AN ESTIMATION METHOD OF VOICE TIMBRE EVALUATION VALUES USING FEATURE EXTRACTION WITH GAUSSIAN MIXTURE MODEL BASED ON REFERENCE SINGER

Soichi Yamane[1]*, Kazuhiro Kobayashi*[1]*, Tomoki Toda*[2]*,*
Tomoyasu Nakano[3]*, Masataka Goto*[3]*, Satoshi Nakamura*[1] *

[1]Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
[2]Information Technology Center, Nagoya University, Japan
[3]National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

This paper presents an estimation method of voice timbre evaluation values for arbitrary singer's singing voices generated with a singing voice synthesis system towards the development of a singing voice retrieval system. The voice timbre evaluation values are numerical values corresponding to voice timbre expression words, such as "Age" and "Gender", and they usually need to be manually assigned to individual singers' singing voices through listening. To make it possible to automatically estimate them from given singer's singing voices, an acoustic feature to well capture only each singer's voice timbre is extracted with a Gaussian mixture model trained using parallel data between singing voices sung by many pre-stored target singers and same voices sung by a reference singer. Then, the voice timbre evaluation values are estimated from the extracted feature using regression models. The experimental results showed that the proposed method is capable of accurately estimating those values for some expression words, such as "Age" and "Gender", and nonlinear regression is effective for the expression words, "Powerfulness" and "Uniqueness."

Index Terms— singing voice synthesis, voice timbre, estimation of evaluation values, Gaussian mixture model, reference singer

1. INTRODUCTION

In creating vocal music, a singing voice synthesis system, such as VOCALOID [1], UTAU [2] or Sinsy [3], is often used by many end users. The singing voice synthesis system allows the users to easily synthesize a singing voice as they want by manually inputting score information such as pitch, onset time, and duration to control melody and linguistic information to represent desired lyrics. Moreover, it enables the users to easily change voice timbre of synthesized singing voices to not only manipulate control parameters of voice timbre but also select different singers' singing voice data. In particular, voice timbre of singing voices has a large impact on the vocal music, and therefore, it is important for the users to carefully select it so that voice timbre of synthesized singing voices well suits to the music created by the user. However, a large number of available singing voice data exist currently and it tends to increase more and more. For example, UTAU voice libraries [2] include over 5000 kinds of singing voice data [4]. Consequently, it is difficult to search for the most suitable singing voice data among them. It will be helpful to develop a system to retrieve the suitable singing voice data.

Music information retrieval has been widely studied and various methods focusing on singing voices have also been proposed. For example, several music information retrieval systems based on voice timbre similarity have been proposed to search for music data including a singing voice of which voice timbre is similar to that in input music data [5–10]. As another retrieval method, a singing style retrieval method uses an input singer's singing voice to search for singing voices sung in a similar singing style to that of the input one based on a similarity measure calculated with a probability distribution on the phase plane (i.e., $f_0 - \Delta f_0$ plane) to express dynamic variations of f_0 patterns [11]. In these methods, the user basically needs to input reference singing voice or music data as a query of which a singing style or voice timbre is similar to the target one that the user wants to search for. Therefore, it is still difficult to search for the desired singing voices used in the singing voice synthesis system if the user cannot find such a suitable reference data.

To develop a singing voice retrieval method with no need of reference data, it is essential to define a measure to describe target characteristics of singing voices, such as a singing style or voice timbre. In this paper, we focus on voice timbre because a singing style can be well controlled by the users in the singing synthesis system. As related work, there have been several attempts at describing voice quality of speaking voices using an evaluation value on voice quality expression words [12] and singing impression words [13]. According to the article [12], several word pairs expressing voice quality, such as husky/clear (clearness) and elder/younger (age), have been selected by applying factor analysis to a result of a large-scaled perceptual evaluation using many speakers' natural voices. Each word pair can be used to manually assign a 5-scaled evaluation value (-2: disagree, -1: disagree a little, 0: neither, 1: agree a little, 2: agree) to describe voice quality of individual natural voices. It has been reported that this description method is helpful for developing a speech synthesis system or a voice conversion system making it possible for the users to intuitively control voice quality of synthesized/converted speech as they want [14, 15].

