Interspeech 2018
2-6 September 2018, Hyderabad

A Comparative Study of Statistical Conversion of Face to Voice Based on Their Subjective Impressions

Yasuhito Ohsugi, Daisuke Saito, Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo
{yasuhito.ohsugi, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Recently, various types of Voice-based User Interfaces (VUIs) including smart speakers have been developed to be on the market. However, many of the VUIs use only synthetic voices to provide information for users. To realize a more natural interface, one feasible solution will be personifying VUIs by adding visual features such as face, but what kind of face is suited to a given quality of voice or what kind of voice quality is suited to a given face? In this paper, we test methods of statistical conversion from face to voice based on their subjective impressions. To this end, six combinations of two types of face features, one type of speech features, and three types of conversion models are tested using a parallel corpus developed based on subjective mapping from face features to voice features. The experimental results show that each subject judge one specific and subject-dependent voice quality as suited to different faces, and that the optimal number of mixtures of face features is different from the numbers of mixtures of voice features tested.

Index Terms: subjective impressions, face to voice conversion, conditional variational autoencoder, Gaussian mixture model, probabilistic canonical correlation analysis

1. Introduction

In the last two decades, a variety of spoken dialogue systems (SDSs) with embodied conversational agents (ECAs) have been developed [1–6]. In those SDSs, human agents talked to and listened to users on a screen and the agants can play the roles of salesman, language teacher, medical doctor, etc. More recently, a small device such as smart speaker has powerful speech input/output modality and it has become popular even without a human-shaped body. In the future, smart speakers will be personified because ECAs are useful to realize more natural human-machine interaction as discussed in [7]. On the other hand, the modality of spoken language has been technically introduced into doll toys, who have become able to entertain children and even elderly people by speaking. When a smart speaker is personified, the question to ask is, what kind of face should be suited to a given voice quality? Also, when one wants to have his/her doll toy talk, an inverse question is possible, what kind of voice quality should be suited to a given face?

In this paper, we examine methods of statistical conversion from face to voice in order to give an adequate quality of voice to a *voiceless* but human-shaped object, such as toy dolls, text-based dialogue systems with a cyber agent, etc. For conversion from face to voice, we put a focus on subjective impressions of face and voice. In [8–12], it was implied that, based on subjective impressions of the face of a person, human subjects can imagine the voice quality of that person. It is very natural, however, that the real voice quality of that person is different from the imagined voice quality. In this paper, we emphasize use of the imagined voice quality, not the real one.

The paper is organized as follows. In Section 2, we describe some related works about face features, voice features, and models for statistical conversion. In Section 3, we introduce our tested methods. After that, in Sections 4, and 5, collection of subjective mappings from face to voice and evaluation of the tested methods are described, respectively. Finally, in Section 6, this paper is summarized.

2. Related works

As for face features, in this paper, we refer to a top-down method and a bottom-up method. The former firstly detects face landmarks with Constrained Local Neural Field (CLNF) [13] and by using the landmarks, the outlines of facial parts such as nose and mouth are obtained. The latter views a face image as a set of dots and, by using Conditional Variational Autoencoder (CVAE) [14], it can embed an input face image to a latent feature vector using external labels that represent non-facial features of that image. By using this model, from a latent feature vector with external labels, a face image can be generated or restored. As discussed in [15, 16], different latent vectors can be converted into different faces and we can discuss the function of each dimension of the latent vector.

Eigenvoice is one of the well-known methods to represent speaker characters [17], which we consider to be the voice quality of interest. Eigenvoice is calculated through Principal Component Analysis (PCA) performed over supervectors of speaker-dependent Gaussian Mixture Models (SD-GMMs). Speech samples of a given eigenvector can be generated by Eigenvoice Conversion (EVC) [18]. In this paper, we formulate face to voice conversion as face to eigenvoice conversion.

