INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Voice quality control using perceptual expressions for statistical parametric speech synthesis based on cluster adaptive training

Yamato Ohtani, Koichiro Mori and Masahiro Morita

Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation
yamato.ohtani@toshiba.co.jp

Abstract

This paper describes novel voice quality control of synthetic speech using cluster adaptive training (CAT). In this method, we model voice quality factors labeled with perceptual expressions such as "Gender," "Age" and "Brightness." In advance, we obtain the intensity scores of the perceptual expressions by conducting a listening test, which evaluates differences of voice qualities between synthetic speech of average voice and that of the target. Then we build perceptual expression (PE) clusters that we call PE models (PEM) under the conditions that the average voice model is used as the bias cluster and the PE intensity scores are employed as the CAT weights. In synthesis, we can generate controlled synthetic speech by the linear combination of PEMs and the existing speaker's model. Subjective results demonstrate that the proposed method can control the voice qualities with PEs in many cases and the target synthetic speech modified by PEMs achieves comparatively good speech quality.

Index Terms: speech synthesis, hidden Markov model, voice quality, perceptual expression, cluster adaptive training

1. Introduction

One of the desired functionalities for speech synthesis is manual manipulation of voice characteristics such as emotional expression, speaking style and voice quality. In terms of voice quality control, it is necessary to create a new speaker's voice and edit the voice quality of the existing speaker for many speech applications such as dialogue systems and broadcasting systems.

Statistical parametric speech synthesis has potential for flexible speech generation with various voice characteristics. Speech synthesis based on a hidden semi-Markov model (HSMM) [1, 2] has achieved the manual control of voice characteristics by multiple regression analysis [3, 4]. Multiple regression HSMM (MRHSMM)-based speech synthesis systems [5, 6] represent a mean vector of the distribution by using the regression matrix and a weight vector. MRHSMM-based speech synthesis for voice quality control uses perceptual expressions (PEs) composed of pairs of words such as "Smoothness (Nonsmooth–Smooth)" and "Warmness (Cold–Warm)" [7]. In training, the MRHSMM is constructed using PE intensity scores, which are obtained by conducting a listening test using recorded speech data of the training speakers. In synthesis, mean vectors of the distributions are generated based on an intensity score vector determined by the user. Then, we obtain synthetic speech with arbitrary speaker individualities. However, this method is difficult to apply to the control of voice characteristics of a certain existing speaker because the training procedure does not assume such a purpose. In addition, the number of model parameters related to each PE is not optimal because the MRHSMM uses the same decision trees among the respective PE parameters.

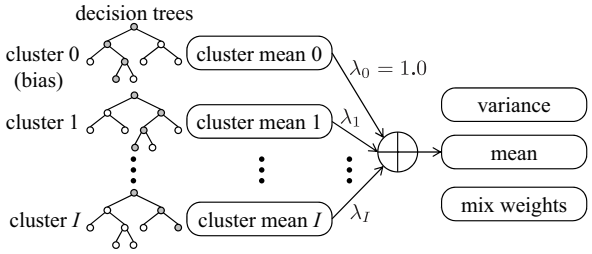


Figure 1: CAT model with cluster-dependent decision trees.

Another type of linear combination method, namely, cluster adaptive training (CAT) has been proposed, which was originally employed for rapid speaker adaptation in speech recognition [8]. The CAT framework expresses mean vectors of distributions by a linear combination of mean vectors from multiple clusters, each of which represents a certain kind of voice characteristics. CAT for speech synthesis has succeeded in some applications such as modeling of speaker variations [9, 10], speaker and language factorization [11], speaker and emotion factorization [12] and transplant of emotions [13, 14].

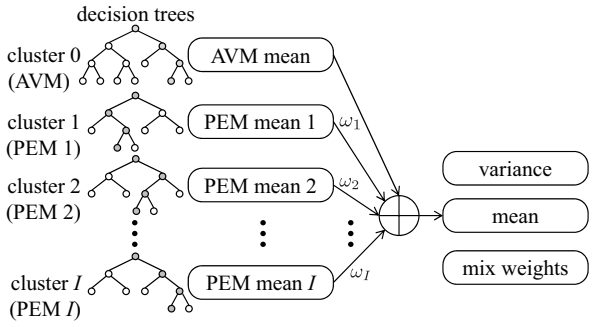
This paper proposes voice quality control based on CAT. In the proposed method, instead of the absolute PE intensity scores used in the conventional MRHSMM-based method, we use relative PE intensity scores between synthetic speech of an average voice model (AVM) [15] and that of the target voice model. In training, the proposed method uses CAT under the conditions that the AVM is set to the bias cluster and the PE intensity scores are used as the weights for the clusters. Consequently, we obtain clusters that model differences between the voice quality of the PE and that of the AVM. We call them perceptual expression models (PEMs) in this paper. Using the PEMs, we can control and edit voice qualities of the existing speakers intuitively.