Inspired by this conventional method, in this paper we also use evaluation values for several word pairs to describe voice timbre of individual singers' singing voices. These voice timbre evaluation values will be effectively used to evaluate similarity of voice timbre between different singing voice data, and also used to intuitively design input query for searching for singing voice data with the desired voice timbre. On the other hand, in order to develop such a singing voice data retrieval system, it is inevitable to assign the voice timbre evaluation values to all existing singing voice data. However, these

*We thank to Suzuki Serif who provided the voice timbre evaluation values intended for the UTAU voice libraries used in this experiments. This work was supported in part by JSPS KAKENHI Grant Number 26280060 and by the JST OngaCREST project.

values basically need to be manually assigned to each singing voice data through listening. In order to reduce a huge amount of effort to do it, it is worthwhile to develop a technique for automatically assigning these values to the existing singing voice data.

In this paper, we propose an automatic estimation method of the voice timbre evaluation values towards the development of a singing voice data retrieval system helpful for the users to find their desired singing voice data to be used in the singing voice synthesis system. To extract an acoustic feature to well capture only voice timbre of each singing voice data, we use the joint probability density modeling method based on a reference singer [16], which was originally proposed as a voice conversion technique [17, 18] and then was successfully applied to singing voice conversion as well [19]. This method makes it possible to separately model voice timbre of singing voices and acoustic variations caused by changes of phones using a Gaussian mixture model (GMM) trained with parallel data sets of synthesized singing voices between the reference singer and many pre-stored target singers. After extracting the acoustic features of individual pre-stored target singers, regression analysis is performed to develop an estimation model of the voice timbre evaluation values from the extracted acoustic feature. We conduct several experimental evaluations to investigate 1) the effectiveness of the proposed estimation method, 2) the effectiveness of using nonlinear regression rather than linear regression, and 3) an effective way to develop parallel data sets used for the GMM training.

2. VOICE TIMBRE FEATURE EXTRACTION BASED ON JOINT PROBABILITY DENSITY MODELING WITH REFERENCE SINGER

As acoustic features to capture voice timbre, we used segmental features, such as a spectral parameter and an aperiodic parameter. However, these features are also affected by phonemes and prosody (i.e., F_0 and duration). Therefore, it is essential to remove their effects on the acoustic features so as to represent only voice timbre. To do so, we apply the joint probability density modeling technique with a reference singer to the voice timbre feature extraction of singing voice data used in a singing voice synthesis system. Figure 1 shows an overview of the proposed feature extraction process.

First, parallel data sets of singing voices sharing the same score information and lyrics between a single reference singer and many pre-stored target singers are synthesized by using a singing voice synthesis system. The segmental features are extracted frame by frame from each singing voice and they are time aligned between each pre-stored target singer and the reference singer. Then, a joint probability density function of these time-aligned segmental features between the reference singer and the s-th pre-stored target singer is modeled with a GMM as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\mu}^{(s)}, \boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(s)\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \quad (1)$$

$$\boldsymbol{\mu}^{(s)} = \left[\boldsymbol{\mu}_1^{(Y)\top}(s), \cdots, \boldsymbol{\mu}_m^{(Y)\top}(s), \cdots, \boldsymbol{\mu}_M^{(Y)\top}(s)\right]^\top, \quad (2)$$

