W2N-AVSC: Audiovisual Extension For Whisper-To-Normal Speech Conversion

Shogo Seki[†,1], Kanami, Imamura[‡], Hirokazu Kameoka[†], Takuhiro Kaneko[†], Kou Tanaka[†], Noboru Harada[†]
[†]*NTT Communication Science Laboratories, NTT Corporation*, Japan
[‡]*Graduate School of Information Science and Technology, The University of Tokyo*, Japan
[1]shogo.seki.va@hco.ntt.co.jp

***Abstract*—In this paper, we extend a method of converting speaking styles for whispered speech, i.e., whisper-to-normal speaking style conversion (W2N-SC). W2N-SC problem is similar but different from a regular voice conversion (VC) task and more challenging due to the characteristics of whispered speech and the deal with different speaking styles. In our previous study, we addressed the task-specific difficulties and developed a variational autoencoder (VAE)-based non-parallel approach called W2N-SC. While W2N-SC demonstrated superior performance to other parallel-data-free approaches, there remains room for improvement in conversion quality. To overcome the limitation, we propose W2N-AVSC, an audiovisual extension of W2N-SC. Unlike the conventional W2N-SC, the proposed W2N-AVSC can take visual information, e.g., lip movements, into account in the conversion of whispered speech. Furthermore, to perform W2N-AVSC, we develop a new audiovisual dataset recording the faces of speakers reading texts in various ways, such as in normals and whispers. Through experimental evaluations using clean and noisy whispered inputs, we reveal an effective representation of visual information, demonstrating that W2N-AVSC perceptually performs better than W2N-SC.**

***Index Terms*—audiovisual signal processing, speaking style conversion, whispered speech, variational autoencoder**

I. Introduction

Whispering is generally used to convey private information without being overheard by third persons or to avoid disturbing others in a quiet place. If whispered speech can be automatically converted to sound like normal speech, listeners can still communicate naturally in these situations. This paper aims to achieve such applications and deals with a whisper-to-normal speaking style conversion (W2N-SC) task [1]–[4], a problem of converting a whispered speech into a normal one.

One of the most relevant tasks to W2N-SC problems is voice conversion (VC), which refers to the problem of converting the non-linguistic or para-linguistic information of an input speech while preserving the linguistic information. However, W2N-SC tasks are different from regular VC tasks, such as converting speaker identities, and more challenging because 1) whispered speech has no/less pitch information and extremely low energy, and 2) two speaking styles can be switched continuously.

To elucidate and address the task-specific difficulties in the W2N-SC problem, we previously developed the method called W2N-SC [4]. W2N-SC is a non-parallel approach based on a variational autoencoder (VAE) [5]. More specifically, W2N-SC employs the frame work of auxiliary classifier VAE-VC (ACVAE-VC) [6], where both whispered and normal voices are virtually treated as different speakers' voices in regular VC tasks. Meanwhile, the encoder-decoder is modified to an any-to-many architecture to accept inputs in different speaking styles, and data augmentation using noisy samples is introduced to improve the robustness against noise. W2N-SC demonstrated superior performance to other non-parallel systems, e.g., those using an autoencoder-based, VAE-based, and generative adversarial network (GAN) [7]-based VCs [8]–[10]. However, the conversion performance is still limited, and much room for improvement existed.

One approach to overcome the limitation is to extend W2N-SC to use multimodal information. The use of multimodal information is a promising approach in various applications, which enables us to improve existing tasks on each modality [11]–[15] and achieve unexplored tasks across multiple modalities [16]–[18]. Among these tasks, lip-to-speech synthesis [19], [20], a task of predicting speech from lip movements, is particularly relevant to W2N-SC.

