# VCVTS: MULTI-SPEAKER VIDEO-TO-SPEECH SYNTHESIS VIA CROSS-MODAL KNOWLEDGE TRANSFER FROM VOICE CONVERSION

*Disong Wang[1*], Shan Yang[2], Dan Su[2], Xunying Liu[1], Dong Yu[2], Helen Meng[1,3]*

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Tencent AI lab, China
[3]Centre for Perceptual and Interactive Intelligence, Hong Kong SAR, China
{dswang, xyliu, hmmeng}@se.cuhk.edu.hk, {shaanyang, dansu, dyu}@tencent.com

## ABSTRACT

Though significant progress has been made for speaker-dependent Video-to-Speech (VTS) synthesis, little attention is devoted to multi-speaker VTS that can map silent video to speech, while allowing flexible control of speaker identity, all in a single system. This paper proposes a novel multi-speaker VTS system based on cross-modal knowledge transfer from voice conversion (VC), where vector quantization with contrastive predictive coding (VQCPC) is used for the content encoder of VC to derive discrete phoneme-like acoustic units, which are transferred to a Lip-to-Index (Lip2Ind) network to infer the index sequence of acoustic units. The Lip2Ind network can then substitute the content encoder of VC to form a multi-speaker VTS system to convert silent video to acoustic units for reconstructing accurate spoken content. The VTS system also inherits the advantages of VC by using a speaker encoder to produce speaker representations to effectively control the speaker identity of generated speech. Extensive evaluations verify the effectiveness of proposed approach, which can be applied in both constrained vocabulary and open vocabulary conditions, achieving state-of-the-art performance in generating high-quality speech with high naturalness, intelligibility and speaker similarity. Our demo page is released here[1].

***Index Terms***— Multi-speaker, Video-to-Speech synthesis, voice conversion, knowledge transfer, vector quantization

## 1. INTRODUCTION

Video-to-Speech (VTS) synthesis aims to reconstruct speech signals from silent video by exploiting their bi-modal correspondences. VTS has various compelling applications such as assistive communication for patients with inability to produce voiced sounds (i.e., aphonia), and voice restoration for videoconferencing when speech signals are corrupted by noise or lost. Previous efforts generally build statistical models to map visual features to acoustic features, then use a vocoder to synthesize the waveform. Typical modelling approaches include Hidden Markov Models [1,2], non-negative matrix factorization [3], maximum likelihood estimation [4] and deep learning methods [1, 5–17]. Most works [1–15] are restricted to small datasets (e.g., GRID [18]) to create single-speaker systems under constrained conditions with limited vocabulary, which hinders their practical deployment. A few studies [16, 17] propose to use speaker representations to capture speaker characteristics and control the speaker identity of generated speech, such that multi-speaker VTS can be achieved in a single system. [16] uses a pre-trained

speaker encoder optimized for speaker recognition task to extract speaker representations, [17] applies adversarial training to achieve speaker disentanglement. Both approaches directly map cropped lips to speech, which treats the deep neural model as a black box, leading to insufficient interpretability of intermediate representations learned by the model. Besides, compared with other speech generation tasks, e.g., text-to-speech (TTS) synthesis or voice conversion (VC), current VTS systems tend to generate less natural-sounding speech.

This paper proposes a novel multi-speaker VTS system referred to as VCVTS (Voice Conversion-based Video-To-Speech), which provides a more legible mapping from lips to speech. This is accomplished by first converting lips to intermediate phoneme-like acoustic units, which are used to accurately restore the spoken content. Besides, VCVTS can generate high-quality speech with flexible control of the speaker identity, leveraging the capabilities of a VC system. Specifically, the proposed approach contains three components: (1) Training a VC system which contains four modules - a *content encoder* using vector quantization with contrastive predictive coding (VQCPC) [19] to derive discrete phoneme-like acoustic units, a *speaker encoder* extracting effective speaker representations to control the speaker identity, a *pitch predictor* inferring the fundamental frequency ($F_0$) to control the pitch contour, and a *decoder* mapping acoustic units, speaker representation and $F_0$ to mel-spectrograms; (2) The discrete phoneme-like acoustic units are treated as knowledge, which is transferred across modalities from speech to image, by training a Lip-to-Index (Lip2Ind) network to predict the index sequence of acoustic units, where the index corresponds to the acoustic unit obtained from the codebook of vector quantization (VQ); and (3) A multi-speaker VTS system, i.e., VCVTS, is formed by concatenating the Lip2Ind network with the VQ-codebook, speaker encoder, pitch predictor and decoder of VC. VCVTS provides a legible mapping from lips to speech and inherits the advantages of VC to effectively control speaker identity and pitch contour, which leads to the generation of high-quality speech outputs.