where $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{\Delta}\boldsymbol{x}_t^\top\right]^\top$ is a joint static and dynamic feature vector of the reference singer, $\boldsymbol{Y}_t^{(s)} = \left[\boldsymbol{y}_t^{(s)\top}, \boldsymbol{\Delta}\boldsymbol{y}_t^{(s)\top}\right]^\top$ is that of the s-th pre-stored target singer, and $\top$ denotes transposition. The normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The total number of mixture components is M and the mixture component index is m. The m-th mixture component weight is α_m. The mean vector of the m-th mixture component consists of the reference singer's mean vector $\boldsymbol{\mu}_m^{(X)}$ and the s-th pre-stored target singer's mean vector $\boldsymbol{\mu}_m^{(Y)}(s)$. The covariance matrix of the m-th mixture component consists of the reference singer's covariance matrix $\boldsymbol{\Sigma}_m^{(XX)}$, the pre-stored target singer's covariance matrix $\boldsymbol{\Sigma}_m^{(YY)}$, and their cross covariance matrices $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$. Note that these parameters except for the pre-stored target singer's mean vectors are shared among all pre-stored target singers, and only the target mean vectors depend on individual pre-stored target singers. Therefore, they are concatenated to develop a super vector $\boldsymbol{\mu}^{(s)}$ as the voice timbre feature vector of the s-th pre-stored target singer. The other parameters are included in a shared parameter set $\boldsymbol{\lambda}$.

To optimize these parameters, first a target-singer-independent GMM is trained using all parallel data sets between the reference singer and the individual pre-stored target singers as follows:

$$\left\{\boldsymbol{\mu}^{(0)}, \boldsymbol{\lambda}^{(0)}\right\} = \arg\max_{\{\boldsymbol{\mu}, \boldsymbol{\lambda}\}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\mu}, \boldsymbol{\lambda}\right), \quad (3)$$

where T_s is the number of time-aligned frames of the parallel data set for the s-th pre-stored target singer and S is the total number of the pre-stored target singers. Then, a singer-dependent GMM for the s-th pre-stored target singer, which is given by Eqs. (1) and (2), is trained by updating only the super vector $\boldsymbol{\mu}^{(0)}$ using only the s-th parallel data set as follows:

$$\boldsymbol{\mu}^{(s)} = \arg\max_{\boldsymbol{\mu}^{(0)}} \prod_{t=1}^{T_s} P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\mu}^{(0)}, \boldsymbol{\lambda}^{(0)}\right). \quad (4)$$

It is noted that each mixture component consistently models the same phone space over all singers, i.e., the reference singer and all pre-stored target singers, thanks to the use of the parameters shared over different singers and the use of parallel data in the parameter optimization [16, 17]. In the joint density modeling, the reference singer's data plays a role as an anchor to align each mixture component to the same phone space over all pre-stored target singers. Consequently, acoustic variations caused by different phones are modeled with different mixture components and only singer-dependent acoustic features are modeled with the super vector in Eq. (2). The use of parallel data sets also sharing prosodic features as well as linguistic information effectively makes differences of the resulting super vector between different pre-stored target singers depend on only differences of their voice timbre.

3. AUTOMATIC ESTIMATION OF VOICE TIMBRE EVALUATION VALUE USING REGRESSION ANALYSIS

A regression analysis is applied to the estimation of voice timbre evaluation values. First, we manually assign the voice timbre evaluation values for pre-determined word pairs to express voice timbre into all of or a part of the pre-stored target singers. Then, we develop a regression model to estimate the voice timbre evaluation values from the voice timbre feature using the pre-stored target singers' data.

3.1. Estimation Method based on Multiple Regression

The voice timbre evaluation values of the s-th pre-stored target singer are stored as a voice timbre evaluation vector $\boldsymbol{w}^{(s)} = \left[w_1^{(s)}, \cdots, w_J^{(s)}\right]^\top$, where $w_j^{(s)}$ is the voice timbre evaluation