In this paper, motivated by the success in lip-to-speech synthesis tasks, we propose an audiovisual extension of W2N-SC, referred to as W2N-AVSC. The proposed W2N-AVSC differs from the conventional W2N-SC in that visual information is considered in the conversion of whispered speech. More specifically, visual information relating to lip movements is used as a prior to the variational posterior of acoustic information, allowing us to convert whispered speech more accurately. One emerging issue in proposing W2N-AVSC would be that, to the best of our knowledge, there exist few datasets including both speech in different speaking styles and audiovisual information [21]. To this end, we develop a new audiovisual dataset that records the faces of speakers reading texts in various ways. A key difference from the existing audiovisual dataset [21] is that our audiovisual dataset includes whispered speech speaking in both quiet and noisy environments. Furthermore, our dataset includes normal speech speaking disfluently for future research and development of diverse speaking style conversions such as a disfluency detection in audio [22]. In the experimental evaluations using the developed audiovisual dataset, we compare and investigate the conventional W2N-SC and the proposed W2N-AVSCs with several visual information, demonstrating the potential of improving the speaking-style conversion performance in W2N-AVSC.

II. Conventional method: W2N-SC

Let the acoustic feature sequence and the one-hot encoded attribute class label, i.e., speaking style, be $\mathbf{A}$ and y, respectively. The Conventional W2N-SC follows a conditional VAE (CVAE) [23] framework, assuming that the (unconditional) audio encoder distribution $q_{\phi_{\mathrm{A}}}(\mathbf{Z}|\mathbf{A})$ and the (conditional) decoder distribution $p_\theta(\mathbf{A}|\mathbf{Z}, y)$ follow Gaussian distributions:

$$q_{\phi_{\mathrm{A}}}(\mathbf{Z}|\mathbf{A}) = \mathcal{N}(\boldsymbol{\mu}_{\phi_{\mathrm{A}}}(\mathbf{A}), \mathrm{diag}\boldsymbol{\sigma}^2_{\phi_{\mathrm{A}}}(\mathbf{A})), \quad (1)$$

$$p_\theta(\mathbf{A}|\mathbf{Z}, y) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{Z}, y), \mathrm{diag}\boldsymbol{\sigma}^2_\theta(\mathbf{Z}, y)), \quad (2)$$

where $\boldsymbol{\mu}_{\phi_{\mathrm{A}}}(\mathbf{A})$ and $\boldsymbol{\sigma}^2_{\phi_{\mathrm{A}}}(\mathbf{A})$ are the audio encoder distribution parameters, and $\boldsymbol{\mu}_\theta(\mathbf{Z}, y)$ and $\boldsymbol{\sigma}^2_\theta(\mathbf{Z}, y)$ are the decoder distribution parameters. ϕ_{A} and θ represent the network parameters of the audio encoder and decoder, respectively. In the conventional W2N-SC, the following variational lower bound to be maximized is used for the training criterion:

$$\mathcal{I} = \mathbb{E}_{(\mathbf{A},y)\sim p(\mathbf{A},y)}[\mathbb{E}_{\mathbf{Z}\sim q_{\phi_{\mathrm{A}}}(\mathbf{Z}|\mathbf{A})}[\log p_\theta(\mathbf{A}|\mathbf{Z}, y)] - \mathcal{D}_{\mathrm{KL}}[q_{\phi_{\mathrm{A}}}(\mathbf{Z}|\mathbf{A})||p(\mathbf{Z})]], \quad (3)$$

where $\mathbb{E}_{(\mathbf{A},y)\sim p(\mathbf{A},y)}[\cdot]$ denotes the sample mean over all the M training pair examples $\{\mathbf{X}_m, y_m\}_{m=1}^M$, and $\mathcal{D}_{\mathrm{KL}}[\cdot||\cdot]$ is the Kullback-Leibler (KL) divergence. We assume the prior distribution $p(\mathbf{Z})$ as a standard Gaussian distribution: $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Similar to ACVAE-VC [6], W2N-SC incorporates the expectation of the mutual information $I(y; \mathbf{A}|\mathbf{Z})$ into the training criterion. Since it is difficult to use the mutual information directory, the following variational lower bound is used instead:

$$\mathcal{J} = \mathbb{E}_{(\mathbf{A}_{\mathrm{s}},y_{\mathrm{s}})\sim p(\mathbf{A},y),\mathbf{Z}\sim q_{\phi_{\mathrm{A}}}(\mathbf{Z}|\mathbf{A}_{\mathrm{s}})}[\mathbb{E}_{(\mathbf{A}_{\mathrm{t}},y_{\mathrm{t}})\sim p(\mathbf{A},y),\mathbf{A}\sim p_\theta(\mathbf{A}|\mathbf{Z},y_{\mathrm{t}})}[\log r_\psi(y_{\mathrm{t}}|\mathbf{A})]], \quad (4)$$

where $r_\psi(y|\mathbf{A})$ is an auxiliary classifier distribution with the network parameter ψ. Moreover, W2N-SC incorporates the cross-entropy:

$$\mathcal{K} = \mathbb{E}_{(\mathbf{A},y)\sim p(\mathbf{A},y)}[\log r_\psi(y|\mathbf{A})]. \quad (5)$$

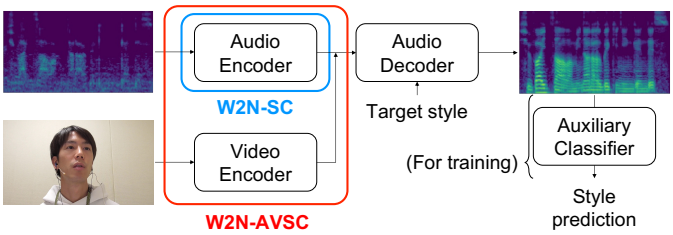
III. Proposed method: W2N-AVSC

Fig. 1 shows an overview and comparison of the conventional W2N-SC and the proposed W2N-AVSC. W2N-AVSC incorporates W2N-SC framework with visual information by employing an lip-to-speech synthesis technique.

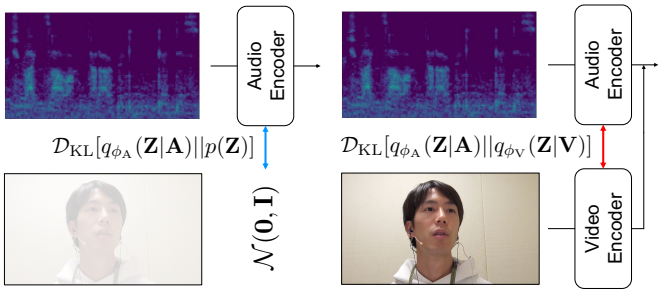
Let the visual feature sequence synchronized with the acoustic feature $\mathbf{A}$ be $\mathbf{V}$. We employ to use an additional encoder network for visual information, i.e., a video encoder, and also assumes that the video encoder distribution $\phi_{\mathrm{V}}(\mathbf{Z}|\mathbf{V})$ follows Gaussian distribution:

$$q_{\phi_{\mathrm{V}}}(\mathbf{Z}|\mathbf{V}) = \mathcal{N}(\boldsymbol{\mu}_{\phi_{\mathrm{V}}}(\mathbf{V}), \mathrm{diag}\boldsymbol{\sigma}^2_{\phi_{\mathrm{V}}}(\mathbf{V})) \quad (6)$$

where $\boldsymbol{\mu}_{\phi_{\mathrm{V}}}(\mathbf{V})$ and $\boldsymbol{\sigma}^2_{\phi_{\mathrm{V}}}(\mathbf{V})$ are the video encoder distribution parameters, and ϕ_{V} represents the network parameters. According to [20], the video encoder distribution is used as the prior instead of a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ at



(a) System overview.



(b) W2N-SC (conventional). (c) W2N-AVSC (proposed).

