Interspeech 2018
2-6 September 2018, Hyderabad

Audio-visual voice conversion using deep canonical correlation analysis for deep bottleneck features

Satoshi TAMURA[1], *Kento HORIO*[1], *Hajime ENDO*[1], *Satoru HAYAMIZU*[1], *Tomoki TODA*[2]

[1]Gifu University, Japan [2]Nagoya University, Japan

tamura@info.gifu-u.ac.jp

Abstract

This paper proposes Audio-Visual Voice Conversion (AVVC) methods using Deep BottleNeck Features (DBNF) and Deep Canonical Correlation Analysis (DCCA). DBNF has been adopted in several speech applications to obtain better feature representations. DCCA can generate much correlated features in two views, and enhance features in one modality based on another view. In addition, DCCA can make projections from different views ideally to the same vector space. Firstly, in this work, we enhance our conventional AVVC scheme by employing the DBNF technique in the visual modality. Secondly, we apply the DCCA technology to DBNFs for new effective visual features. Thirdly, we build a cross-modal voice conversion model available for both audio and visual DCCA features. In order to clarify effectiveness of these frameworks, we carried out subjective and objective evaluations and compared them with conventional methods. Experimental results show that our DBNF- and DCCA-based AVVC can successfully improve the quality of converted speech waveforms.

Index Terms: statistical speech conversion, audio-visual processing, deep learning, bottleneck feature, canonical component analysis.

1. Introduction

Voice Conversion (VC) is a technique to convert speech waveforms pronounced by a source speaker into those of a target speaker [1]. Considering to develop VC applications in real environments, noise-robust techniques are essential, just like speech recognition. Some of authors have already proposed an Audio-Visual Speech Recognition (AVSR) method, exploiting a visual modality i.e. lip images, in order to ensure acoustic noise robustness [2, 3]. According to the success of AVSR, Audio-Visual Voice Conversion (AVVC) methods employing the similar technique as AVSR have been proposed [4, 5, 6]. In addition, there are some related works focusing on speech synthesis from visual cues [7, 8]. We tested our approach [6] in noisy situations, and found the AV technology is also useful for VC.

Recently, deep learning has attracted attentions in speech processing fields. In VC, the most simplest way to employ the technology is to replace the statistical model into a Deep Neural Network (DNN); a deep-learning conversion model is made in which its input layer corresponds to acoustic features of a source speaker, and the output layer corresponds to features for a target speaker. In this case, some networks such as deep belief network and deep bidirectional long short-term memory are chosen as a conversion model [9, 10]. Since these schemes require a huge training data set, there is another strategy to use DNNs for feature extraction: Deep BottleNeck Features (DBNFs). Deep bottleneck features are proposed to extract much more powerful acoustic features compared to conventional features in speech processing such as speech recognition. A DNN having a bottleneck layer, on which there are relatively fewer perceptrons, is built using a large corpus. After model training, output values from every units on the bottleneck layer are composed into a new feature set. We have applied this technique to speech recognition, lipreading and AVSR [2, 3]. In order to obtain noise-robust acoustic DBNFs, speech data overlapped with noise signals are used for DNN training. A visual DNN is also made to obtain effective visual representations.