The main contributions of this work are: (1) Derivation of discrete phoneme-like acoustic units by using VQCPC under the VC framework; (2) Development of a Lip2Ind network via cross-modal knowledge transfer to map lips to acoustic units for reconstructing spoken content; and (3) Development of a novel multi-speaker VTS system that can process diverse, unconstrained vocabularies and complicated image scenarios, e.g., on LRW dataset [20].

## 2. PROPOSED APPROACH

Fig. 1 shows the diagram of proposed approach, which contains three components that are elaborated in the following subsections. Assuming that there are $K$ videos with 25 frames per second (FPS),
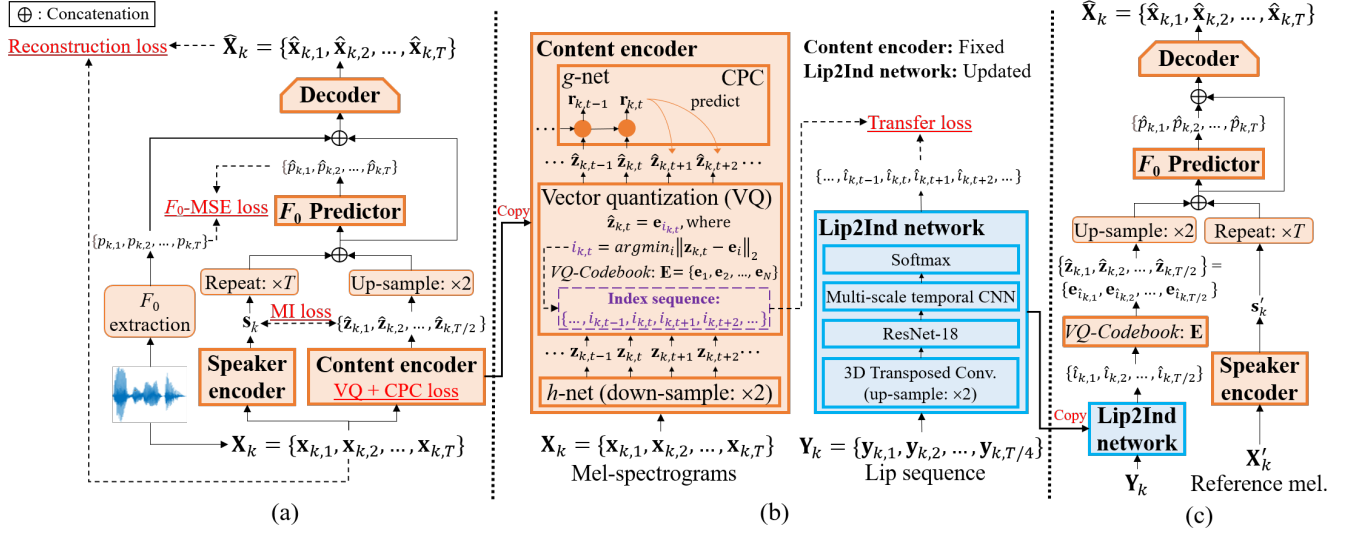
---

[1]Demo: https://wendison.github.io/VCVTS-demo/

**Fig. 1**. Diagram of the proposed approach: (a) VC system training by using VQCPC for the content encoder to derive discrete acoustic units; (b) Cross-modal knowledge transfer from the content encoder of VC to a Lip2Ind network; (c) A multi-speaker VTS system, i.e., VCVTS, which is formed by concatenating the Lip2Ind network with the VQ-codebook, speaker encoder, $F_0$ predictor and decoder of VC system.

each video contains two streams: speech and image. For the $k^{th}$ video, the speech stream is processed to mel-spectrograms $\mathbf{X}_k = \{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, ..., \mathbf{x}_{k,T}\}$ computed with a shift size of 10ms (i.e., 100 FPS) to have a length of $T$, the image stream is processed to obtain the lip sequence $\mathbf{Y}_k = \{\mathbf{y}_{k,1}, \mathbf{y}_{k,2}, ..., \mathbf{y}_{k,T/4}\}$ with a length of $T/4$.