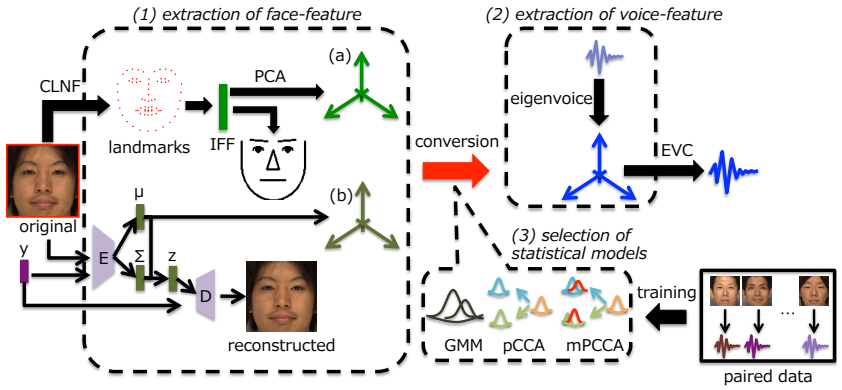
A GMM of joint vectors of $\{\boldsymbol{X}\}$ and $\{\boldsymbol{Y}\}$ is often used to implement conversion between the two spaces. The relation between $\{\boldsymbol{X}\}$ and $\{\boldsymbol{Y}\}$ is modeled explicitly as covariance terms of each Gaussian distribution in the GMM. Voice conversion or speaker identity conversion is often implemented using GMM (GMM-VC) [19]. On the other hand, in order to search for the latent variable between the two spaces, which can capture latent relations between them, Canonical Correlation Analysis (CCA) was proposed [20]. Probabilistic CCA (pCCA) [21] and mixture of pCCA (mPCCA) [22] were also proposed, where $\{\boldsymbol{X}\}$, $\{\boldsymbol{Y}\}$, and their latent variable are assumed to follow Gaussian distributions. In [22], articulatory-to-acoustic conversion was examined based on mPCCA. In this paper, conversion based on deep neural networks is not examined because of the size of training data.

3. Tested features and models

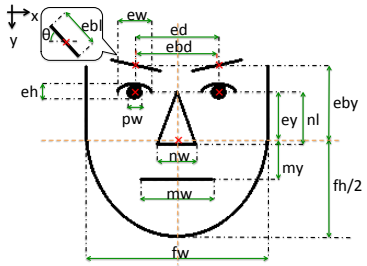
3.1. The overview of the tested methods

In this paper, six combinations of two face features, eigenvoice, and three conversion models are examined. The overview of the experiments is shown in Figure 1a. The two face features tested

10.21437/Interspeech.2018-2005



(a) *The overview*



(b) *IFFs*

Figure 1: *(a) The overview of our tested statistical conversion from face to voice, and (b) Iconified Face Features (IFFs)*

are CVAE features and new features, which are landmark-based and proposed as Iconified Face Features (IFFs) in this paper. IFFs will be explained shortly in Section 3.2. For conversion, the three models of GMM, pCCA, and mPCCA are tested. Generation of voices is done using the framework of EVC.

To realize face-to-voice conversion, a parallel corpus of face features and voice features has to be prepared. In this paper, not a real mapping, e.g. the face and the voice of a person, but subjective mappings from faces to voices are used. Collection of subjective mappings is explained in Section 4.

3.2. Iconified Face Feature (IFF)

CLNF converts a given face image to a set of dots that can capture face landmarks, shown in Figure 1a. To extract their IFFs, the landmarks are converted into iconified facial parts, illustrated in Figure 1b, and the locations of the parts are automatically detected. The IFFs are 15 geometrical parameters to represent the shape and the location of these parts. Finally, a face image is converted into a 15-dimensional vector.