The rest of this paper is organized as follows. Section 2 reviews the CAT framework in speech synthesis. The details of the proposed method are described in Section 3. Section 4 shows the results of evaluation experiments. Finally, we conclude this paper in Section 5.

2. Cluster adaptive training

In speech synthesis, unlike the original CAT [8], the CAT model includes several clusters with different decision trees as shown in Figure 1. The cluster 0 is defined as the bias and always weighted with 1.0. A weighted sum of all clusters makes a mean vector of the probability distribution. The acoustic feature at frame t $\boldsymbol{o}_t^{(s)}$, which includes static and delta parameters for a given speaker s, is modeled with a CAT model included in $I+1$

http://dx.doi.org/10.21437/Interspeech.2016-290



PEM: perceptual expression model

Figure 2: Proposed CAT-model with the perceptual expression clusters.

clusters as follows:

$$\mathrm{P}\left(\boldsymbol{o}_t^{(s)}|m,\boldsymbol{\lambda}_m^{(s)},\mathcal{M}_m\right)=\mathcal{N}\left(\boldsymbol{o}_t^{(s)};\boldsymbol{\mu}_m^{(s)},\boldsymbol{\Sigma}_{v(m)}\right), \quad (1)$$

$$\boldsymbol{\mu}_m^{(s)}=\sum_{i=0}^{I}\lambda_{i,w(m)}\boldsymbol{\mu}_{c_i(m)},\lambda_{w_0(m)}=1.0, \quad (2)$$

where, $\mathcal{N}(\cdot)$ means a Gaussian distribution, and $\boldsymbol{\lambda}_m^{(s)} = \left[\lambda_{0,w(m)},\lambda_{1,w(m)},\cdots,\lambda_{I,w(m)}\right]^\top$ ($\top$: transpose) denotes a vector of weights for the respective clusters (CAT weights) for the speaker s associated with the regression class $w(m)$. $\boldsymbol{\mu}_{c_i(m)}$ and $\boldsymbol{\Sigma}_{v(m)}$ are the mean vector of the i^{th} cluster and the covariance matrix determined by the bias cluster, respectively. $\mathcal{M}_m$ is the canonical parameter set, i.e., cluster means and covariance of component m. The CAT model can represent voice characteristics of various speakers using appropriate $\boldsymbol{\lambda}$.

In training, the CAT framework iteratively updates CAT weights and canonical parameters. In initialization of the model, the framework iteratively builds each cluster's decision tree using multiple training speakers' data and their initial CAT weights [12] based on the minimum description length (MDL) [16] and the cross-validation [17]. At that time, it updates only the parameters of a single target cluster [18]. Then, we iteratively optimize the CAT weight set $\left\{\boldsymbol{\lambda}_m^{(1)},\boldsymbol{\lambda}_m^{(2)},\cdots,\boldsymbol{\lambda}_m^{(S)}\right\}$ and the canonical parameter set $\mathcal{M}_m$ based on the maximum likelihood criterion in a complementary way [9].

In synthesis, the CAT framework selects cluster means from respective clusters based on an input context and builds a distribution for each state written above using the CAT weights for the target speaker. An acoustic parameter sequence is generated using the parameter generation algorithm [1, 19] and we then obtain the synthetic speech of the target speaker.