## 2.1. VC system training

As shown in Fig. 1 (a), we propose a new VC system that is modified from VQMIVC [21]. Four modules are contained: speaker encoder, content encoder, pitch predictor and decoder, where the first three modules produce the speaker representation, discrete acoustic units and $F_0$ respectively, which are fed into the decoder to reconstruct mel-spectrograms. The pitch predictor is newly added to infer $F_0$ values that are difficult to be predicted from visual features [1]. VQCPC is used for content encoding as shown in Fig. 1 (b). The $h$-net of content encoder first maps $\mathbf{X}_k$ to $\mathbf{Z}_k = \{\mathbf{z}_{k,1}, \mathbf{z}_{k,2}, ..., \mathbf{z}_{k,T/2}\}$ with a down-sampling factor of 2, then applies VQ on $\mathbf{Z}_k$ to obtain the discrete acoustic units $\hat{\mathbf{Z}}_k = \{\hat{\mathbf{z}}_{k,1}, \hat{\mathbf{z}}_{k,2}, ..., \hat{\mathbf{z}}_{k,T/2}\}$:

$$\hat{\mathbf{z}}_{k,t} = \mathbf{e}_{i_{k,t}}, i_{k,t} = \arg\min_i \|\mathbf{z}_{k,t} - \mathbf{e}_i\|_2 \quad (1)$$

where $\mathbf{e}_i$ is $i^{th}$ unit of the VQ-codebook $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N\}$. To learn a content-related codebook $\mathbf{E}$, we minimize the VQ loss [22] to impose an information bottleneck:

$$L_{VQ} = \frac{2}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T/2} \|\mathbf{z}_{k,t} - sg(\hat{\mathbf{z}}_{k,t})\|_2^2 \quad (2)$$

where $sg(\cdot)$ denotes the stop-gradient operator. Besides, a $g$-net is added after $\hat{\mathbf{Z}}_k$ to obtain $\mathbf{R}_k = \{\mathbf{r}_{k,1}, \mathbf{r}_{k,2}, ..., \mathbf{r}_{k,T/2}\}$ to distinguish a positive sample $\hat{\mathbf{z}}_{k,t+m}$ that is $m$ steps in the future from negative samples drawn from the set $\Omega_{k,t,m}$ by minimizing the contrastive predictive coding (CPC) loss [21, 23]:

$$L_{CPC} = -\frac{1}{KT'M} \sum_{k=1}^{K} \sum_{t=1}^{T'} \sum_{m=1}^{M} \log \left[ \frac{\exp(\hat{\mathbf{z}}_{k,t+m}^T \mathbf{W}_m \mathbf{r}_{k,t})}{\sum_{\tilde{\mathbf{z}} \in \Omega_{k,t,m}} \exp(\tilde{\mathbf{z}}^T \mathbf{W}_m \mathbf{r}_{k,t})} \right] \quad (3)$$

where $T' = T/2 - M$, $\mathbf{W}_m$ ($m=1,2,...,M$) is a trainable projection matrix. CPC forces the discrete acoustic units $\hat{\mathbf{Z}}_k$ to capture 'slow
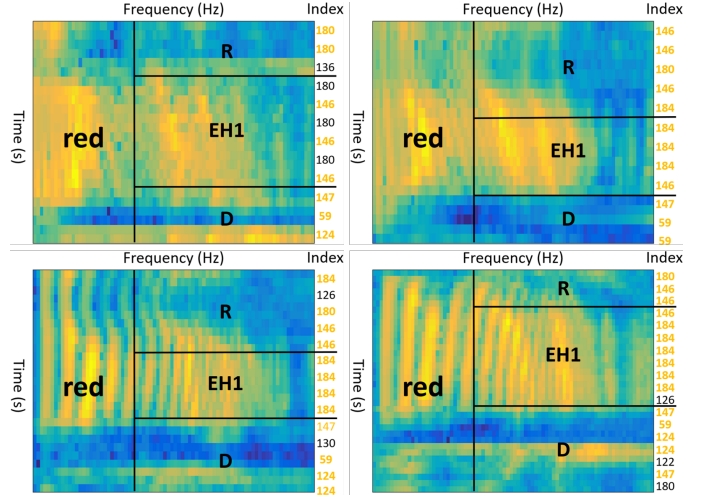


