AUDIO-VISUAL VOICE CONVERSION USING NOISE-ROBUST FEATURES

Kohei Sawada, Masanori Takehara, Satoshi Tamura and Satoru Hayamizu

Department of Engineering, Gifu University
1-1 Yanagido, Gifu, Gifu, 501-1193 Japan

ABSTRACT

Voice Conversion (VC) is a technique to convert speech data of source speaker into ones of target speaker. VC has been investigated and statistical VC is used for various purposes. Conventional VC uses acoustic features, however, the audio-only VC has suffered from the degradation in noisy or real environments. This paper proposes an Audio-Visual VC (AVVC) method using not only audio features but also visual information, i.e. lip images. Eigenlip feature is employed in our scheme as visual feature. We also propose a feature selection approach for audio-visual features. Experiments were conducted to evaluate our AVVC scheme comparing with audio-only VC, using noisy data. The results show that AVVC can improve the performance even in noisy environments, by properly selecting audio and visual parameters. It is also found that visual VC is also successful. Furthermore, it is observed that visual dynamic features are more effective than visual static information.

Index Terms— voice conversion, audio-visual processing, noise robustness, feature selection.

1. INTRODUCTION

Voice Conversion (VC), namely speech conversion, is a technique to convert speech signals of source speaker into ones of target speaker [1]. VC is expected in various applications, and one of them is for handicapped people who cannot speak speech due to laryngectomy. By VC techniques, their electrolaryngeal speech can be converted into normal speech signals [2]. In order to develop such a system, VC must be available on mobile devices in real environments. However, there is a crucial issue that the quality of converted speech heavily decreases particularly in the situation where noise exists. To enhance the noise robustness, we have proposed new acoustic feature and succeeded to increase the quality in noisy conditions [3]. As the other noise-robust method, VC using Non-negative Matrix factorization (NMF) has been investigated [4].

For speech recognition in noisy environments, multi-modal Automatic Speech Recognition (ASR) has been investigated [5-7]. It utilizes not only speech signals but also additional information obtained from the other modality. A visual information e.g. lip pictures is typically employed as another modality. Since visual infor-mation is not affected by audio noise, the recog-nition performance can be improved in real conditions.

From these viewpoints, we focus on Audio-Visual Voice Conversion (AVVC). In the paper [8], speech conversion was conducted and facial animation was generated using audio-visual features of source speakers. In this paper, we further concentrate on the noise robustness of AVVC, in order to build a noise-robust VC system in real environments. In particular, feature extraction methods are examined. Experiments were conducted using audio-visual parallel data with noises, to compare audio-only and visual-only VC approaches as well as to evaluate the robustness of the feature extraction and AVVC itself.

This paper is organized as follows: Conventional VC techniques are introduced in Section 2. Section 3 describes our AVVC method, particularly feature extraction. We conducted evaluation experiments in Section 4. Finally Section 5 concludes this paper.

2. VOICE CONVERSION

The section summarizes a statistical VC method using conventional audio-only features [1]. A cross-speaker model is firstly built from training data. Secondly, using the model, output acoustic features of target speaker are obtained from input features of source speaker. Finally speech signal is generated using estimated parameters. Figure 1 illustrates the procedure.

Before training, the same sentences spoken by source and target speakers are collected as a training data set. Feature vectors are extracted for each utterance pair made by source and target speakers (**T1, T3**). Let us denote a source feature by $\boldsymbol{X}_t$ and a target feature by $\boldsymbol{Y}_t$, where t indicates time. Mel-CEPstrum (MCEP) coefficients are often employed as input and output acoustic features. Frame-level time alignments between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ are then obtained applying dynamic time warping technique (**T4**). A cross-speaker model is subsequently built using these features, in which a joint probability $p(\boldsymbol{X}_t, \boldsymbol{Y}_t)$ is computed (**T5**). GMM (Gaussian Mixture Model) is typically employed as the model. In the training, the following equation is applied:

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

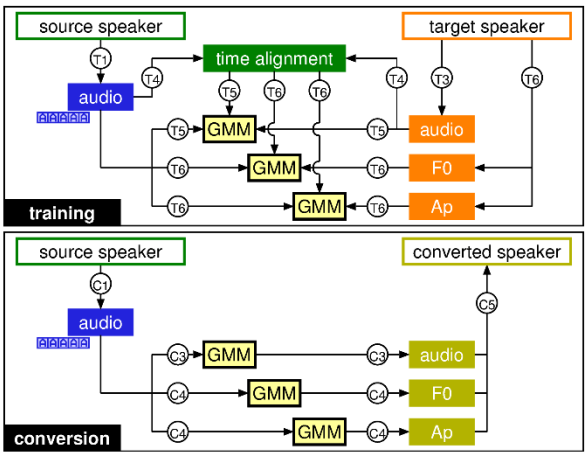

Fig. 1. Conventional audio-only voice conversion.

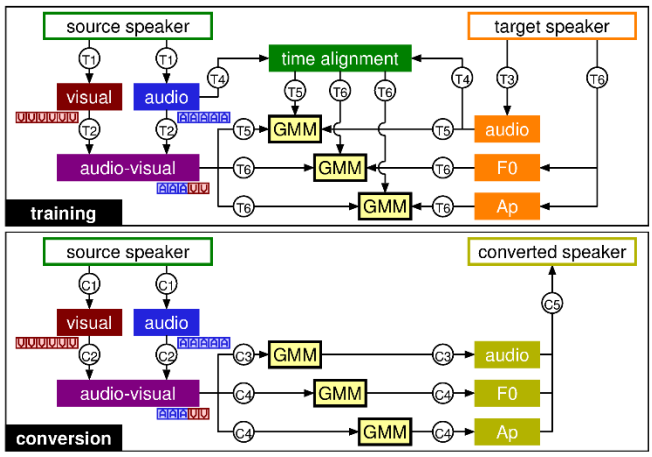

Fig. 2. Proposed audio-visual voice conversion.

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \prod_{t} p(\boldsymbol{X}_t, \boldsymbol{Y}_t|\lambda) \tag{1}$$

where λ denotes model parameters. F0 and aperiodic for speech generation are also extracted from target speech data and modeled by GMM (**T6**).

A source feature sequence $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top\right]^\top$ is computed from input speech signals (**C1**), where $\top$ indicates transposition of a vector. Using the trained GMM, output acoustic features are computed based on maximum likelihood estimation (**C3**); the output sequence $\hat{\boldsymbol{Y}}$ can be estimated as:

$$\hat{\boldsymbol{Y}} = \operatorname{argmax} P(\boldsymbol{Y}|\boldsymbol{X}, \lambda) \tag{2}$$

where $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_T^\top\right]^\top$ is a target sequence. Finally the voice conversion is done using the obtained output features as well as F0 and aperiodic parameters (**C4**, **C5**).

3. AUDIO-VISUAL VOICE CONVERSION

In this paper, we employ audio-visual features as input source features in the statistical VC described in Section 2. In audio-visual speech recognition / voice activity detection [9,10], visual features contribute to increase the robustness against acoustic noises. Therefore visual features are also expected to play the same role in VC. AVVC has another benefit. Larygectomees have a problem that the quality of electrolaryngeal speech is quite low. Since visual features are not affected, it is possible for the laryngectomees to acquire more clean speech, converting the electrolaryngeal speech.

3.1. AVVC summary

Basically, our proposed AVVC method has the similar framework as the audio-only VC method in Section 2. Figure 2 depicts our AVVC training and conversion techniques. Model training is conducted as below:

T1. Extract audio and visual features of source speaker.
T2. Combine both audio and visual features obtained in **T1** into an audio-visual feature.
T3. Extract audio features of target speaker.
T4. Obtain time alignments for the audio features of both speakers in **T1** and **T3.**
T5. Using the features in **T2** and **T3** as well as the alignments in **T4**, build a GMM.
T6. Make F0 and aperiodic GMMs in the same way.

Voice conversion is performed as follows:

C1. Extract audio and visual features of source speaker.
C2. Combine both audio and visual features obtained in **C1** into an audio-visual feature.
C3. Using the audio-visual features in **C2** and the model in **T5**, estimate the output feature for target speaker.
C4. Estimate F0 and aperiodic parameters using features in **C2** and GMMs in **T6**.
C5. Generate converted speech from estimated parameters in **C3** and **C4**.