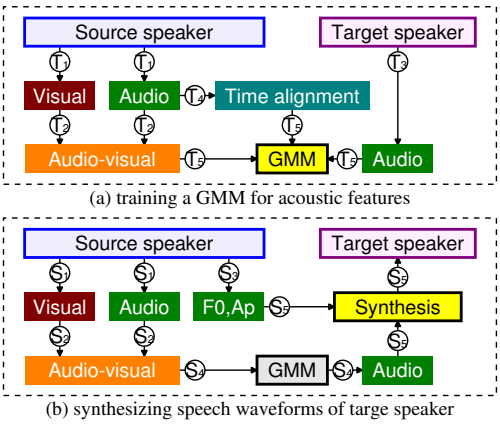


Figure 1: *Audio-visual speech conversion.*

This paper firstly proposes a new AVVC method, using the DBNF technique. Because applying the technique we can obtain better visual features, the VC quality is expected to be improved in noisy environments. In addition, we try to further improve the method by employing Deep Canonical Correlation Analysis (DCCA) [11]. DCCA is a non-linear extension of CCA, which projects vectors obtained from two views to have a higher correlation score. We utilize this technique to obtain better feature representation. Finally, we also try to propose a cross-modal VC technique based on the DCCA technology. Since features from two modalities are ideally projected into one single vector space, one common cross-modal conversion model is now available for both audio and visual modalities. We compare these methods with conventional VC schemes using our AV data, by means of subjective and objective evaluation.

2. Voice Conversion

We briefly summarize the statistical audio-only VC method [1], as well as our audio-visual VC method [6]. Because the AVVC approach is based on the audio-only VC framework, we would like to introduce the AVVC scheme depicted in Figure 1.

For an i-th pair (an audio-visual movie of source speaker and a corresponding speech waveform of target speaker) in a training data set, an audio feature $\boldsymbol{a}_{i,t}$ and a visual feature $\boldsymbol{v}_{i,t}$ of source speaker (t is a frame index) are extracted ($\mathbf{T}_1$). An audio-visual feature vector is then obtained by concatenating these vectors frame by frame ($\mathbf{T}_2$):

$$\boldsymbol{x}_{i,t} = (\boldsymbol{a}_{i,t}{}^{\top}, \boldsymbol{v}_{i,t}{}^{\top})^{\top} \quad (1)$$

10.21437/Interspeech.2018-2286

This integration is skipped in the audio-only VC, that is:

$$\boldsymbol{x}_{i,t} = \boldsymbol{a}_{i,t} \tag{2}$$

An acoustic feature of the target waveform, Mel CEPstral coefficients (**MCEP**) denoted by $\boldsymbol{y}_{i,t}$, is extracted ($\mathbf{T}_3$). A frame-level time alignment L_i between $X_i = (\boldsymbol{x}_{i,t})$ and $Y_i = (\boldsymbol{y}_{i,t})$ is subsequently obtained applying a dynamic time warping technique ($\mathbf{T}_4$). A cross-speaker Gaussian Mixture Model (GMM) is finally built using all pairs (X_i, Y_i) and corresponding labels L_i, in which a joint probability $p(\boldsymbol{x}, \boldsymbol{y})$ can be computed ($\mathbf{T}_5$).

When synthesizing speech signals of target speaker, audio and visual features are generated ($\mathbf{S}_1$), followed by feature concatenation to obtain source features $X = (\boldsymbol{x}_t)$, according to Eq.(1) or Eq.(2) ($\mathbf{S}_2$). F0 and aperiodic data which are necessary for speech synthesis are simultaneously obtained from source speeches ($\mathbf{S}_3$). Applying the GMM, we can estimate audio features of target speaker, denoted by $\hat{Y} = (\hat{\boldsymbol{y}}_t)$, according to the following equation ($\mathbf{S}_4$):

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, p(\boldsymbol{y}|\boldsymbol{x}_t) \tag{3}$$

Finally the voice conversion is done using $\hat{Y}$ as well as the F0 and aperiodic parameters ($\mathbf{S}_5$):