Fig. 1: (a) System overview of whisper-to-normal speaking style conversion framework and comparison of (b) conventional W2N-SC and (c) proposed W2N-AVSC. A standard Gaussian distribution is used as the prior in W2N-SC, and a video encoder distribution is used instead in W2N-AVSC.

the KL divergence term in the variational lower bound. Eq. (3) can be rewritten as:

$$\mathcal{I} = \mathbb{E}_{(\mathbf{A},\mathbf{V},y)\sim p(\mathbf{A},\mathbf{V},y)}[\mathbb{E}_{\mathbf{Z}\sim q_{\phi_{\mathrm{A}}}(\mathbf{Z}|\mathbf{A})}[\log p_\theta(\mathbf{A}|\mathbf{Z}, y)] - \mathcal{D}_{\mathrm{KL}}[q_{\phi_{\mathrm{A}}}(\mathbf{Z}|\mathbf{A})||q_{\phi_{\mathrm{V}}}(\mathbf{Z}|\mathbf{V})]], \quad (7)$$

where $\mathbb{E}_{(\mathbf{A},\mathbf{V},y)\sim p(\mathbf{A},\mathbf{V},y)}[\cdot]$ denotes the sample mean over all the M training triplet examples $\{\mathbf{A}_m, \mathbf{V}_m, y_m\}_{m=1}^M$. This encourage the audio encoder to encode speech contents more accurately by taking the visual information into account. The whole training criterion is then given by replacing eq. (3) with eq. (7):

$$\mathcal{I} + \lambda_{\mathcal{J}}\mathcal{J} + \lambda_{\mathcal{K}}\mathcal{K}, \quad (8)$$

where $\lambda_{\mathcal{J}} \geq 0$ and $\lambda_{\mathcal{K}} \geq 0$ are regularization parameters, which weigh the importances of the regularization terms.

Note that, through our preliminary experiments, we found that using not only the audio encoder output $\mathbf{Z} \sim q_{\phi_{\mathrm{A}}}(\mathbf{Z}|\mathbf{A})$ but also the video encoder output $\mathbf{Z} \sim q_{\phi_{\mathrm{V}}}(\mathbf{Z}|\mathbf{V})$ as input for the decoder provides us better performance. Thus, we simply fuse two encoder outputs by manipulating element-wise summation, and use the accumulated output as the decoder input $\mathbf{Z}$.

Once all the network parameters are trained, a source audio and video features $\mathbf{A}_{\mathrm{s}}$ and $\mathbf{V}_{\mathrm{s}}$ can be converted by using a target attribute class labels y_{t}:

$$\hat{\mathbf{A}}_{\mathrm{t}} = \boldsymbol{\mu}_\theta(\boldsymbol{\mu}_{\phi_{\mathrm{A}}}(\mathbf{A}_{\mathrm{s}}) + \boldsymbol{\mu}_{\phi_{\mathrm{V}}}(\mathbf{V}_{\mathrm{s}}), y_{\mathrm{t}}). \quad (9)$$

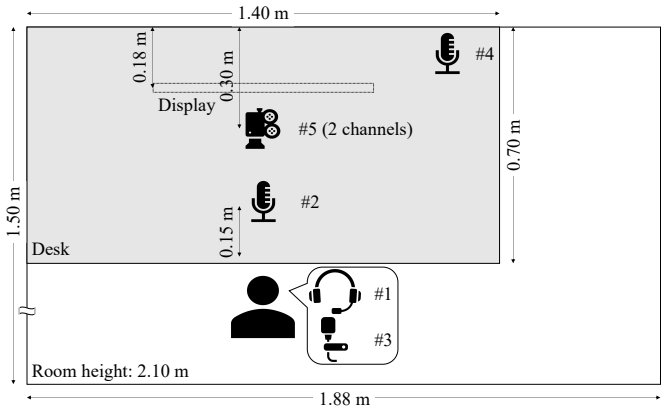


Fig. 2: Recording room layouts.

IV. Audiovisual dataset for W2N-AVSC

Although there exist several datasets that contain whispered speech [24], [25] and movies recording speakers speaking [11]–[13], [26], to the best of our knowledge, there are few datasets that satisfy both conditions [21]. To this end, we collected multimodal recordings of multiple speakers reading the same sentences in multiple speaking styles and built a dataset.