3.3. GMM-based conversion from face to voice

Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ denote a d_x-dimensional vector of the input face feature and a d_y-dimensional vector of the output voice feature, respectively. If a parallel corpus $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ are given, the parameters of GMM $\boldsymbol{\lambda} = \{\alpha_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma_m}^{(z)}\}_{m=1}^M$ are trained to maximize the joint probability as follows,

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \prod_{i=1}^{N} \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}_i; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma_m}^{(z)}), \quad (1)$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \ \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}, \quad (2)$$

where $\boldsymbol{z}_i = [\boldsymbol{x}_i^{\mathrm{T}}, \boldsymbol{y}_i^{\mathrm{T}}]^{\mathrm{T}}$ and M is the number of mixtures.

By using the trained parameters $\hat{\boldsymbol{\lambda}}$, a face feature $\boldsymbol{x}$ can be converted to a voice feature $\hat{\boldsymbol{y}}$ as follows,

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}, \lambda). \quad (3)$$

In GMM-VC, variance and covariance matrices can be cross diagonal because both of input and output features are speech features. However in GMM-based conversion from face to voice, the domain of $\boldsymbol{x}$ is different from that of $\boldsymbol{y}$. Therefore, covariance matrices, $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ have to be full matrices.

3.4. mPCCA-based conversion from face to voice

[22] formulated the probabilities of mPCCA as

$$p(\boldsymbol{x}|\boldsymbol{h}, z_m) = \mathcal{N}(\boldsymbol{U}_m\boldsymbol{h} + \boldsymbol{b}_m, \boldsymbol{\Gamma}_m), \quad (4)$$

$$p(\boldsymbol{y}|\boldsymbol{h}, z_m) = \mathcal{N}(\boldsymbol{V}_m\boldsymbol{h} + \boldsymbol{d}_m, \boldsymbol{\Lambda}_m), \quad (5)$$

$$p(\boldsymbol{h}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad (6)$$

$$p(z_m) = \phi_m, \quad (7)$$

where $\boldsymbol{h}$ and z_m are latent variables. If a parallel corpus $\{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^N$ are given, the parameters of M-mixture mPCCA $\boldsymbol{\Theta}$ are trained to maximize the joint probability as follows,

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{y}_n), \quad (8)$$

where $\boldsymbol{\Theta}=\{\boldsymbol{U}_m, \boldsymbol{b}_m, \boldsymbol{\Gamma}_m, \boldsymbol{V}_m, \boldsymbol{d}_m, \boldsymbol{\Lambda}_m, \phi_m\}_{m=1}^M$. The variance matrices $\boldsymbol{\Gamma}_m$ and $\boldsymbol{\Lambda}_m$ can be diagonal because $\boldsymbol{x}$ and $\boldsymbol{y}$ are not a joint vector. These equations are available when pCCA-based conversion is tested because pCCA is mPCCA in the case of M=1. In Eq. (8), $\boldsymbol{\Theta}$ can be initialized by pCCA parameters $\boldsymbol{\theta}$ which can also be calculated by using Eq. (8), and $\boldsymbol{\theta}$ can be initialized by deterministic values proposed in [21].

By using the trained parameters $\hat{\boldsymbol{\Theta}}$, a face feature $\boldsymbol{x}$ can be converted to a voice feature $\hat{\boldsymbol{y}}$ as follows,

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \ln p(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{\Theta}}). \quad (9)$$

Eq. (9) can be solved by EM algorithm and in the M-step of Eq. (9) $\hat{\boldsymbol{y}}$ can be updated as follows,

$$\hat{\boldsymbol{y}} = \left(\sum_m \boldsymbol{G}_m\right)^{-1} \sum_m \boldsymbol{G}_m \left(\boldsymbol{\psi}_m + \boldsymbol{d}_m\right), \quad (10)$$

$$\boldsymbol{G}_m = \left\langle z_m|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old} \right\rangle \boldsymbol{\Lambda}_m^{-1}, \quad (11)$$

$$\boldsymbol{\psi}_m = \boldsymbol{V}_m \left\langle \boldsymbol{h}|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}, z_m \right\rangle, \quad (12)$$

where $\langle\cdot\rangle$ denotes an expectation and $\hat{\boldsymbol{y}}^{old}$ means $\hat{\boldsymbol{y}}$ in the previous M-step. Although $\hat{\boldsymbol{y}}^{old}$ can be initialized by values based on GMM as in [22], in order to consider the latent variables explicitly, we propose the initialization of $\boldsymbol{G}_m$ and $\boldsymbol{\psi}_m$ based on expectations of $\boldsymbol{x}$, $\boldsymbol{h}$, and z_m as follows.