3. Deep learning for AVVC

3.1. Deep bottleneck feature

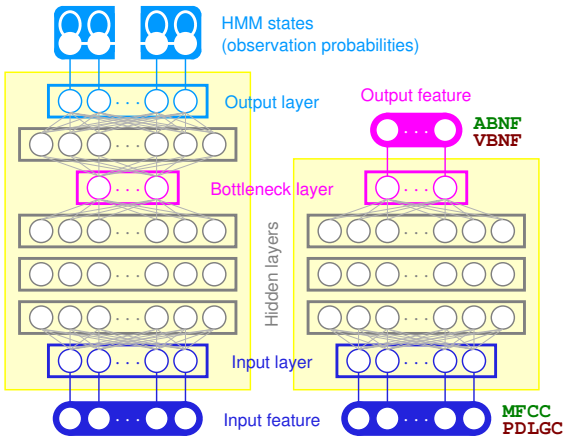
Figure 2 depicts an architecture for DBNFs in [2, 3]. Before training a DNN for DBNF extraction, Hidden Markov Models (HMMs) are built in a conventional manner, in order to get state-level time alignment data. The DNN having a bottleneck layer on which there are relatively few perceptrons compared to the other hidden ones, is then built using a large-scale training data. Its input layer corresponds to training feature vectors, while the output layer corresponds to the state-level time alignment. After completing the training, all the layers beyond the bottleneck layer are removed, so that the bottleneck layer could become a new output layer. We can now apply the DNN to convert input features into new feature vectors, called DBNFs.

In our previous works [2, 3], we exploited this DBNF framework to generate effective audio and visual features for AVSR. Not only clean but also noisy speech data were prepared to give the DNN robustness against acoustic noises. We also collected many kinds of basic visual features: appearance-based cues such as Principal Component Analysis (**PCA**), Discrete Cosine Transform (**DCT**), Linear Discriminant Analysis (**LDA**) and our original feature (**GIF**) [12], in addition to one shape-based parameters having COORDinates of lip contours (**COORD**). We then concatenate these features into an input vector (**PDLGC**) of another DNN (see also Figure 4). We then adopted DBNFs, i.e. **ABNF** and **VBNF**, to make audio-visual features.

3.2. Deep canonical correlation analysis

CCA is a technique to make projections from two modalities so that the correlation between transformed vectors should be maximized. The most conventional one is the linear CCA. Let us denote vectors in two modalities by $\boldsymbol{a}$ and $\boldsymbol{v}$, respectively. CCA defines two matrices A and V that maximize the cross correlation between $A\boldsymbol{a}$ and $V\boldsymbol{v}$. Similar to PCA, CCA finds a first canonical component corresponding to the first rows of A and V. Second and following canonical components are subsequently obtained subject to existing components.

In order to obtain higher correlation, non-linear analyses such as kernel-based CCA and deep-learning-based CCA have been proposed [11, 13]. In this paper, we focus on the latter one, DCCA. Figure 3 illustrates a framework of DCCA. We assume DNNs each having two hidden layers. In the first modality, one DNN is prepared which transforms $\boldsymbol{a}$ into $\boldsymbol{a}' = \boldsymbol{f}(\boldsymbol{a})$ non-linearly. We have another DNN for the second modality to convert $\boldsymbol{v}$ into $\boldsymbol{v}' = \boldsymbol{g}(\boldsymbol{v})$. All the parameters appearing in $\boldsymbol{f}$ and $\boldsymbol{g}$ are optimized such that a correlation score for training data is maximized like CCA.



(a) DNN training (b) feature extraction

Figure 2: *Deep bottleneck feature.*

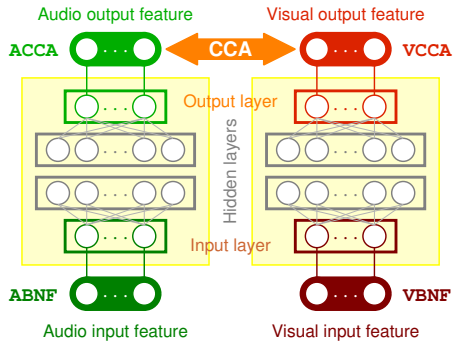


Figure 3: *Deep canonical correlation analysis.*

4. Proposed AVVCs

4.1. Method 1 — visual DBFs

We had chosen **PCA** as a visual feature in our previous AVVC scheme [6]. In order to improve the visual feature set, in this work, we employ the DBNF architecture in the visual modality. Since it has been proven that exploiting visual DBNFs significantly improves lipreading recognition accuracy (**PCA**: 42.52% $\rightarrow$ **VBNF**: 73.66%) [3], it is expected that adopting **VBNF** can also enhance the quality of AVVC. **MCEP** and **VBNF** vectors are finally concatenated frame by frame into an audio-visual feature vector. Table 1 summarizes features and integration schemes used in this paper.