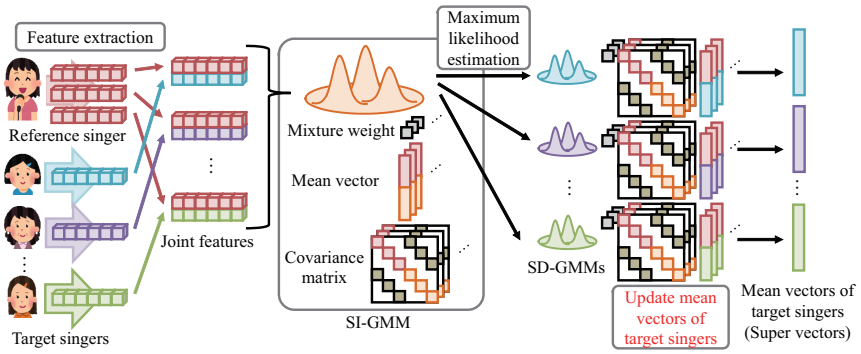

Fig. 1. Extraction of voice timbre features using parallel data

value for the j-th pre-determined word pairs, and the number of voice timbre evaluation values (i.e., the number of pre-determined word pairs) is J. In multiple regression analysis, the voice timbre evaluation vector $\boldsymbol{w}^{(s)}$ is estimated from the corresponding super vector $\boldsymbol{\mu}^{(s)}$ as follows:

$$\hat{\boldsymbol{w}}^{(s)} = \boldsymbol{A}\boldsymbol{\mu}^{(s)} + \boldsymbol{b}, \tag{5}$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ are regression parameters, which are determined with minimum mean square error estimation between the given voice timbre evaluation vectors $(\boldsymbol{w}^1, \cdots, \boldsymbol{w}^S)$ and the estimated ones $(\hat{\boldsymbol{w}}^1, \cdots, \hat{\boldsymbol{w}}^S)$.

3.2. Estimation Method based on Kernel Regression

In kernel regression analysis, the voice timbre evaluation vector of s-th target singer is estimated from the corresponding super vector as follows:

$$\hat{\boldsymbol{w}}^{(s)} = \boldsymbol{V}^\top \boldsymbol{\phi}\left(\boldsymbol{\mu}^{(s)}\right), \tag{6}$$

where $\boldsymbol{\phi}(\cdot)$ is a function to map the super vector to a higher dimensional feature space, and $\boldsymbol{V}$ is a regression parameter in the higher dimensional feature space, which is given by

$$\boldsymbol{V} = \sum_{s=1}^{S} \boldsymbol{\phi}\left(\boldsymbol{\mu}^{(s)}\right) \boldsymbol{Z}_s^\top, \tag{7}$$

where $\boldsymbol{Z}_s$ is a weighting parameter for the s-th mapped super vector. From Eqs. (6) and (7), the estimate of $\boldsymbol{w}^{(s)}$ is written as

$$\hat{\boldsymbol{w}}^{(s)} = \boldsymbol{Z}\boldsymbol{k}\left(\boldsymbol{\mu}^{(s)}\right), \tag{8}$$

$$\boldsymbol{k}\left(\boldsymbol{\mu}^{(s)}\right) = \left[k\left(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(s)}\right), \cdots, k\left(\boldsymbol{\mu}^{(S)}, \boldsymbol{\mu}^{(s)}\right)\right]^\top, \tag{9}$$

where $\boldsymbol{Z} = [\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_S]$, and $k(\cdot, \cdot)$ is a kernel function. In this paper, we use Gaussian kernel as the kernel function, which is given by

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{\sigma^2}\right), \tag{10}$$

where σ is a parameter of any positive value. The weighting parameter $\boldsymbol{Z}$ is determined with minimum mean square error estimation also using regularization as follows:

$$\boldsymbol{Z} = \boldsymbol{W}\left(\boldsymbol{K} + r\boldsymbol{I}\right)^{-1}, \tag{11}$$

where $\boldsymbol{K} = \left[\boldsymbol{k}\left(\boldsymbol{\mu}^{(1)}\right), \cdots, \boldsymbol{k}\left(\boldsymbol{\mu}^{(S)}\right)\right]^\top$, $\boldsymbol{W} = \left[\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(S)}\right]$, and r is a regularization parameter.