$$\boldsymbol{G}_m = \langle z_m|\boldsymbol{x}\rangle \boldsymbol{\Lambda}_m^{-1}, \quad (13)$$

$$\boldsymbol{\psi}_m = \boldsymbol{V}_m \langle \boldsymbol{h}|\boldsymbol{x}, z_m\rangle, \quad (14)$$

This initialization was inspired by a deterministic method in [22].

4. Collection of face-to-voice mappings

At fist, we prepared a corpus of face images of Japanese male adults and another corpus of Japanese male voices independently. Then, for the conversion experiments, subjective mappings from faces to voices were collected from 17 Japanese male adult subjects. They were asked to select the voice best suited subjectively to the impression of a given face. The voices used are those of 73 speakers selected from the JNAS database [23]. The face images are those of 71 persons, which are provided by the Morishima laboratory of Waseda University. In order to simplify the experimental procedure, a binary eigenvoice-based clustering tree was built for the voices in advance, and it was used to select the best voice for a given face efficiently.

As a result, it was revealed that the relationship between faces and voices was not one-to-one, but many-to-one. Each subject selected one specific speaker for various faces and the specific speaker depends on subjects. For example, one subject selected one speaker for 24 faces. This many-to-one relationship supports the finding obtained in [12] about human performance of identity matching. The authors said that some people look and sound more similar than others.

5. Experiments

In this section, we conducted three experiments about extraction of face features, extraction of voice features, and statistical conversion from the former features to the latter features.

5.1. Extraction of face features

5.1.1. Experimental conditions

In the first experiment, two different kinds of face features were extracted by using two feature extractors. To train these two extractors, another huge corpus of face images was used. The corpus is the MORPH, which contains 54,147 face images [24]. For each face image, the location of its face was automatically detected by OpenCV[1] and the detected faces were used for training the two extractors. In the case of IFFs, they were converted from the 68 face landmarks, which were extracted with CLNF implemented in OpenFace [25]. The obtained IFFs were compressed through PCA. In the case of CVAE, the encoder and decoder were implemented using four and five layers' Convolutional Neural Networks (CNNs), respectively. 50,000 face images and 4,147 face images were used for training and validation, respectively. We used a latent variable $\boldsymbol{\mu}$ of CVAE as a face feature which denotes the mean vector of the posterior distribution. It characterizes an input image. The dimension of $\boldsymbol{\mu}$, namely d_μ, was 3 or 10. The method for optimization was the Adam [26], the batchsize was 100, the activation function of the last layer of encoder was the identity function, that of the last layer of decoder was sigmoid function, and that of the other layers was the ReLU [27].

5.1.2. Experimental results

In the case of IFFs, as shown in Figure 2, face landmarks can be detected correctly from a given face image, which is abstracted moderately to be an icon image based on IFFs. The cumulative contribution ratio of PCA was over 95% by using only the first

[1] `http://opencv.jp/` [Accessed 19 January 2017]

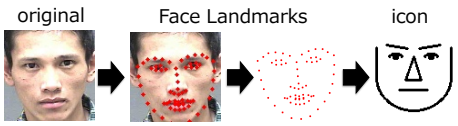


Figure 2: *Face landmarks and the icon image based on IFFs*

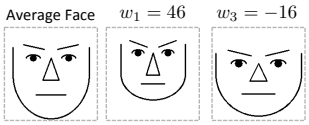


Figure 3: *Face icons with different values of the i-th principal component w_i*

to the third components. Figure 3 shows several examples of face icons that are generated by changing the values of principal components. More examples are available at `https://goo.gl/4NCvFQ`.