The difference from the audio-only VC is adding visual features to the input feature vectors for source speaker. Note that visual information of target speaker is not necessary in this framework. In the rest of this section, audio and visual feature extractions and feature combination are explained.

3.2. Feature extraction

3.2.1. Audio feature

For source features, Fast Fourier Transform (FFT) is simply applied in order to reduce computational consumption. In contrast, for target features, STRAIGHT [11] is used to obtain high-resolution components, because the resolution of output features directly affects the quality of converted speech. Subsequently, mel-cepstral input features $\boldsymbol{X}_{At}$ and output features $\boldsymbol{Y}_t$ are computed.

3.2.2. Visual feature

In our AVVC, an eigenlip feature is employed as a visual parameters. The calculation of visual feature is based on the CENSREC-1-AV baseline method [12]. Principal component analysis is applied to lip images in a training data set in order to obtain eigenvectors (eigenlips). Using the eigenvectors, we can get eigenvalue components for each lip image. Because the visual frame rate is lower than the acoustic one, feature interpolation (upsampling) is conducted to the components so as to synchronize the feature rate. In this process, three-dimensional spline interpolation is adapted. Dynamic feature values are also obtained. Consequently, an input visual feature vector $\boldsymbol{X}_{Vt}$, consisting of interpolated eigenvalue components as well as their derivatives, is completed.

3.2.3. Combination of audio and visual features

Audio and visual features are concatenated into an audio-visual feature:

$$\boldsymbol{X}_t = \left(\boldsymbol{X}_{At}{}^{\top}, \boldsymbol{X}_{Vt}{}^{\top}\right)^{\top} \quad (3)$$

Higher dimension might cause the performance degradation in VC; it becomes difficult to estimate output features because the feature space is too wide due to the higher dimension [13].

Therefore, feature selection is proposed in our AVVC scheme. In both audio and visual features, parameters in lower dimensions are more informative than those in higher dimensions. Thus we extract subvectors from lower parts of original audio and visual features respectively, and afterwards the concatenation is performed. For visual features, static and dynamic parameter selection is also considered.

4. EXPERIMENT

Experiments were conducted to evaluate our method, comparing with audio-only and visual-only VCs. The feature selection scheme is also investigated.

4.1. Experimental condition

Table 1 indicates experimental setup. Target data were chosen from an audio-visual corpus CENSREC-1-AV [12], whereas speech data for source speaker were recorded in clean condition. Note that CENSREC-1-AV includes 42 training speaker, however, read texts are different by speakers. In order to collect parallel data, we recorded the utterances of which sentences are the same for one speaker in the database. Two kinds of noises were prepared, white noises at SNR=15dB and SNR=10dB, to disturb audio features of source speaker. Note that GMMs used in the experiments were trained under the clean condition.

Each audio feature had 25 mel-cepstral coefficients, whereas visual feature was prepared consisting of 10 eigenvalues with 10 Δ and 10 ΔΔ parameters. We tested 10 feature conditions, as shown in Table 2. The dimension was fixed as 25 among all the conditions. **A1** corresponds to conventional audio-only VC. In **AV1**-**AV3**, 15-dimensional audio feature and 10-dimensional visual feature were combined. The visual feature included either of static, Δ, and ΔΔ coefficients. Similarly, 5-dimensional audio and 20-dimensional visual features were used in **AV4**-**AV6**. In these cases, two of the three components (static, Δ, and ΔΔ) were chosen. In **V1**-**V3**, visual-only VCs were performed. To adjust the feature size, five coefficients were removed from the original visual features.

Table 1. Experimental setup.

Data	Task	Japanese continuous digit
	Source spkr	1 male (recorded)
	Target spkr	1 male (CENSREC-1-AV)
	# utterances (train)	66
	# utterances (test)	10
Audio	Sample data [Hz]	16,000
	Frame size [msec]	5
	Frame length [msec]	5
Visual	Frame rate [fps]	30 (source), 29.97 (target)
	Image size	40×27 (1,080 dim)
	# pictures (PCA)	4,620
GMM	# mixtures	16 (F0), 32 (others)

Table 2. Feature conditions.