**Fig. 2**. Visualization of mel-spectrograms and acoustic units (denoted by the index sequence) derived by the content encoder of VC from four speakers (selected from the GRID dataset) uttering the same word 'red' with phoneme transcription 'R EH1 D', the common index that occurs at least twice in the four index sequences within the same phoneme is marked as orange.

features' [24] that span many time steps, e.g., phonemes. To separate $\hat{\mathbf{Z}}_k$ from speaker representation $\mathbf{s}_k$, their mutual information (MI) $L_{MI}$ [25, 26] is estimated and minimized during the training. Together with the $F_0$ prediction loss $L_{F_0-MSE}$ and mel-spectrograms reconstruction loss $L_{REC}$, the total loss used for training the VC system is given in Eq. (4). More training details can be found in [21].

$$L_{VC} = L_{VQ} + L_{CPC} + L_{MI} + L_{F_0-MSE} + L_{REC} \quad (4)$$

## 2.2. Cross-modal knowledge transfer

We illustrate some acoustic units derived by a well-trained content encoder of VC as shown in Fig. 2, the acoustic units are denoted by the index sequence $\{i_{k,t}\}$ that is obtained by Eq. (1). We can observe that same indices tend to appear within the same phoneme, and

**Table 1**. Objective and subjective evaluation results of different VTS systems on testing speakers, where 'Seen' and 'Unseen' denote that testing speakers are respectively seen and unseen during training, and subjective results are MOS with 95% confidence intervals for Speech Naturalness (MOS-SN) and Speaker Similarity (MOS-SS).

| Dataset | System | Vocoder | Speakers | Objective | | | | | Subjective | |
|---------|--------|---------|----------|-----------|--|--|--|--|------------|--|
| | | | | PESQ ↑ | STOI ↑ | ESTOI ↑ | MCD ↓ | $F_0$-RMSE ↓ | MOS-SN ↑ | MOS-SS ↑ |
| GRID | XTS [17] | GL | Seen | 1.471 | 0.494 | 0.268 | 8.52 | 55.24 | 2.59 ± 0.09 | 3.56 ± 0.11 |
| | Lip2Wav [16] | GL | Seen | 1.690 | 0.610 | 0.434 | 7.59 | 49.33 | 3.41 ± 0.11 | 4.30 ± 0.11 |
| | VCVTS (ours) | GL | Seen | **1.816** | **0.691** | **0.512** | 6.15 | **43.16** | 3.46 ± 0.12 | 4.38 ± 0.11 |
| | VCVTS (ours) | PWG | Seen | 1.670 | 0.673 | 0.481 | **5.23** | 52.14 | **3.78 ± 0.10** | **4.47 ± 0.12** |
| | XTS [17] | GL | Unseen | 1.295 | 0.439 | 0.160 | 12.04 | 88.97 | 2.33 ± 0.11 | 1.99 ± 0.18 |
| | Lip2Wav [16] | GL | Unseen | 1.293 | 0.419 | 0.202 | 10.11 | 55.35 | 3.13 ± 0.12 | 2.44 ± 0.16 |
| | VCVTS (ours) | GL | Unseen | **1.417** | **0.582** | **0.330** | 8.36 | **50.80** | 3.25 ± 0.11 | 2.66 ± 0.17 |
| | VCVTS (ours) | PWG | Unseen | 1.386 | 0.581 | 0.326 | **7.30** | 54.60 | **3.60 ± 0.10** | **2.75 ± 0.16** |
| LRW | Lip2Wav [16] | GL | Unseen | 1.197 | 0.543 | 0.344 | 9.18 | 71.29 | 2.65 ± 0.10 | 3.05 ± 0.18 |
| | VCVTS (ours) | GL | Unseen | **1.352** | **0.628** | **0.458** | 7.34 | **47.76** | 3.22 ± 0.11 | 3.68 ± 0.13 |
| | VCVTS (ours) | PWG | Unseen | 1.281 | 0.608 | 0.429 | **6.81** | 55.72 | **3.38 ± 0.12** | **3.72 ± 0.14** |

different indices tend to occur in different phonemes, this indicates that the index sequence is strongly associated with the underlying linguistic content, e.g., phonemes. Therefore, the knowledge, i.e., index, can be transferred to the VTS task, such that the accurate spoken content can be appropriately inferred from lip features.