Fig. 2 shows the recording room and recording device layouts. We used four different (#1: headset, #2: desktop, #3: pin, and # 4: stand) microphones and a two-channel microphone in a camcorder (#5), where the sampling rate of each microphone is set to 48 kHz. For video recording, we used one camcorder recording speakers' faces, and the image size and the frame per seconds is set to 1920×1080 and 59.94, respectively. Note that audio signals were synchronously recorded except for microphone #5 and the audio and video recordings were manually synchronized in a post-processing.

We collected 25 Japanese speakers, including 12 females and 13 males, and the speakers are distributed from their twenties to sixties. Each speaker is asked to read scripts on a display. For reading scripts, we selected 300 sentences in a public domain Japanese text corpus called inter-field task accelarating (ITA) corpus[1].

Different from [21], we collected two types of whispering: (standard) whispering in a quiet environment and whispering in a noisy environment, i.e., whispering with the Lombard effect. We recorded the whispered speech with the Lombard effect by playing loud noise on a headset and presenting it to speakers. Furthermore, we collected another speaking-style speech: speech with intentional fillers, i.e., disfluent speech, for future research and development. We recorded five types of disfluent speech by allocating several kinds of fillers in different positions[2].

[1]https://github.com/mmorise/ita-corpus

[2]We plan to release the dataset in the future.

V. Experimental Evaluation

We evaluated the proposed W2N-AVSC under clean and noisy conditions with a comparison of visual information representations and conducted objective and subjective evaluations for each input condition. In the evaluation, we dealt with a single-speaker case and left further developments, such as multi-speaker and multi-style extensions, as our future work.

A. Experimental settings

We used a subset of the collected data, which consists of movies of one Japanese male speaker reading the same utterances in normal and whispered styles. We used the first 16 and the last 256 sentences from these sentences for evaluation and training, respectively. For audio information, all the speech signals were resampled at 16 kHz, and 80-dimensional log mel-spectrograms were extracted with a 64 ms frame length and a 16 ms frameshift. For video information, we evaluated three types of visual features related to lip regions: facial action units (FAU), 2D landmark positions (LMK), and gray-scale images (IMG). Feature extraction was conducted through by OpenFace [27]–[29], an open source facial feature extraction toolkit[3], then each feature was vectorized. The number of feature dimensions of AU, landmark positions, and images was 9, 40 (20×2), and 4096 (64×64×1), respectively. Note that the visual features were aligned with acoustic features by selecting the nearest frame at each frameshift.

We used conventional W2N-SC as a baseline and compared it with the proposed W2N-AVSC with three different visual features. The audio encoder, decoder, and auxiliary classifier networks consisted of four-layer convolutional, four-decovolutional, and six-convolutional architectures with gated linear units (GLUs) [30]. All the dimensions of hidden and latent features were set to 64 and 16, respectively. We used the same architecture for the video encoder network, where the number of input features differs depending on visual features.

For training, all the regularization parameters were set at $\lambda_{\mathcal{J}} = \lambda_{\mathcal{K}} = 1$. We used the Adam optimizers [31], where the learning rates were set at 1.0×10^{-3}. All the models were trained for around 30k iterations. For the generation of time-domain signals, the HiFi-GAN vocoder was used [32]. HiFi-GAN vocoder was prepared from a publicly available implementation[4], where "V2" network architecture was used.

We employed a data augmentation approach for noisy data inputs and used the DEMAND dataset [33][5], where the noises categorized as "public" were used for evaluation, and the rest were used for training. The SNR range for the data augmentation was set to 0-10 dB, and the SNR of noisy inputs was set to 5 dB.

B. Objective evaluation

As the evaluation metrics, the average of the Mel-cepstral distortions (MCDs) between the converted and target signals

[3]https://github.com/TadasBaltrusaitis/OpenFace

[4]https://github.com/kan-bayashi/ParallelWaveGAN

[5]https://zenodo.org/record/1227121

TABLE I: Objective evaluation results on clean/noisy inputs.