4.2. Method 2 — DCCA-based visual features

We further try to improve visual features in AVVC by adjusting **VBNF**. After computing audio and visual DBNFs, i.e. **ABNF** and **VBNF**, DCCA is applied to get new features, i.e. **ACCA** and **VCCA** respectively. In this method, **VCCA** is employed as a new visual feature set while **MCEP** is still used, because of visual feature comparison.

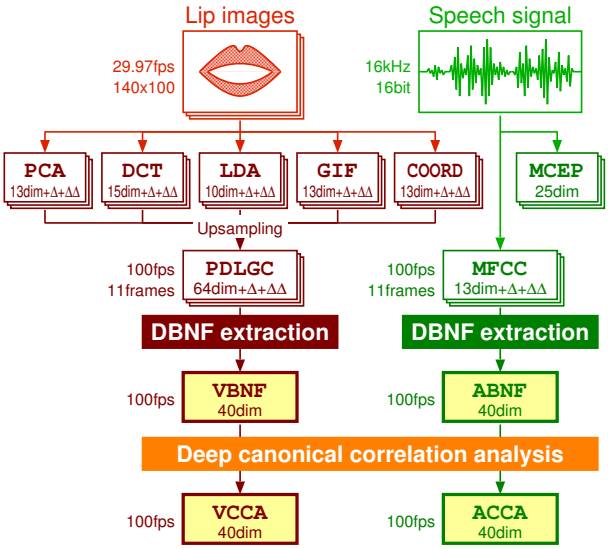


Figure 4: *Feature extraction in our AVVC.*

Table 1: *Audio and visual features/model in AVVC methods.*

	Feature		AV integration
	Audio	Visual	
Conv. [6]	MCEP	PCA	feature concatenation
Method 1	MCEP	VBNF	feature concatenation
Method 2	MCEP	VCCA	feature concatenation
Method 3	ACCA	VCCA	cross-modal model

4.3. Method 3 — a cross-modal model for DCCAs

As an extension of the last method, we can choose **ACCA** instead of **MCEP**, with composing **ACCA** and **VCCA** as an audio-visual feature vector. Such the scheme may be able to improve the conversion quality, on the other hand, we can adopt another strategy for these features. DCCA generates non-linear transformations for original audio and visual features so that projected features could have high cross correlation. That means DCCA ideally projects audio and visual vectors into one single space. In this method, we train a GMM using **VCCA** feature vectors. Not only **VCCA** but also **ACCA** vectors can be then converted to acoustic features of target speaker using this cross-modal GMM.

In existing AVSR systems, an audio-visual balancing architecture is widely employed; for example, multi-stream HMMs which can emphasize an audio or a visual modality according to stream weight factors are often adopted. Since DCCA may transform features in both modalities into the same vector space as already described, a different and simple fusion technique is chosen in this work. Let us denote **ACCA** and **VCCA** vectors by $\boldsymbol{a}'$ and $\boldsymbol{v}'$ respectively. Here, a new feature vector $\boldsymbol{c}$ can be obtained as:

$$\boldsymbol{c} = \lambda \boldsymbol{a}' + (1 - \lambda)\boldsymbol{v}' \quad (4)$$

where λ is a linear combination parameter ($0 \leq \lambda \leq 1$).

5. Experiment

We carried out subjective and objective evaluations to clarify effectiveness of proposed AVVC frameworks.