In the case of CVAE, although d_μ was smaller than the values adopted in previous works [15, 16], it was experimentally confirmed that CVAE-based reconstructed images preserved facial identities well even if d_μ=3. In order to examine how various face images can be generated by changing the values of $\boldsymbol{\mu}$, the following variation of $\boldsymbol{\mu}$ was tested,

$$\boldsymbol{\mu}^{(i)} = [0, 0, ..., 0, \hat{\mu_d}^{(i)}, 0, ..., 0], \tag{15}$$

$$\hat{\mu_d}^{(i)} = \mu_d^{(min)} + i \times (\mu_d^{(max)} - \mu_d^{(min)})/10, \tag{16}$$

where i=0, 1, ..., 10 and $\mu_d^{(max)}$ and $\mu_d^{(min)}$ were the maximum and minimum values of the d-th dimension in the validation set, respectively. Figure 4 shows the results when d_μ=3 and it is easily found that, by manipulating the value of $\boldsymbol{\mu}$, various face images can be generated.

From the above, the principal components of IFFs (PC-IFF) and $\boldsymbol{\mu}$ of CVAE (CVAE-$\boldsymbol{\mu}$) can be regarded as face features that represent facial impressions.

5.2. Extraction of voice features

5.2.1. Experimental conditions

In the second experiment, eigenvoices were extracted through PCA performed over 127 supervectors of 256 mixture SD-GMMs. SD-GMMs of a speaker of the JNAS database were trained by using the 1st to 24th dimensional Mel-cepstrum coefficients and their delta features of his speech samples, which were digitized at 16 kHz and 16 bits. The coefficients were extracted by using WORLD [28] and SPTK[2]. In order to examine how various voice qualities can be generated by changing the values of eigenvoices, speech samples were synthesized by using the framework of EVC.

5.2.2. Experimental results

The cumulative contribution ratio was over 50% by using the first to the 6-th eigenvoices and over 80% by using the first up to the 33rd eigenvoices. It was confirmed that by manipulating the values of eigenvoice, various voice qualities can be generated. Therefore, eigenvoices can be regarded as voice features that represent vocal impressions.

[2] `http://sp-tk.sourceforge.net/` [Accessed 19 January 2017]

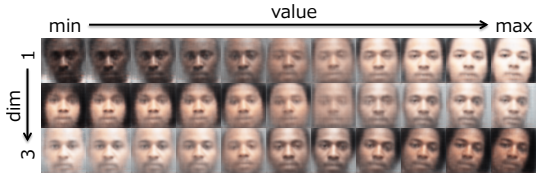


Figure 4: *Various face images generated when d_μ=3*

Table 1: *The mean Mel-cepstrum distortion [dB] over subjects and faces: The number in parentheses denotes the dimension of face features.*

	PC-IFF(3)	CVAE-$\boldsymbol{\mu}$(3)	CVAE-$\boldsymbol{\mu}$(10)
GMM	**1.89**	**1.88**	2.24
pCCA	1.90	1.89	**1.96**
mPCCA	1.98	1.98	2.12

5.3. Statistical conversion from face to voice

5.3.1. Experimental conditions

In the last experiment, PC-IFF or CVAE-$\boldsymbol{\mu}$ was converted statistically to 6-dimensional eigenvoice based on 2-mixture GMM, pCCA, or 2-mixture mPCCA. The dimension of PC-IFF was 3 and d_μ was 3 or 10. The dimension of $\boldsymbol{h}$ of Eq. (6) was 3. Statistical models were trained by using subject-dependent 64 paired data of Section 4. By using the remaining 7 paired data, comparisons were made between automatically converted eigenvoices and manually specified eigenvoices. The comparisons were based on mean Mel-cepstrum distortion (MCD) performed over 53 speeches generated through EVC. The MCD was calculated as follows,

$$\mathrm{MCD\ [dB]} = 10/\ln 10 \sqrt{2\sum_{d=1}^{24}\left(c_d^{(s)} - c_d^{(m)}\right)^2}, \qquad (17)$$

where $c_d^{(s)}$ and $c_d^{(m)}$ were d-th Mel-cepstrum coefficient based on statistical conversion and manual selection, respectively.