Method	MCD [dB]	LFC [-]	MOSnet [-]
Source feature	9.78 / 8.05	-0.03 / -0.25	2.84 / 2.92
W2N-SC [4]	8.57 / 8.59	0.10 / -0.10	2.88 / 2.85
W2N-AVSC (FAU)	9.02 / 8.79	0.08 / -0.12	2.93 / 2.91
W2N-AVSC (LMK)	8.56 / 8.61	0.21 / 0.14	2.80 / 2.81
W2N-AVSC (IMG)	8.40 / 8.68	0.14 / 0.18	3.09 / 3.02
Target feature	7.09	0.85	2.93

was used in the objective evaluation, where the dynamic time warping (DTW) was applied to align Mel-cepstral sequence pairs in advance. The frame-level MCDs were averaged to obtain the utterance-level MCDs for each converted signal. We also used log-scaled fundamental frequency correlations (LFCs) and scores obtained from pre-trained MOSnet [34][6]. In addition to conventional W2N-SC and proposed W2N-AVSCs, synthesized signals obtained from source features and target features were also evaluated.

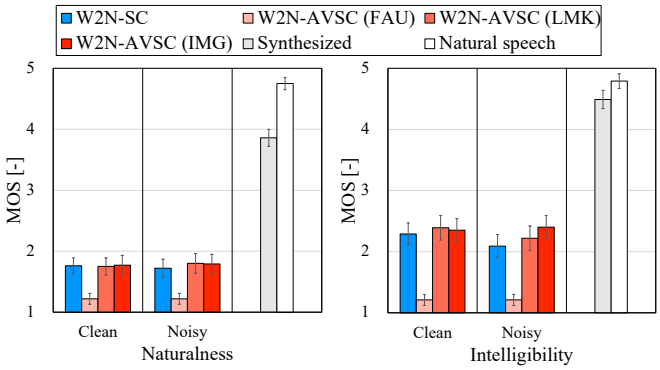
Table I shows a comparison of the conversion performance of each method using clean and noisy whispered speech. From a comparison of source and target features, we can see that 1) the MCD is oddly improved when using noisy source features[7], 2) there is no significant difference in the MOSnet scores of source and target features, and 3) the LFCs work reasonably. When then comparing the LFCs of W2N-SC and W2N-AVSCs, it can be seen that 1) W2N-AVSC (FAU) consistently underperform W2N-SC, and 2) W2N-AVSC (LMK) and W2N-AVSC (IMG) provides performance improvements, performing the best in W2N-AVSC (LMK). This demonstrates the potential of W2N-AVSC when using appropriate feature representations.

C. Subjective evaluation

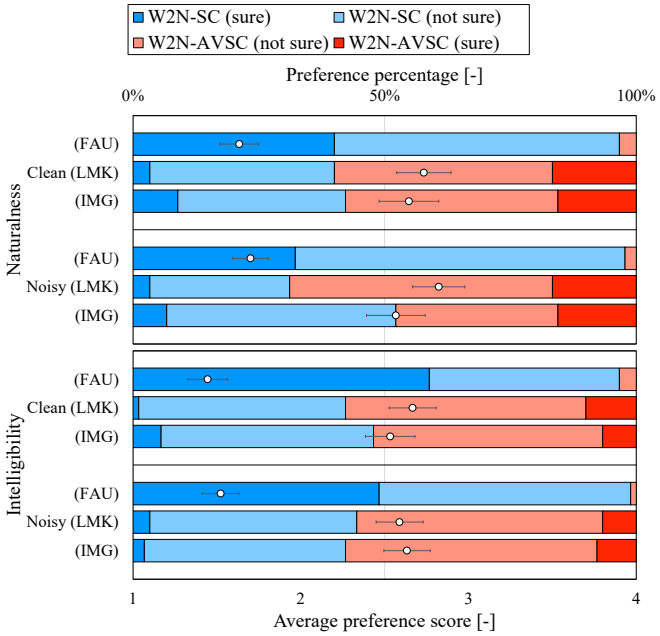
Subjective evaluation tests on naturalness and intelligibility were conducted to investigate perceptual quality. In the subjective evaluation, ten converted samples for each were used to reduce the evaluation cost. Mean opinion score (MOS) tests were conducted for both naturalness and intelligibility, where four different conversion methods for clean and noisy inputs, natural speech, and synthesized signals from natural features were evaluated. This amounted to 10 systems. Ten subjects participated in the test, and the subjects were asked to assign a score by selecting "1: Bad", "2: Poor", "3: Fair", "Good", or "5: Excellent". We also conducted preference (ABX) tests for both naturalness and intelligibility. Each proposed W2N-AVSC was compared with conventional W2N-SC using clean and noisy inputs, resulting in six pairs of conversion methods. Ten subjects were joined; each was presented with a target utterance at first and two converted utterances, A and B. Then, each subject was asked to assign a score by selecting "1: A (sure)", "2: A (not sure)", "3: B (not sure)", or "4: B (sure)" in terms of better perceptual quality. Note that the order of two converted samples, A and B, was randomly shuffled.