Therefore, for the VQ-codebook $\mathbf{E}$ containing $N$ units, a Lip2Ind network is designed to infer indices from lip sequence as shown in Fig. 1 (b), which is a $N$-way classification problem. Given the video with mel-spectrograms $\mathbf{X}_k$ and lip sequence $\mathbf{Y}_k$, the content encoder takes in $\mathbf{X}_k$ to obtain the index sequence $\{i_{k,1}, i_{k,2}, ..., i_{k,T/2}\}$, which is treated as the learning target of the Lip2Ind network that takes $\mathbf{Y}_k$ as the input. So the Lip2Ind network is trained to minimize the following knowledge transfer loss (i.e., cross-entropy):

$$L_{Trans} = -\frac{2}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T/2} \log q_{i_{k,t}} \qquad (5)$$

where $q_{i_{k,t}}$ is the output of the Lip2Ind network at time $t$ and denotes the posterior probability of index $i_{k,t}$.

### 2.3. Multi-speaker VTS system

The proposed VCVTS is shown in Fig. 1 (c), where the well-trained Lip2Ind network with VQ-codebook can produce accurate acoustic units for spoken content reconstruction, so it can be concatenated with the speaker encoder, pitch predictor and decoder of VC to form a multi-speaker VTS system. This configuration leverages the power of VC to generate the speech with the speaker identity controlled by effective speaker representation $\mathbf{s}'_k$, which is produced by the speaker encoder from a reference utterance, and pitch contour that is controlled by realistic $F_0$ values inferred by the pitch predictor.

### 3. EXPERIMENTS

Experiments are separately conducted on GRID [18] and LRW [20] datasets. GRID contains 33 speakers with 1K videos per speaker, the vocabulary has 52 words which are recorded in a constrained indoor condition. We use 4 speakers (s1, s2, s4 and s29) as testing speakers, and consider two partitions of speakers used for training: all 33 speakers and 29 speakers excluding 4 testing speakers, which measures the performance of VTS for in-domain and out-of-domain speakers, respectively. Videos of each speaker are randomly split into train/dev/test with ratios of 80%/10%/10%. LRW is obtained in unconstrained outdoor conditions, with the vocabulary more than 500 words spoken by hundreds of speakers. We use the default splits

of train/dev/test of LRW for experiments, and assume that speakers in different splits are exclusive considering that the speaker label of each video is unavailable. For speech features, 80-dim mel-spectrograms are calculated with 400-point fast Fourier transform and 10ms shift size that is also used for extracting $F_0$. For image features, Dlib toolkit [27] is employed to obtain facial landmarks that are used to crop and resize the lip region to have a size of 96×96.

The VC system follows most settings of VQMIVC [21], except that the speaker encoder is taken from Resemblyzer[2] and outputs a 256-dim vector, the VQ-codebook of content encoder has 200 160-dim units, $F_0$ predictor contains 2-layer 1D-convolutional network with channel of 384, kernel size of 3, ReLU and layer normalization, which is followed by a linear layer to predict 1-dim $F_0$ values. Besides, the decoder of VC is based on Conformer [28], which consists of 4 blocks, each block has 384-dim attention, 2 attention heads, 1536-dim feed forward (FFN) layers, and 31x1 convolutional kernel size. The speech streams of GRID and LRW are separately used to train different VC systems by Adam [29] for 150 epochs with the first 15-epoch warmup increasing the learning rate from 1e-6 to 1e-3, and batch size of 256. The Lip2Ind network is modified from [30], and consists of a 3D-transposed convolutional network to up-sample lip sequence with a factor of 2, a ResNet-18 [31], a 4-layer multi-scale temporal CNN with each temporal convolution composed of 3 branches that have the kernel size of 3, 5 and 7 respectively, and a softmax layer to predict the probability distribution of index. The Lip2Ind network is trained by Adam for 80 epochs using a cosine scheduler with the initial learning rate of 3e-4, a weight decay of 1e-4 and batch size of 32, where we also use data augmentation including random cropping of 88×88 pixels, random horizontal flip and mixup [32]. For the proposed VCVTS, mel-spectrograms are converted to waveform by using Griffin-Lim (GL) [33] or Parallel WaveGAN (PWG) [34] vocoder that is trained on LibriTTS [35]. We compare VCVTS with two multi-speaker VTS baseline systems, i.e., Lip2Wav [16] and XTS [17], both of which use GL following their papers, where XTS is only implemented on GRID as its training requires speaker labels that are unavailable in LRW.