3.3. Estimation of Voice Timbre Evaluation Value for Arbitrary Target Singer

In order to automatically estimate the voice timbre evaluation values for a given target singer, first we generate a parallel data set between the reference singer and the given target singer using a singing voice synthesis system, and then, the target singer's super vector is extracted in the same manner as shown in Eq. (4). Finally, the voice timbre evaluation values are estimated from the extracted super vector using the trained regression model.

The proposed method can easily be applied to natural singing voices as well. Thanks to the singing voice synthesis system, a parallel data set consisting of the reference singer's singing voices and the natural singing voices is easily developed by generating the reference singer's singing voices corresponding to the natural ones. Therefore, the voice timbre evaluation values for the given natural singing voice are estimated in the same manner as mentioned above. It is also possible to extract the super vector of the given natural singing voices without generating the parallel data set. The probability density function of only the target acoustic features $P\left(\boldsymbol{Y}_t^{(s)}|\boldsymbol{\mu}^{(0)}, \boldsymbol{\lambda}^{(0)}\right)$ is easily derived from the target-singer-independent GMM $P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\mu}^{(0)}, \boldsymbol{\lambda}^{(0)}\right)$ in Eq. (4) by marginalizing out the reference singer's acoustic features $\boldsymbol{X}_t$. Then, the super vector is optimized by using only acoustic features of the given natural singing voices and the marginalized probability density function. Well-known model adaptation techniques, such as maximum *a posteriori* estimation [20], maximum likelihood linear regression [21], and eigenvoice [22], are available in this optimization.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Conditions

We used 40 kinds of UTAU voice libraries as the singing voice data. Three different types of synthesized voices, "Monosyllabic voice", "Talking voice", and "Singing voice", were generated by using UTAU singing voice synthesis system and they were used for training and evaluation. "Monosyllabic voice" consisted of 100 kinds of Japanese syllables generated with 7 different F_0 values, i.e., 700 samples in total. "Talking voice" consisted of 50 synthesized voices generated by manually mimicking prosody of Japanese normal speech in the ATR phoneme balanced sentence set [23]. "Singing voice" consisted of 60 phrases extracted from 10 songs using MIDI data of the RWC Japanese popular music database [24]. The length of each sample of "Monosyllabic voice", "Talking voice", and "Singing voice" was about 2, 5, and 20 seconds, respectively.

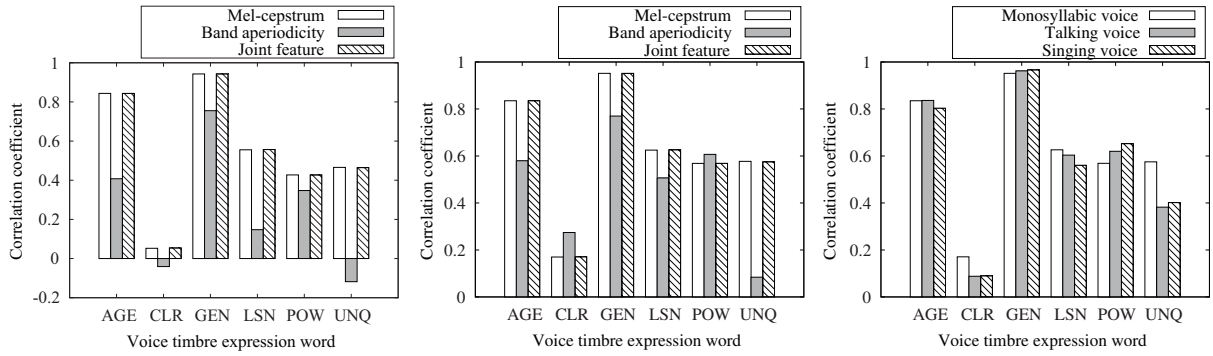

Fig. 2. Correlation coefficients between the estimated and target voice timbre evaluation values when using left) multiple regression, middle) kernel regression, and right) different types of parallel data.