[6]https://github.com/lochenchou/MOSNet

[7]One possible reason might be because that the energy of background noise positively affected the source features to resemble with target features.



(a) MOS test result.



(b) ABX test result.

Fig. 3: Subjective evaluation results on naturalness and intelligibility, where the average preference scores in the ABX test are denoted as white dots. The error bars indicate the 95% confidence intervals.

Fig. 3 shows a comparison of the conversion performance of each method on naturalness and intelligibility. First, W2N-AVSC (FAU) consistently failed to perform well. We found that the converted speech by W2N-AVSC (FAU) did not preserve speech context correctly. This might be because the FAU feature is too compressed to represent lip movement and can be treated as noise. Hence, we hereafter focus on two W2N-AVSCs, W2N-AVSC (LMK) and W2N-AVSC (IMG). From the MOS test results, it can be seen that these W2N-AVSCs perform comparably with W2N-SC on naturalness and slightly better on intelligibility. We also found that, from the ABX test, these two W2N-AVSCs are preferable to W2N-SC. We can conclude from these results that these W2N-AVSCs perform perceptually better than W2N-SC. Moreover, in a comparison of the two W2N-AVSCs, we found that W2N-AVSC (LMK) tends to be better than W2N-AVSC (IMG).

VI. Conclusion

This paper proposed an audiovisual extension of W2N-SC called W2N-AVSC. The proposed W2N-AVSC differs from the conventional W2N-SC in that visual information about lip movements is considered in the conversion of whispered speech. Furthermore, for the purpose of performing a wide variety of tasks, including W2N-AVSC tasks, we also developed a new audiovisual dataset. From the experimental evaluations, we revealed that W2N-AVSC performs perceptually better than W2N-SC when using an appropriate visual representation.

ACKNOWLEDGEMENT

This work was partly supported by JST, CREST Grant Number JPMJCR19A3, Japan.

References

[1] Hailun Lian, Yuting Hu, Jian Zhou, Huabin Wang, and Liang Tao, "Whisper to Normal Speech Based on Deep Neural Networks with MCC and F0 Features," in *International Conference on Digital Signal Processing*, 2018, pp. 1–5.

[2] Hailun Lian, Yuting Hu, Weiwei Yu, Jian Zhou, and Wenming Zheng, "Whisper to Normal Speech Conversion Using Sequence-to-Sequence Mapping Model With Auditory Attention," *IEEE Access*, vol. 7, pp. 130495–130504, 2019.

[3] Mihir Parmar, Savan Doshi, Nirmesh J. Shah, Maitreya Patel, and Hemant A. Patil, "Effectiveness of Cross-Domain Architectures for Whisper-to-Normal Speech Conversion," in *European Signal Processing Conference*, 2019, pp. 1–5.

[4] Shogo Seki, Hirokazu Kameoka, Takuhiro Kaneko, and Kou Tanaka, "Non-parallel whisper-to-normal speaking style conversion using auxiliary classifier variational autoencoder," *IEEE Access*, vol. 11, pp. 44590–44599, 2023.

[5] Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations*, 2014, pp. 1–14.