5.1. Experimental setup

5.1.1. Data

We chose CENSREC-1-AV [14] consisting of audio-visual connected-digit sentences. To make pre-trained HMMs and DNNs, we used all data in the CENSREC-1-AV training data set including 3,234 utterances by 42 speakers. We randomly selected four speakers as source speakers and extracted their sentences, from the CENSREC-1-AV test data set. We then asked four male subjects, as target speakers, to utter the same digit sequences.

In order to make additional acoustic training data and to evaluate VC methods in acoustically difficult environments, input speeches of source speakers were contaminated by adding several kinds of acoustic noises. We chose in-car noises [14] for DNN training while we obtained various noises from [15] for testing, at several Signal-to-Noise Ratios (SNRs).

Table 2: *DNN setup.*

		DBNF		DCCA	
		Audio	Visual	Audio	Visual
# of units on layer	Input	429	2,112	40	40
	Hidden	2,048	2,048	1,600	1,600
	Bottleneck	40	40	—	—
	Output	179	179	40	40
Pre-training	Batch size	256	256	—	—
	Max epochs	10	10	—	—
	Learning rate	0.004	0.004	—	—
Fine tuning	Batch size	256	256	400	400
	Max epochs	50	50	50	50
	Learning rate	0.006	0.006	0.001	0.001

5.1.2. HMM

We made HMMs for state-level time alignment. HMMs were prepared according to [14]; each digit HMM had 16 states while a silence HMM consisted of 3 states. There were thus 179 states in total. In training, clean and noisy data were exploited [2].

5.1.3. DNN

Table 2 summarizes our DNN setup. We built DNNs for **ABNF** and **VBNF** separately, as [2, 3], also shown in Figure 2; there were five hidden layers, one of which was a bottleneck layer having only 40 perceptrons. Each unit on the output layer corresponded to a pre-trained HMM state. We concatenated features at neighbor frames as an input vector to these DNNs: a 429-dimensional vector for **ABNF** (**MFCC** features from current, 5 previous and 5 next frames) and a 2,112-dimensional vector for **VBNF** (**PDLGC** coefficients $\times$ 11 frames).

Two DNNs for DCCA were also prepared as Figure 3. The audio DNN received an **ABNF** vector followed by generating a 40-dimensional output vector **ACCA**. There were two fully-connected hidden layers in this DNN. The same architecture was applied to the visual modality. Since these DNNs were relatively shallow, only fine-tuning was conducted in this work.

5.1.4. Evaluation

In this paper, a Mel-Cepstrum Distortion (MCD) score was used for objective evaluation. A small MCD score means that the quality of generated speeches is successful. For subjective evaluation, Mean Opinion Score (MOS) was used. We also carried out transcription test; we asked subjects to listen to converted speeches and write down their transcriptions, followed by calculating accuracy of the transcriptions.

5.2. Results for Method 1

We at first evaluated **Method 1**, using 14 types of noises from [15] and white noise. Note that MCD scores of the conventional

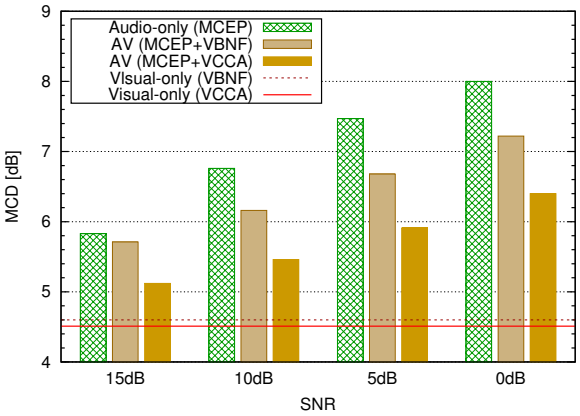


Figure 5: *MCD scores of audio-only, visual-only and audio-visual VCs (**Method 1** and **method 2**).*