Table 1. Word pairs to express voice timbre

Specification	Word pairs
Adultness (AGE)	Young – Adult
Clearness (CLR)	Noisy – Clear
Gender (GEN)	Feminine – Masculine
Listenability (LSN)	Lisping – Lucidly
Powerfulness (POW)	Tender – Powerful
Uniqueness (UNQ)	Universality – Peculiarity

STRAIGHT analysis [25] was used to extract spectral envelope, which was further parameterized into the 1st through 24th mel-cepstral coefficients as a spectral feature. As an excitation feature, STRAIGHT analysis [26] was also used to extract aperiodic components, which were averaged in five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6 and 6-8 kHz. The frame shift was set to 5 ms. The sampling frequency was set to 16 kHz. The number of mixture components for the spectral feature was 128, and that for the aperiodic feature was 16.

Table 1 shows 6 kinds of word pairs to express voice timbre used in this evaluation. A 7-scaled evaluation value was used to describe voice timbre corresponding to each word pair (e.g. "Adultness (AGE)" is annotated from 1-Young to 7-Adult). 19 annotators manually assigned these values to individual 40 singers' singing voice data, and then values averaged over all annotators were used as the voice timbre evaluation values.

For each acoustic feature (i.e., the spectral feature or the aperiodic feature), the target-singer-independent GMM was trained using all singers' singing voice data. Then, the super vector for each singer was extracted. In regression analysis, leave-one-out cross-validation was performed to train the regression model and evaluate its estimation accuracy of the voice timbre evaluation values. The parameters of the kernel regression were optimized manually.

4.2. Experimental Results

Figure 2 shows the results of estimation of voice timbre evaluation values by the multiple regression (left graph) and the kernel regression (middle graph) when using "Monosyllabic voice" as the parallel data. Each figure shows results using the spectral feature (Mel-cepstrum), the aperiodic feature (Band aperiodicity), and their joint feature as the acoustic feature modeled by the GMM. We can see that the spectral feature yields higher estimation accuracy than the aperiodic feature. Even if using the joint feature, its estimation accuracy is almost the same as that of the spectral feature. Therefore, the aperiodic feature is not effective for the estimation of the voice timbre evaluation values.

We can see that the correlation coefficients for "Adultness (AGE)" and "Gender (GEN)" are over 0.8 and 0.9, respectively. On the other hand, very weak correlation is observed for "Clearness (CLR)." These results are consistent in both the multiple regression and the kernel regression. On the other hand, the correlation coefficients for "Powerfulness (POW)" and "Uniqueness (UNQ)" tend to be increased from 0.4 to 0.6 using the kernel regression rather than the multiple regression. These results imply that there exists nonlinearity in the mapping from the super vector into these voice timbre evaluation values although the number of dimensions of the super vector is already quite high.

The result when using different types of the parallel data is also shown in Fig. 2 (right graph). In this experiment, we used the joint feature and the kernel regression. We can see that the difference of estimation accuracy caused by using different types of the parallel data is small except for estimation of the voice timbre evaluation score on "Uniqueness (UNQ)." Because of these small differences, we may flexibly select a type of parallel data according to that of available singing voice data.

5. CONCLUSION

This paper has presented an estimation method of voice timbre evaluation values from given singing voice data used in a singing voice synthesis system. To extract an acoustic feature to well represent only voice timbre while minimizing the effects of acoustic variations caused by different phones or prosody, we have successfully applied the joint probability density modeling using reference singer's singing voices to the proposed estimation process. The regression analysis have also been used for estimating the voice timbre evaluation values from the extracted voice timbre feature. Experimental evaluations for the estimation of 6 voice timbre evaluation values for "Adultness", "Clearness", "Gender", "Listenability", "Powerfulness" and "Uniqueness" have been conducted, demonstrating that 1) very high estimation accuracy is achieved for "Adultness" ($r > 0.8$) and "Gender" ($r > 0.9$) using the mel-cepstral coefficients as the acoustic feature in the proposed method; 2) improvements in estimation accuracy (from $r = 0.4$ to $r = 0.6$) is observed for "Powerfulness" and "Uniqueness"; and 3) various types of parallel data, such as "Monosyllabic voice", "Talking voice", and "Singing voice", can also be used in the proposed method. We plan to develop a singing voice retrieval system for a singing voice synthesis system using the proposed method.