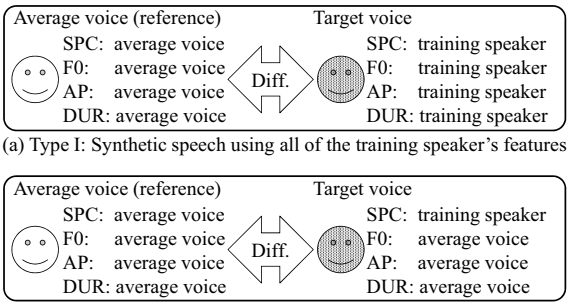
3. CAT based on perceptual expressions

3.1. Definition

Figure 2 shows the proposed CAT-model based on perceptual expressions. The proposed model structure includes an average voice model [15] as the bias cluster and perceptual expression models (PEMs) as other clusters. A PEM represents the voice quality difference regarding a certain perceptual expression (PE). In the proposed method, the probability density function for speaker s is formulated as follows:

$$\mathrm{P}\left(\boldsymbol{o}_t^{(s)}|m,\boldsymbol{\omega}_m^{(s)},\mathcal{M}_m^{(E)},\mathcal{M}_m^{(A)}\right)=\mathcal{N}\left(\boldsymbol{o}_t^{(s)};\tilde{\boldsymbol{\mu}}_m^{(s)},\boldsymbol{\Sigma}_{a(m)}\right), \quad (3)$$

$$\tilde{\boldsymbol{\mu}}_m^{(s)}=\boldsymbol{\mu}_{a(m)}+\sum_{i=1}^{I}\omega_{i,w(m)}^{(s)}\boldsymbol{\mu}_{c_i(m)}, \quad (4)$$



(a) Type I: Synthetic speech using all of the training speaker's features

(b) Type II: Synthetic speech using some of the training speaker's features

Figure 3: Stimuli for the scoring of the perceptual expression. (b) shows an example of the scoring for the spectral feature.

where, $\mathcal{M}_m^{(E)} \in \left\{\boldsymbol{\mu}_{c_1(m)},\boldsymbol{\mu}_{c_2(m)},\cdots,\boldsymbol{\mu}_{c_I(m)}\right\}$ is a parameter sets of the PEMs and $\mathcal{M}_m^{(A)} \in \left\{\boldsymbol{\mu}_{a(m)},\boldsymbol{\Sigma}_{a(m)}\right\}$ represents that of the AVM. The weight vector for the speaker s $\boldsymbol{\omega}_m^{(s)} = \left[\omega_{1,w(m)}^{(s)},\omega_{2,w(m)}^{(s)},\cdots,\omega_{I,w(m)}^{(s)}\right]^\top$ means the PE intensity scores related to the component m.

3.2. Scoring of perceptual expression intensity

To train the PEMs, we use the relative PE intensity scores between each training speaker's voice and the average voice. In this paper, to obtain the relative scores, we use synthetic speech both of each training speaker's voice and the average voice. Because original speech includes richer voice characteristics than synthetic speech, we thus consider that it is difficult to obtain proper PE intensity scores by comparing original speech of a training speaker and a synthetic speech of the AVM. In addition, by using synthetic speech, we are able to get detailed PE intensity scores for respective types of synthetic parameters such as spectral feature (SPC), fundamental frequency (F0), aperiodic component (AP) [20] and duration (DUR).

In scoring for relative PE intensities, we conduct subjective evaluations on our crowdsourcing system. The subjects listen to a pair of synthetic speech samples of average voice (reference) and target voice, and they then evaluate a relative intensity at 11 levels for each PE expressed by a pair of words. For example, in the case of "Brightness (Dark–Bright)," +5 means "Bright," and -5 means "Dark." As the target voice, we employ two types of synthetic speech as shown in Figure 3. Type I (Figure 3 (a)) is intended to capture the total expression of the training speaker's voice. Therefore, the target voice is generated using all the synthetic parameters of the training speaker. On the other hand, Type II (Figure 3 (b)) is aimed at obtaining the detailed PE scores for respective synthetic parameters. The target voice is generated using parameters of the training speaker for one parameter type and those of the AVM for remaining parameter types. This evaluation is conducted for respective synthetic parameter types, i.e., SPC, F0, AP and DUR. Note that the training speakers' models are constructed by speaker adaptation using the AVM [15] in the present work.