### 3.1. Objective evaluation results

Following previous works [16, 17], we adopt three standard speech quality metrics: perceptual evaluation of speech quality (PESQ) [36] to measure the quality, Short-Time Objective Intelligibility (STOI) [37] and Extended STOI (ESTOI) [38] to measure the speech intelli-
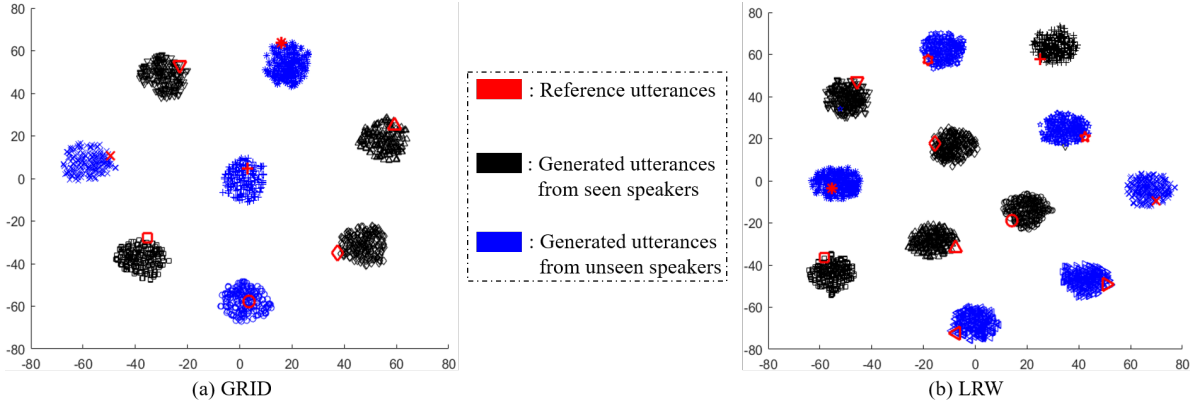
---

[2]https://github.com/resemble-ai/Resemblyzer

**Fig. 3**. Visualization of speaker representations of reference and generated utterances, where reference utterances from different speakers are highlighted with larger, red markers with different shapes, generated utterances are highlighted with smaller markers and the same shape as their corresponding reference utterance that is used to control the speaker identity.

gibility. We also compute mel-cepstral Distortion (MCD) and Root Mean Square Error of $F_0$ ($F_0$-RMSE). Table 1 shows the objective evaluation results of different VTS systems on testing speakers that are seen or unseen during training.

For the GRID, we can observe that compared with results for the unseen testing speakers, all VTS systems achieve better performance on the seen testing speakers by a great margin, this shows the susceptibility of VTS systems to out-of-domain speakers. The proposed VCVTS using GL achieves highest scores for both seen and unseen speakers on PESQ, STOI, ESOI and $F_0$-RMSE, which shows that the Lip2Ind network can accurately infer the indices of discrete acoustic units for generating the speech with higher quality and intelligibility, and the proposed pitch predictor can be used to infer realistic $F_0$ values to control the pitch contour to be closer to that of original speech. Besides, although the proposed VCVTS using the PWG vocoder outperforms baseline systems on most metrics, it is inferior to the proposed VCVTS using GL except for MCD, as PWG is based on neural networks and tends to introduce acoustic artifacts that degrade the objective performance. For the LRW, it is encouraging to see that VCVTS using GL or PWG outperforms the start-of-the-art VTS system, i.e., Lip2Wav, this indicates that under unconstrained conditions, the proposed system can still employ the Lip2Ind network and VQ-codebook to obtain accurate acoustic units used to restore the spoken content. Consequently, the generated speech has more accurate spectral details and pitch contour.

### 3.2. Subjective evaluation results

Subjective listening tests have been conducted to give mean opinion score (MOS) for Speech Naturalness (MOS-SN) and Speaker Similarity (MOS-SS). 20 subjects are recruited to give 5-point MOS, i.e., 1-bad, 2-poor, 3-fair, 4-good, 5-excellent. For both GRID and LRW, 60 generated utterances are randomly selected from different systems for MOS tests, where 15 utterances are selected for each of 4 testing speakers of GRID, and the results are reported in Table 1.

