-based speech synthesis. Proceedings of ICASSP, 4238–4241.

Zen, H., K. Tokuda, and A. Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51 (11): 1039–1064.