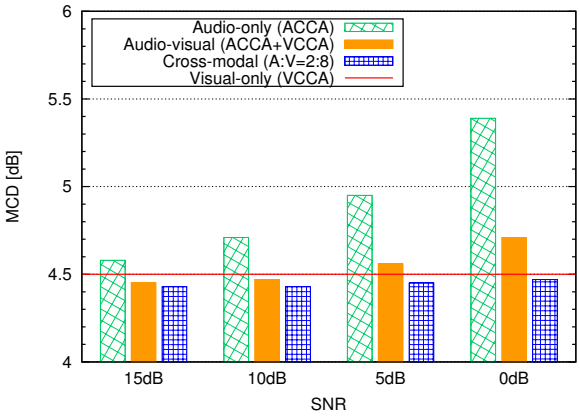


Figure 6: *MCD scores of audio-only, visual-only, audio-visual and cross-modal VCs (**Method 3**) in in-car noises.*

Table 3: *Transcription accuracy at SNR=15dB, for audio-only, conventional and proposed audio-visual VCs (**Method 1**).*

	Stationary	In-crowd	Impact
Audio-only (**MCEP**)	86%	96%	79%
AV (**MCEP**+**PCA**)	92%	97%	80%
AV (**MCEP**+**VBNF**)	93%	99%	92%

Table 4: *Transcription accuracy for audio-visual VCs (**Method 1** and **Method 2**).*

	SNR=15dB	SNR=5dB
AV (**MCEP**+**VBNF**)	93%	77%
AV (**MCEP**+**VCCA**)	94%	92%

Table 5: *MOS scores at SNR=15dB/5dB for audio-visual VCs (**Method 1** and **Method 2**).*

	Stationary	In-crowd	Impact
AV (**MCEP**+**VBNF**)	3.07/2.38	3.01/2.35	2.94/2.31
AV (**MCEP**+**VCCA**)	3.07/2.70	3.39/2.89	3.25/2.82

audio-only VC in clean condition and the visual-only VC using **PCA** were 4.31dB and 5.02dB, respectively.

Figure 5 indicates mean MCD scores among all the noisy conditions at different SNRs. Focusing on visual-only results it is easily found that **VBNF** is better than **PCA**. Furthermore, our AVVC scheme using **VBNF** got better speech quality compared to the conventional audio-only method at all the SNRs. We conducted the subjective evaluation at 15dB. We categorized the noises into three classes "stationary," "in-crowd" and "impact." Table 3 shows transcription test results given by six subjects. Although no significant difference was observed in MOS evaluation, in this test it is found that applying **VBNF** is effective particularly against impact noises.

5.3. Results for Method 2

Next, we evaluated **Method 2**. Figure 5 also includes its results. In both visual-only and audio-visual conditions, we could improve MCD scores compared to **Method 1**. Table 4 shows results of the transcription test obtained from 10 subjects. In the table, we can see that using **VCCA** significantly improved the accuracy at 5dB. Finally Table 5 indicates MOS scores at 15dB and 5dB in the three noise kinds. Except the stationary noise at 15dB, we could improve the scores from **VBNF**.

It is often observed that, even if one modality is only available for training, a unimodal processing in another modality can be improved. Basically the audio modality is much more useful to discriminate phonemes than the visual modality in ideal environments. In DCCA, the audio information thus implicitly refined visual feature extraction, resulting the better VC quality.

5.4. Results for Method 3

Finally we tested **Method 3**. In this evaluation we only adopted two kinds of in-car noises at SNR=15, 10, 5 and 0dB. Based on preliminary experiments, we set $\lambda = 0.2$.

Figure 6 shows MCD scores of the following VC schemes:

1. audio-only VC using **ACCA**
2. visual-only VC using **VCCA**
3. audio-visual VC using **ACCA** and **VCCA**
4. the cross-modal VC **Method 3** utilizing the GMM trained from **VCCA** for linear-combined features obtained from **ACCA** and **VCCA** at $\lambda = 0.2$

As we expected the audio-visual scheme based on **Method 2**, using **ACCA** instead of **MCEP** seemed to work well. Furthermore, it turns out that the cross-modal approach **Method 3** can significantly improve the VC quality even in heavily noisy environments. It is also remarkable that this cross-modal VC achieved slightly better performance than audio-only and visual-only VCs. This means such the audio-visual balancing architecture is quite useful in AVVC, just like AVSR.