3.3. Training of perceptual expression model

In the proposed method, we employ the AVM trained in advance and only build the PEMs based on the relative PE intensity scores in Sec. 3.2. In initialization of the PEMs, we construct the MRHSMM that is similar to [6]. The regression matrix $\boldsymbol{H}^{(E)}$ is calculated using the AVM, the training speakers'

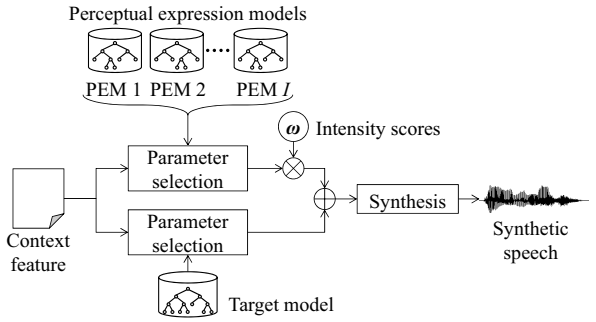


Figure 4: Synthesis process with the PEMs.

models and their corresponding PE intensity scores as follows:

$$\hat{\boldsymbol{H}}^{(E)} = \arg\min_{\boldsymbol{H}^{(E)}} \sum_{s=1}^{S} \left|\left| \left(\overline{\boldsymbol{\mu}}^{(s)} - \overline{\boldsymbol{\mu}}^{(a)}\right) - \boldsymbol{H}^{(E)}\boldsymbol{\omega}^{(s)} \right|\right|, \quad (5)$$

where, $\overline{\boldsymbol{\mu}}^{(s)}$ and $\overline{\boldsymbol{\mu}}^{(a)}$ denote supervectors of the training speaker s and the AVM, respectively. Then each column vector is assigned to the initial mean vectors of the corresponding PEM.

Next, we optimize decision trees in each PEM. In the proposed method, in order to capture the common voice quality included in the training speakers, we apply the shared context clustering approach [21] to each PEM. Then, the total PEM parameter set $\boldsymbol{\mathcal{M}}^{(E)}$ is updated based on the maximum likelihood criterion as follows:

$$\hat{\boldsymbol{\mathcal{M}}}^{(E)} = \arg\max_{\boldsymbol{\mathcal{M}}^{(E)}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} \mathrm{P}\left(\boldsymbol{o}_t^{(s)} | \boldsymbol{\omega}^{(s)}, \boldsymbol{\mathcal{M}}^{(E)}, \boldsymbol{\mathcal{M}}^{(A)}\right). \quad (6)$$

The auxiliary function for the PEM $\boldsymbol{\mathcal{M}}^{(E)}$ is defined as follows:

$$Q\left(\boldsymbol{\mathcal{M}}^{(E)}, \hat{\boldsymbol{\mathcal{M}}}^{(E)}\right) = -\frac{1}{2} \sum_{\substack{m,\, s \\ t \in T_s}} \gamma_{m,t}^{(s)} \left(\boldsymbol{o}_t^{(s)} - \tilde{\boldsymbol{\mu}}_m^{(s)}\right)^{\top} \boldsymbol{\Sigma}_{a(m)}^{-1} \left(\boldsymbol{o}_t^{(s)} - \tilde{\boldsymbol{\mu}}_m^{(s)}\right) + C, \quad (7)$$

where, $\gamma_{m,t}^{(s)}$ denotes the occupancy probability for speaker s of component m at frame t and C is a constant value. The estimated mean vector in the i^{th} PEM $\hat{\boldsymbol{\mu}}_{c_i(m)}$ is given by

$$\hat{\boldsymbol{\mu}}_{c_i(m)} = \boldsymbol{G}_{ii}^{(m)^{-1}} \left(\boldsymbol{k}_i^{(m)} - \sum_{j=1, j\neq i}^{I} \boldsymbol{G}_{ij}^{(m)} \boldsymbol{\mu}_{c_j(m)} \right), \quad (8)$$

where,

$$\boldsymbol{G}_{ij}^{(m)} = \sum_{s, t\in T_s} \gamma_{m,t}^{(s)} \omega_{i,w(m)}^{(s)} \boldsymbol{\Sigma}_{a(m)}^{-1} \omega_{j,w(m)}^{(s)}, \quad (9)$$

$$\boldsymbol{k}_i^{(m)} = \sum_{s, t\in T_s} \gamma_{m,t}^{(s)} \omega_{i,w(m)}^{(s)} \boldsymbol{\Sigma}_{a(m)}^{-1} \left(\boldsymbol{o}_t^{(s)} - \boldsymbol{\mu}_{a(m)}\right). \quad (10)$$

3.4. Synthesis with perceptual expression model

Figure 4 shows the synthesis process using an arbitrary target speaker's model. According to an input context, we select model parameters from the target model and the PEMs, respectively. Based on the PE intensity scores from the user, a mean vector of the target speaker is controlled as follows:

$$\boldsymbol{\mu}_m = \boldsymbol{\mu}_{c(m)}^{(tar)} + \sum_{i=1}^{I} \omega_{i,w(m)} \boldsymbol{\mu}_{c_i(m)}, \quad (11)$$

where, $\boldsymbol{\mu}_{c(m)}^{(tar)}$ and $\boldsymbol{\mu}_m$ are the original and controlled target mean vector, respectively. Then we perform the parameter generation [1] and finally obtain target synthetic speech with voice quality control.