		A1	**AV1**	**AV2**	**AV3**	**AV4**	**AV5**	**AV6**	**V1**	**V2**	**V3**
Audio feature (MCEP)		25	15			5					
Visual feature	static		10			10		10	10	5	10
	Δ			10		10	10		10	10	5
	ΔΔ				10		10	10	5	10	10

4.2. Objective evaluation

Each feature was evaluated by objective evaluation score: Mel-Cepstral Distortion (Mel-CD) [1]. The 0th value corresponding to a power coefficient is not included. A small Mel-CD score means the conversion was successfully done.

Figure 3 shows experimental results of audio-only, visual-only and audio-visual VCs in the three environments. In these figures, the vertical axes indicate Mel-CD [dB]. According to the first figure (1), audio-only and audio-visual VCs achieved almost the same performance (around 4.2dB), followed by visual-only VC. However, it is obvious that audio-only VC (**A1**) was drastically degraded in noisy conditions. Compared to the audio-only result, audio-visual VCs (**AV2**, **AV5**) could improve the quality of generated speech. This indicated that visual information can work effectively in noisy conditions. It is also notable that visual-only VC achieved roughly 4.5dB performance (see (4)). This might be caused by small vocabulary, but still there is a possibility to realize a large-vocabulary system converting from lip movies to speech signals.

In the second result (2) for **AV1**, **AV2** and **AV3**, the method using 15 MCEPs and visual static parameters (**AV1**) was strongly affected by noise, while the others were not. And in the third result (3) for **AV4**, **AV5**, and **AV6**, the

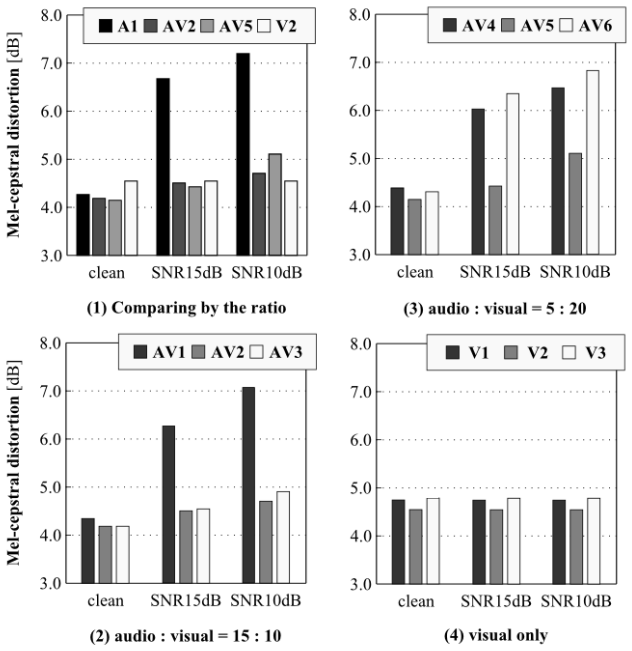

Fig. 3. Results of objective evaluation.

method employing 20-dimensional visual derivatives in addition to 5 MCEPs (**AV5**) obviously achieved the best performance. Since the rest two methods included static features, it turns out that visual static components are not so useful, on the other hand, visual derivative components, particularly Δ parameters, are effective. In other words, if we denote the performances of static, Δ, and ΔΔcomponents by E_s, E_d, and $E_a$respectively, the relationship:

$$E_s < E_a < E_d \quad (4)$$

can be experimentally obtained. According to previous works [14, 15], ASR in noisy environments has the same tendency; dynamic features are more reliable than static one. We further investigate this phenomenon in near future.

4.3. Subjective evaluation

We conducted subject evaluations for speech quality and speaker individuality. In the test of speech quality, samples of converted speech by the conventional method [1] and by the proposed method were presented to listeners as A and B in a random order. Listeners were asked which sample sounded more natural. As the test of speaker individuality, an XAB test was conducted. First, we presented a target speech X, and converted speeches by the conventional method and by the proposed method as A and B in a random order. Note that both A and B are converted from the source speaker into the target speaker X. In each XAB set, we presented speech samples of the same utterance. Listeners were asked to choose which of A or B sounded more similar to X in terms of speaker individuality. In the both experiments, each subject gave a relative assessment by choosing three-scale scores: (1) A, (2) neutral, and (3) B.