6. REFERENCES

[1] H. Kenmochi and H. Ohshita, "VOCALOID - commercial singing synthesizer based on sample concatenation," *Proc. INTERSPEECH*, pp. 4011–4012, Aug 2007.

[2] Ameya, "UTAU - singing voice synthesis tool," http://utau2008.web.fc2.com.

[3] K. Oura, A. Mase, S. Muto, Y. Nankaku, and T. Tokuda, "Recent development of the HMM-based singing voice synthesis system - Sinsy," *SSW7*, pp. 211–216, Sep 2010.

[4] Ruto, "UTAU voice libraries database," http://ruto.yu.to/.

[5] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," *Proc. ISMIR*, Sep 2007.

[6] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 15, pp. 519–530, Feb 2007.

[7] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 18, pp. 638–648, Mar 2010.

[8] W. H. Tsai and H. P. Lin, "Background music removal based on cepstrum transformation for popular singer identification," *IEEE Trans Audio Speech and Lang. Process.*, vol. 19, pp. 1196–1205, July 2010.

[9] M. Lagrange, A. Ozerov, and E. Vincent, "Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning," *Proc. ISMIR*, Oct 2012.

[10] T. Nakano, K. Yoshii, and M. Goto, "Vocal timbre analysis latent dirichlet allocation and cross-gender vocal timbre similarity," *Proc. ICASSP*, pp. 5239–5243, May 2014.

[11] T. Kako, Y. Ohishi, H. Kameoka, K. Kashino, and K. Takeda, "Automatic identification for singing style based on sung melodic contour characterized in phase plane," *Proc. ISMIR*, pp. 393–398, Oct 2009.

[12] H. Kido and H. Kasuya, "Everyday expressions associated with voice quality normal utterance extraction by perceptual evaluation," *The Acoustical Society of Japan*, pp. 337–344, May 2001, (in Japanese).

[13] A. Kanato, T. Nakano, M. Goto, and H. Kikuchi, "An automatic singing impression estimation method using factor analysis and multiple regression," *Proc. ICMC SMC*, pp. 1244–1251, Sep 2014.

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

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

[16] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, pp. 1249–1252, Apr 2007.

[17] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," *Proc. INTERSPEECH*, pp. 2446–2449, Sep 2006.

[18] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, Nov 2007.

[19] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-tosinging synthesis system," *Proc. APSIPA ASC*, Nov 2012.

[20] J. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. SAP*, vol. 2, pp. 291–298, Apr 1994.

[21] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Proc. CSL*, vol. 9, no. 2, pp. 171–185, Feb 1995.

[22] R. Kuhn, J. C. Junqua, P. Nugyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. SAP*, vol. 8, no. 6, pp. 695–707, Nov 2000.

[23] K. Iso, T. Watanabe, and H. Kuwabara, "Design of a Japanese sentence list for a speech database," *Preprints, Spring Meeting of Acous. Soc. Jpn.*, vol. 2-2-19, pp. 88–90, Mar 1988, (in Japanese).

[24] M. Goto, T. Nishimura, H. Hashiguchi, and R. Oka, "RWC music database : Music genne database and musical instrument sounddatabase," *Proc. ISMIR*, pp. 229–230, Oct 2003.

[25] K. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, Apr 1999.

[26] H. Kawahara and H. Katayose, "Scat generation research program based on straight a high-quality speech analysis modification and synthesis system," *J. of IPSJ*, vol. 43, no. 2, pp. 208–218, Feb 2002, (in Japanese).