6. Conclusion

For audio-visual voice conversion, in this paper, at first we improved visual features by employing the DBNF architecture. We subsequently refined visual representations by introducing the DCCA framework. By the aid of bottleneck network and audio information, the AVVC quality could significantly increase in noisy environments. We further proposed to use a cross-modal AVVC model based on DCCA. Objective and subjective experimental results indicate our approach is quite successful compared to existing VC schemes.

Our future work includes evaluation of our AVVC performance in visually difficult situations and in large-vocabulary tasks, in addition to investigation of automatic optimization of the balancing parameter.

7. Acknowledgments

The authors would like to thank Dr. Karen Livescu at Toyota technological Institute at Chicago (TTIC), for her technological supports about deep canonical correlation analysis.

8. References

[1] 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.2222-2225 (2007).

[2] H.Ninomiya, N.Kitaoka, S.Tamura, Y.Iribe and K.Takeda, "Integration of deep bottleneck features for audio-visual speech recognition," Proc. INTERSPEECH2015, pp.563-567 (2015).

[3] S.Tamura, H.Ninomiya, N.Kitaoka, S.Osuga, Y.Iribe, K.Takeda and S.Hayamizu, "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," Proc. APSIPA ASC 2015, pp.575-582 (2015).

[4] A.Barbulescu, T.Hueber, G.Bailly and R.Ronfard, "Audio-visual speaker conversion using prosody features," Proc. AVSP2013, pp.11-16 (2013).

[5] K.Masaka, R.Aihara, T.Takiguchi and Y. Ariki, "Multimodal voice conversion using non-negative matrix factorization in noisy environments," Proc. ICASSP2014, pp.1561-1565 (2014).

[6] K.Sawada, M.Takehara, S.Tamura and S.Hayamizu, "Audio-visual voice conversion using noise-robust features," Proc. ICASSP2014, pp.7949-7953 (2014).

[7] T.Hueber, E.Benaroya, B.Denbt and G.Chollet, "Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface," Proc. INTERSPEECH 2011, pp.593-596 (2011).

[8] H.Akbari, H.Arora, L.Cao and N.Mesgarani, "LIP2AUD-SPEC: Speech reconstruction from silent lip movements video" Proc. ICASSP2018, pp.2516-2520 (2018).

[9] T.Nakashika R.Takashima, T.Takiguchi and Y. Ariki, "Voice Conversion in high-order eigen space using deep belief nets," Proc.INTERSPEECH2013, pp.369-372 (2013).

[10] L.Sun, S.Kang, K.Li and H.Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," Proc.ICASSP2015, pp.4869-4873 (2015).

[11] G.Andrew, R.Arora, J.Bilmes and K.Livescu, "Deep canonical correlation analysis," Proc. ICML2013, pp.1247-1255 (2013).

[12] N. Ukai, T. Seko, S. Tamura and S. Hayamizu, "GIF-LR: GA-based informative feature for lipreading," Proc. APSIPA ASC 2012, PS.3-IVM.7.5 (2012).

[13] R.Arora and K.Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," Proc. MLSLP2012 (2012).

[14] S.Tamura, C.Miyajima, N.Kitaoka, T.Yamada, S.Tsuge, T.Takiguchi, K.Yamamoto, T.Nishiura, M.Nakayama, Y.Denda, M.Fujimoto, S.Matsuda, T.Ogawa, S.Kuroiwa, K.Takeda and S.Nakamura, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," Proc. AVSP2010, pp.85-88 (2010).

[15] Denshikyo noise databese, `http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE`.