Table 1: Perceptual expressions and their related pairs of words. In the questionnaire, the words on the left are scored with negative values and those on the right are given positive values.

PE type	Pair of words
Gender	Female – Male
Age	Young – Old
Brightness	Dark – Bright
Tightness	Soft – Tight
Intelligibleness	Muffled – Intelligible
Fluency	Halting – Fluent
Clarity	Hoarse – Clear

4. Experiments

4.1. Conditions

To determine a proper small set of PEs for representing the voice quality, we conducted a preliminary questionnaire on our crowdsourcing system. In this questionnaire, we used PEs described in [22, 23]. In the result of the questionnaire, we employed 7 PEs and their pairs of words shown in Table 1. PE intensity scores of gender and age were obtained by Type I as illustrated in Figure 3 and remaining PEs were scored by conducting both Type I and II. All PE intensity scores were normalized within -1.0–1.0.

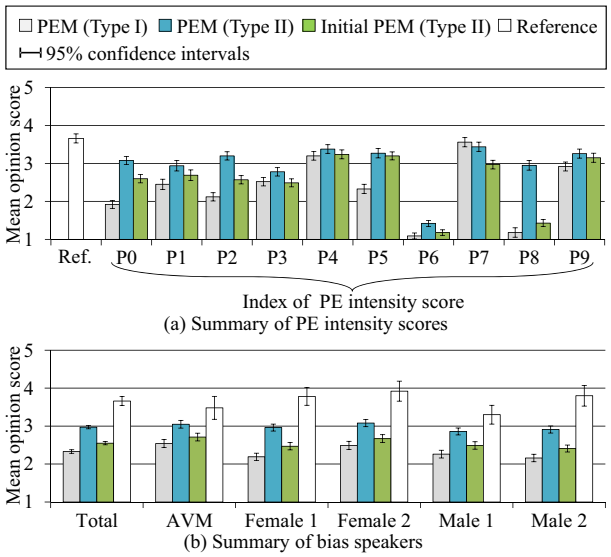
We used the HSMM with 5 states, left-to-right and no skip structure for all models. The AVM was trained in advance using speech data of 7 males and 8 females, which contained 12332 utterances in total. The PEMs were constructed using speech data of 16 males and 24 females including the training speakers for the AVM. For training the PEMs, we randomly selected 100 utterances at most from subsets of each training speaker and then used 12427 utterances. We employed 2 males and 2 females as the target speakers not included in the training speakers. The target speakers respectively uttered 376 sentences and their models were built by MLLR-based speaker adaptation from the AVM.

The sampling frequency was 22.05 [kHz]. The speech waveforms were analyzed by a pitch-synchronous Fourier transform with 1024 points. Mel-scaled line spectral pairs were extracted from pitch-synchronous spectra, and the respective parameters included delta coefficients. Banded aperiodic component was calculated with the pitch-scaled harmonic filter [24] and consisted of 20-dimensional components that are divided by the same interval on the linear scale, and their delta components. A log-F0 vector was composed of a static, delta and delta-delta components.

4.2. Results of subjective evaluations

4.2.1. Speech quality in the combination of PEMs and target speakers

First, speech quality was subjectively evaluated by the 5-level mean opinion score (MOS) test (1: poor and 5: excellent) on our crowdsourcing system. We used the AVM and target speakers' models as the bias speakers. To make controlled synthetic speech, 10 types of PE intensity scores were automatically generated at random within -1.0–1.0. In addition, we used synthetic speech of the AVM or the target speaker models as the reference. In this test, we employed three types of PEMs such as one using Type I score ("PEM (Type I)"), one using Type II score ("PEM (Type II)") and the initial PEM using Type II ("Initial PEM (Type II)"). Note that "Initial PEM (Type II)" is regarded as the conventional MRHSMM because it has the same struc-



(a) Summary of PE intensity scores

(b) Summary of bias speakers

Figure 5: Results of speech quality tests with several target speakers.