5.3.2. Experimental results

As shown in Table 1, by comparing MCDs for each kind of face feature, CVAE-$\boldsymbol{\mu}$ when d_μ=3 is slightly better than PC-IFF, and both of them are better than CVAE-$\boldsymbol{\mu}$ when d_μ=10. As shown in Section 5.1, CVAE is better than IFF in that various photographic face images can be generated. However, since PC-IFF features are obtained after PCA, the function of each dimension of PC-IFF can be interpreted. This may make it possible to examine what face features play more important roles in face-to-voice conversion. This is one of our future works.

Next, we compare the conversion performance of the three models. In the case of CVAE-$\boldsymbol{\mu}$ with d_μ=10, pCCA is better than GMM and mPCCA. On the other hand, in the case of PC-IFF and CVAE-$\boldsymbol{\mu}$ with d_μ=3, GMM is slightly better than pCCA, and both of them are better than mPCCA. These results are considered to be because the size of the collected mappings is not sufficient to train GMM and mPCCA effectively, especially in the case of CVAE-$\boldsymbol{\mu}$ with d_μ=10. However, we obtained the following finding from results of analysis. Although the same number of mixtures, M, is assumed in mPCCA between face features and voice features, analytical results show that the optimal number of mixtures depends on the domain

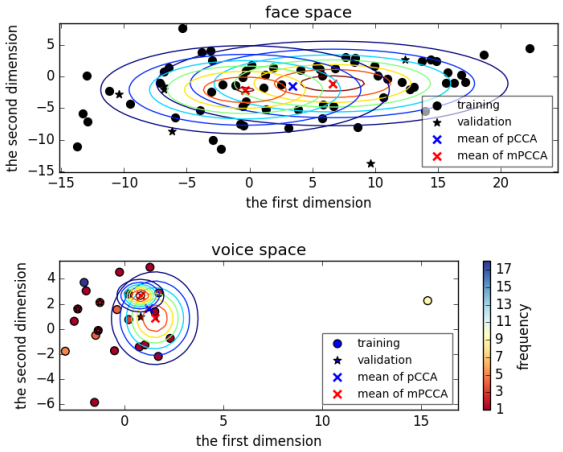


Figure 5: *One mean of pCCA and two means of mPCCA and two Gaussian distributions estimated by the mPCCA. The top is for PC-IFF face features and the bottom is for eigenvoice.*

of features. For example, as shown in Figure 5, in the case of mPCCA (M=2), it is observed that, for voice features, one large distribution seems good enough, but that two large distributions are needed for face features. As future work, we are interested in introducing domain- or feature-dependent number of mixtures for mPCCA, which will characterize the relations between the two spaces more adequately.

6. Conclusion

In this paper, we tested six methods of statistical conversion from face features to voice features based on their subjective impressions. As for face features, two different kinds of features were used. One was IFF, which represents the layout of facial parts, and the other was CVAE, which can embed a photographic face image to a latent feature vector. As for voice features, eigenvoice was used, which represents speaker characters. To realize statistical conversion, manual face-to-voice mappings were collected from Japanese adult subjects. Based on the subjective mappings, PC-IFF or CVAE-$\boldsymbol{\mu}$ was converted statistically to eigenvoice based on GMM, pCCA, or mPCCA. Results of manual mappings implied that each subject has his *favorite* voice and various faces were mapped to that specific voice. This result supported the finding obtained in [12] about identity matching where the difficulty in matching a person's face to its own voice is dependent on the person. As experimental results of statistical conversion, it was found that CVAE-$\boldsymbol{\mu}$ with d_μ=3 is slightly better than PC-IFF, and it was also suggested that the optimal number of mixtures of face features is different from that of voice features.