[6] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "ACVAE-VC: Non-Parallel Voice Conversion With Auxiliary Classifier Variational Autoencoder," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1432–1443, 2019.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1–9, 2014.

[8] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss," in *International Conference on Machine Learning*, 2019, pp. 5210–5219.

[9] Wen-Chin Huang, Hsin-Te Hwang, Yu-Huai Peng, Yu Tsao, and Hsin-Min Wang, "Voice Conversion Based on Cross-Domain Features Using Variational Auto Encoders," in *International Symposium on Chinese Spoken Language Processing*, 2018, pp. 51–55.

[10] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks," in *IEEE Spoken Language Technology Workshop*, 2018, pp. 266–273.

[11] Arsha Nagrani, Joon-Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech*, 2017, pp. 2616–2620.

[12] Joon-Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018, pp. 1086–1090.

[13] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "VoxCeleb: Large-Scale Speaker Verification in the Wild," *Computer Science and Language*, p. 101027, 2019.

[14] Triantafyllos Afouras, Joon-Son Chung, and Andrew Zisserman, "The Conversation: Deep Audio-Visual Speech Enhancement," in *Interspeech*, 2018, pp. 3244–3248.

[15] Sadeghi Mostafa and Alameda-Pineda Xavier, "Mixture of Inference Networks for VAE-Based Audio-Visual Speech Enhancement," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1899–1909, 2021.

[16] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, "Seeing Voices and Hearing Faces: Cross-modal biometric matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8427–8436.

[17] Hirokazu Kameoka, Takuhiro Kaneko, Shogo Seki, and Kou Tanaka, "CAUSE: Crossmodal Action Unit Sequence Estimation from Speech," in *Interspeech*, 2022, pp. 506–510.

[18] Yasunori Ohishi, Marc Delcroix, Tsubasa Ochiai, Shoko Araki, Daiki Takeuchi, Daisuke Niizumi, Akisato Kimura, Noboru Harada, and Kunio Kashino, "ConceptBeam: Concept Driven Target Speech Extraction," in *ACM International Conference on Multimedia*, 2022, pp. 4252–4260.

[19] Renukanand Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and C.V. Jawahar, "Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13793–13802.

[20] Ravindra Yadav, Ashish Sardana, Vinay P Namboodiri, and Rajesh M Hegde, "Speech Prediction in Silent Videos Using Variational Autoencoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7048–7052.

[21] T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 8101–8105.

[22] Shaolei Wang, Wanxiang Che, and Ting Liu, "A neural attention model for disfluency detection," in *International Conference on Computational Linguistics (COLING)*, 2016, pp. 278–287.

[23] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, "Semi-Supervised Learning with Deep Generative Models," in *Advances in Neural Information Processing Systems*, 2014, pp. 3581–3589.

[24] Boon Pang Lim, *Computational Differences between Whispered and Non-Whispered Speech*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 2011.

[25] Fred Cummins, Marco Grimaldi, Thomas Leonard, and Juraj Simko, "The CHAINS Corpus: Characterizing individual speakers," in *International Conference on Speech and Computer*, 2006, pp. 431–435.

[26] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[27] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018, pp. 59–66.

[28] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson, "Cross-Dataset Learning and Person-Specific Normalisation for Automatic Action Unit Detection," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015, pp. 1–6.

[29] Amir Zadeh, Yao Chong Lim, Tadas Baltrušaitis, and Louis-Philippe Morency, "Convolutional Experts Constrained Local Model for 3D Facial Landmark Detection," in *IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2519–2528.

[30] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, "Convolutional Sequence to Sequence Learning," in *International Conference on Machine Learning*, 2017, pp. 1243–1252.

[31] Diederik P Kingma and Jimmy Lei Ba, "Adam: A Method for Stochastic Gradient Optimization," in *International Conference on Leanrning Representations*, 2015, pp. 1–15.

[32] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems*, 2020, pp. 17022–17033.

[33] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *International Congress on Acoustics*, 2013, pp. 1–5.

[34] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Interspeech*, 2019, pp. 1541–1545.