Table 2: Results of identification tests. (%)

(a) PEM (Type I)				(b) PEM (Type II)		
Changed (correct)	Changed (incorrect)	Not changed	PE word	Changed (correct)	Changed (incorrect)	Not changed
95.11	1.63	3.26	Female	92.96	2.35	4.69
97.62	0.95	1.43	Male	95.79	1.40	2.80
95.69	3.35	0.96	Young	96.28	2.13	1.60
88.77	2.67	8.56	Old	85.19	4.17	10.65
95.17	2.90	1.93	Dark	92.42	2.84	4.74
94.81	1.42	3.77	Bright	93.05	1.07	5.88
32.45	43.09	24.47	Soft	84.48	3.45	12.07
42.25	36.36	21.39	Tight	76.78	6.16	17.06
57.84	19.46	22.7	Muffled	64.26	10.64	25.11
46.92	36.97	16.11	Intelligible	86.55	4.20	9.24
84.62	3.42	11.97	Halting	87.20	2.84	9.95
86.73	10.90	2.37	Fluent	81.13	9.43	9.43
43.87	39.15	16.98	Hoarse	29.57	20.97	49.46
45.45	49.20	5.35	Clear	45.53	16.17	38.30

ture as the MRHSMM. The number of test utterances was four, and the total number of listeners was 43, in this evaluation.

Figure 5 shows the results of the speech quality tests. Figure 5 (a) shows the MOS scores summarized for each combination of PE intensity scores. According to this, synthetic speech of "PEM (Type I)" has worse speech quality than the other PEMs in most cases. This indicates that the scoring based on Type I may not always give proper intensities to each synthetic parameter and then "PEM (Type I)" is not modeled adequately. On the other hand, "PEM (Type II)" keeps good speech quality in all combinations of PE intensity scores except P6, and is even better than "Initial PEM (Type II)" in many combinations. Thus, the PEM is able to control voice qualities robustly in terms of speech quality. However, all PEMs degrade speech quality in the case of P6. The generated spectra become distorted because absolute values of all PE intensity scores in P6 are larger than 0.5. From Figure 5 (b), which summarizes the results for each bias speaker, "PEM (Type II)" has better speech quality than the other models for all the bias speakers, and these results thus suggest that "PEM (Type II)" can change voice quality even of arbitrary speakers while keeping good speech quality.

Table 3: Result of 5-level MOS test between the traditional CAT model and the PEM.

Traditional CAT model	2.99±0.0656
PEM (Type II)	3.00±0.0673

4.2.2. Identification of perceptual expressions

To evaluate the controllability of the PEMs, we conducted identification tests for the PEs. In these evaluations, we gave each listener a pair of reference and target speech, and asked the listener to choose one from three options. For example, in the case of the brightness, listeners should select "change to dark," "change to bright" or "not changed." In this evaluation, we used synthetic speech generated from the AVM as the reference, and synthetic speech generated from a "PEM (Type II)" (bias is AVM) with a certain combination of the controlled PE intensity score as the target. In each identification test, we set the intensity scores to -0.5 and 0.5. We used five test sentences and had 55 listeners in total on the crowdsourcing system.

Table 2 presents the results of the identification tests. In several tests, the listeners can perceive the change of the voice qualities labeled with the PEs correctly. However, some PE words such as "Soft," "Tight," "Intelligible," "Hoarse" and "Clear" cannot be perceived correctly in "PEM (Type I)." These results suggest that it is difficult to model these PEMs using PE intensity scores of Type I. On the other hand, "Soft," "Tight," and "Intelligible" are identified correctly in "PEM (Type II)." Therefore, "PEM (Type II)" achieves modeling superior to "PEM (Type I)." However, "Clarity (Hoarse–Clear)" is not identified correctly even in "PEM (Type II)." Thus, improvement of the modeling of this PE is a subject for future work.

4.2.3. Comparison of traditional CAT model

Finally, we compared "PEM (Type II)" with the traditional CAT model in terms of speech quality by the 5-level MOS. We achieved voice quality control for the traditional CAT model by converting the PE intensity scores into CAT weights based on multiple regression analysis. We used the AVM as the target speaker's model and employed the same PE intensity scores and test sentences as in Section 4.2.1. The number of listeners was 20.