In future works, we will collect a larger number of manual mappings of face to voice and examine the effect of the size of data on conversion performance. We should also examine what face features are important in face-to-voice conversion, and how the optimal numbers of mixtures of mPCCA are dependent on the two domains and the features used for those domains.

7. Acknowledgements

The authors would like to thank Prof. S. Morishima of Waseda University, Japan, for permission to use his collection of Japanese face images.

8. References

[1] J. Beskow, K. Elenius, and S. McGlashan, "Olga-a dialogue system with an animated talking agent," in *Proceedings of the Fifth European Conference on Speech Communication and Technology*, 1997.

[2] K. Isbister, H. Nakanishi, T. Ishida, and C. Nass, "Helper agent: Designing an assistant for human-human interaction in a virtual meeting space," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2000, pp. 57–64.

[3] J. Rickel and W. L. Johnson, "Task-oriented collaboration with embodied agents in virtual worlds," *Embodied Conversational Agents*, 2000.

[4] H. Morton and M. A. Jack, "Scenario-based spoken interaction with virtual agents," *Computer Assisted Language Learning*, vol. 18, no. 3, pp. 171–191, 2005.

[5] S. Miyake and A. Ito, "A spoken dialogue system using virtual conversational agent with augmented reality," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1–4.

[6] A. Lee, K. Oura, and K. Tokuda, "Mmdagent—a fully open-source toolkit for voice interaction systems," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8382–8385.

[7] E. André and C. Pelachaud, *Interacting with Embodied Conversational Agents*. Boston, MA: Springer US, 2010, pp. 123–149.

[8] R. M. Krauss, R. Freyberg, and E. Morsella, "Inferring speakers ' physical attributes from their voices," *Journal of Experimental Social Psychology*, vol. 38, no. 6, pp. 618 – 625, 2002.

[9] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "'putting the face to the voice': Matching identity across modality," *Current Biology*, vol. 13, no. 19, pp. 1709 – 1714, 2003.

[10] L. Lachs and D. B. Pisoni, "Crossmodal source identification in speech perception," *Ecological Psychology*, vol. 16, no. 3, pp. 159–187, 2004.

[11] L. W. Mavica and E. Barenholtz, "Matching voice and face identity from static images," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 2, pp. 307–312, 2013.

[12] H. M. J. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, "Matching novel face and voice identity using static and dynamic facial images," *Attention, Perception, & Psychophysics*, vol. 78, no. 3, pp. 868–879, Apr 2016.

[13] T. Baltrušaitis, P. Robinson, and L. P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *2013 IEEE International Conference on Computer Vision Workshops*, Dec 2013, pp. 354–361.

[14] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, 2014, pp. 3581–3589.

[15] R. Yeh, Z. Liu, D. B. Goldman, and A. Agarwala, "Semantic Facial Expression Editing using Autoencoded Flow," *ArXiv e-prints*, 2016.

[16] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 1133–1141.

[17] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, Nov 2000.

[18] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, 2007, pp. IV–1249–IV–1252.

[19] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, Nov 2007.

[20] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[21] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

[22] H. Uchida, D. Saito, and N. Minematsu, "Acoustic-to-articulatory mapping based on mixture of probabilistic canonical correlation analysis," *Interspeech*, 2017.

[23] "Japanese newspaper article sentences: Jnas," http://research.nii.ac.jp/src/JNAS.html.

[24] K. Ricanek and T. Tesafaye, "Morph: a longitudinal image database of normal adult age-progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, April 2006, pp. 341–345.

[25] T. Baltrušaitis, P. Robinson, and L. P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–10.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15, 2011, pp. 315–323.

[28] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.