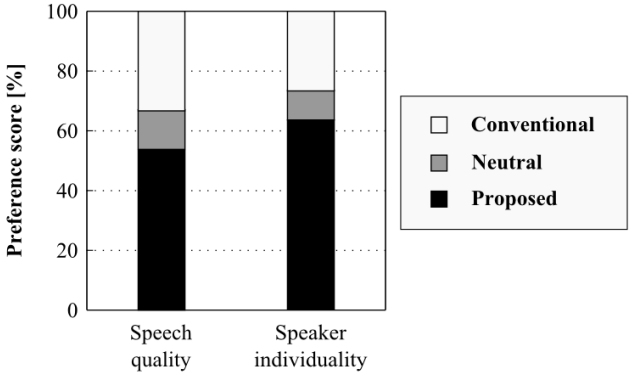

Fig. 4. Results of subjective evaluation.

The number of utterances was two in the evaluation set, and the number of subjects was 15. The converted speeches in the subject evaluations were generated under the noise environment (SNR10dB). And we employed **AV2** as a proposed method because it was the best performance in the noise environment (SNR10dB).

Figure 4 shows the results of subject evaluation. The performances for speech quality and speaker individuality of the proposed method were better than those of the conventional method. When authors listened each converted speech, we felt that the converted speeches by the conventional method were artificial, while the converted speeches by the proposed method included block noise and natural speech.

5. CONCLUSION

This paper proposes audio-visual voice conversion to improve quality of converted speech in noisy environments. Several audio and visual feature combinations were tested. Experimental results indicate our proposed method can successfully improve the performance, and that visual dynamic features are especially effective.

Our future works include: evaluation of our scheme in other noise conditions and visually difficult situations, development of real-time system applicable in real environments, investigation of the relation between the performance and static/dynamic features in several modalities. Global variance and dynamic feature might be explored.

6. ACKNOWLEDGMENT

This work was supported by A-STEP (Adaptable and Seamless Technology transfer Program through target-driven R&D) operated by JST (Japan Science and Technology Agency). And the authors are deeply grateful to Prof. Tomoki Toda of Nara Institute of Science and Technology, Japan, for giving valuable advices and providing techniques of voice conversion.

7. REFERENCES

[1] T.Toda et al., "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language*, vol. 15, no. 8, pp. 2222-2225, 2007.

[2] K. Nakamura et al., "The use of air-pressure sensor in electrolaryngeal speech enhancement based on statistical voice conversion," in *Proc. INTERSPEECH 2010*, pp. 1628-1631.

[3] K.Sawada et al., "Statistical voice conversion using GA-based informative feature," in *Proc. APSIPA ASC 2012*, PS.5-SLA. 18.9.

[4] R. Takashima et al., "Exemplar-based voice conversion in noisy environment," in *Proc. SLT 2012*, pp. 313-317.

[5] G.Potamianos et al., "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. ICASSP '98*, pp. 3733-3736.

[6] K.Iwano et al., "Bimodal speech recognition using lip movement measured by optical-flow analysis," in *Proc. HSC 2001*, pp. 187-190.

[7] C.Miyajima et al., "Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights," in *Proc. ICSLP 2000*, pp. 1023-1026.

[8] A.Barbulescu et al., "Audio-visual speaker conversion using prosody features," in *Proc. AVSP 2013*, pp. 11-16.

[9] C.Ishi et al., "Real-time audio-visual voice activity detection for speech recognition in noisy environments," in *Proc. AVSP 2010*, pp. 81-84.

[10] S.Takeuchi et al., "Voice activity detection based on fusion of audio and visual information," in *Proc. AVSP 2009*, pp.151-154.

[11] H.Kawahara et al., "Restructuring speech representation using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187-207, 1999.

[12] S.Tamura et al., "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," in *Proc. AVSP 2010*, pp. 85-88.

[13] T. Drugman et al., "Relevant feature selection for audio-visual speech recognition," in *Proc. MMSP 2007*, pp. 179-182.

[14] C.Yang et al., "Static and dynamic spectral features: their noise robustness and optimal weights for ASR," in *Proc. ICASSP 2005*, pp. 241-244.

[15] S.Tamura et al., "Multi-stream acoustic model adaptation for noisy speech recognition," in *Proc. APSIPA ASC 2012*, OS.3-SLA. 1.4.