The result in Table 3 shows that the proposed method has the same speech quality as the traditional CAT. Thus, this indicates that the proposed method can model the voice with performance equivalent to that of the traditional CAT. The proposed method still has an advantage over the traditional CAT, i.e., the proposed method is easy to apply to the control of voice characteristics of any existing speaker structure while the traditional CAT is difficult to apply due to its model structure.

The results of the subjective evaluations suggest that the proposed method with "PEM (Type II)" can achieve good speech quality, robustness to various combinations of PE intensity scores, and good controllability of the PEs.

5. Conclusions

This paper proposed a novel voice quality control method for speech synthesis based on cluster adaptive training. The proposed method builds perceptual expression models (PEMs) using perceptual expression (PE) scores that represent differences between the average voice and training speakers in terms of the voice qualities. Thus, each PEM models differences of acoustic features related to a PE. Subjective experimental results demonstrate that the proposed method can control voice qualities while keeping speech quality.

6. References

[1] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP*, vol. 1, pp. 660–663, May 1995.

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

[3] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Regression approaches to voice quality control based on one-to-many eigenvoice conversion," *6th ISCA Speech Synthesis Workshop (SSW6)*, pp. 101–106, Aug. 2007.

[4] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Adaptive voice-quality control based on one-to-many eigenvoice conversion," *Proc. INTERSPEECH2010*, pp. 2158–2161, Sep. 2010.

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

[6] T. Nose and T. Kobayashi, "An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model," *Speech Communication*, vol. 55, pp. 347–357, Feb. 2013.

[7] M. Tachibana, T. Nose, J. Yamagishi, and T. Kobayashi, "A technique for controlling voice quality of synthetic speech using multiple regression HSMM," *Proc. INTERSPEECH2006-ICSLP*, pp. 2438–2441, Sep. 2006.

[8] M. J. F. Gales, "Cluster adaptive training for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[9] V. Wan, J. Latorre, K. Chin, L. Chen, , M. Gales, H. Zen, K. Knill, and M. Akamine, "Combining multiple high quality corpora for improving HMM-TTS," *Proc. INTERSPEECH2012*, p. Tue.O5d.01, Sep. 2012.

[10] V. Wan, J. Latorre, K. Yanagisawa, M. Gales, and Y. Stylianou, "Cluster adaptive training of average voice models," *Proc. ICASSP 2014*, pp. 280–284, 2014.

[11] H. Zen, , N. Braunshweiler, S. Buchhoz, M. J. F. Gales, K. Knill, Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 6, pp. 1713–1724, Aug. 2012.

[12] J. Latorre, V. Wan, M. Gales, L. Chen, K. Chin, K. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training," *Proc. INTERSPEECH2012*, p. Tue.P4C.04, Sep. 2012.

[13] L. Chen, N. Braunshweiler, and M. J. F. Gales, "Speaker and expression factorization for audiobook data- expressiveness and transplantation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 605–618, Apr. 2015.

[14] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine, "Emotional transplant in statistical speech synthesis based on emotion additive model," *Proc. INTERSPEECH2015*, pp. 274–278, Sep. 2015.

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

[16] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 2, pp. 77–86, Mar. 2000.

[17] T. Shinozaki, "HMM state clustering based on efficient cross validation," *Proc. INTERSPEECH2006-ICSLP*, pp. 1157–1160, Sep. 2006.

[18] K. Saino, "A clustering technique for factor analyzed voice models," *Master thesis, Nagoya Institute of Technology*, 2008.

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

[20] 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," *MAVEBA 2001*, Sep. 2001.

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

[22] H. Kido and H. Kasuya, "Everyday expressions associated with voice quality of normal utterance –extraction by perceptual evaluation–," *J. Acoust. Soc. Jpn*, vol. 57, no. 5, pp. 337–344, Jan. 2001 (in Japanese).

[23] K. Takamuku, M. Higashi, and Y. Tanida, "Modeling of perception and cognitive structure by words of voice impressions," *Proc. Spring meeting of Acoust. Soc. Jpn.*, no. 2-Q5-2, pp. 451–454, Mar. 2014 (in Japanese).

[24] P. Jackson and C. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Sep. 2001.