For the GRID, we observe that compared with results on seen testing speakers, the performance on unseen testing speakers of all VTS systems deteriorates with significant degradation of speaker similarity, which is caused by the limited number of speakers of GRID for training a VTS system that generalizes poorly to out-of-domain speakers. This problem can be alleviated for LRW with hundreds of speakers used for training, where both Lip2Wav and VCVTS achieve higher MOS of speaker similarity on LRW than that on GRID. Besides, We can see that VCVTS still outperforms the Lip2Wav and XTS with higher MOS-SN and MOS-SS, this shows

that the proposed VCVTS inherits the advantages of VC to generate the speech with higher naturalness and voice similarity. Besides, for the proposed VCVTS, PWG is superior to GL, this shows that PWG tends to generate the waveform that sounds more realistic, which improves the perceptual quality of human.

### 3.3. Visualization of speaker representation

To further contextualize the speaker similarity evaluation results illustrated in section 3.2, we select the reference utterances from different speakers that are seen or unseen during training to control the speaker identities of generated utterances, where lip sequences are randomly selected from testing speakers. Then the speaker representations of reference and generated utterances are extracted by using the speaker encoder used in the proposed VCVTS and visualized via t-SNE [39] in the embedding space, as shown in Fig. 3. We observe that for both GRID and LRW, each reference utterance and its corresponding generated utterances can form a distinct cluster, this further shows that the proposed multi-speaker VTS system can use the speaker encoder to produce the speaker representation that effectively captures the speaker characteristics.

## 4. CONCLUSIONS

We propose a novel multi-speaker VTS system, i.e., VCVTS, which can accurately reconstruct the spoken content from the lip motions and effectively control the speaker identity of generated speech. This is achieved by transferring the knowledge from VC, using the VQCPC-based content encoder of VC to guide the learning of a Lip2Ind network, and using the speaker representation produced by a speaker encoder to capture desired speaker characteristics. Compared with existing works, the proposed VTS system provides a more transparent mapping process from the lips to speech, quantitative and qualitative results show that state-of-the-art performance can be achieved based on various objective and subjective metrics under both constrained (e.g., GRID) and unconstrained (e.g., LRW) conditions. Our future work aims to study more challenging conditions for VTS including cross-domain and multi-lingual scenarios.

# 6. REFERENCES

[1] Ben Milner and Thomas Le Cornu, "Reconstructing intelligible audio speech from visual speech features," in *INTERSPEECH*, 2015, pp. 3355–3359.

[2] Thomas Hueber and Gérard Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance hmm," *Computer Speech & Language*, vol. 36, pp. 274–293, 2016.

[3] Ryo Aihara, Kenta MASAKA, Tetsuya TAKIGUCHI, and Yasuo ARIKI, "Lip-to-speech synthesis using locality-constraint non-negative matrix factorization," in *Proc. MLSLP*, 2015.

[4] Rina Ra, Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, "Visual-to-speech conversion based on maximum likelihood estimation," in *Proc. IAPR-MVA*. IEEE, 2017, pp. 518–521.

[5] Thomas Le Cornu and Ben Milner, "Generating intelligible audio speech from visual speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1751–1761, 2017.

[6] Ariel Ephrat and Shmuel Peleg, "Vid2speech: speech reconstruction from silent video," in *ICASSP*. IEEE, 2017, pp. 5095–5099.

[7] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg, "Improved speech reconstruction from silent video," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 455–462.

[8] Yuki Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based lip-to-speech synthesis using convolutional neural networks," in *Proc. IW-FCV*, 2019.

[9] Yaman Kumar, Mayank Aggarwal, Pratham Nawal, Shin'ichi Satoh, Rajiv Ratn Shah, and Roger Zimmermann, "Harnessing ai for speech reconstruction using multi-view silent video feed," in *ACM MM*, 2018, pp. 1976–1983.

[10] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *ICASSP*. IEEE, 2018, pp. 2516–2520.

[11] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen, "Vocoder-based speech synthesis from silent videos," *arXiv preprint arXiv:2004.02541*, 2020.

[12] Yaman Kumar, Rohit Jain, Khwaja Mohd Salik, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann, "Lipper: Synthesizing thy speech using multi-view lipreading," in *AAAI*, 2019, vol. 33, pp. 2588–2595.

[13] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Video-driven speech reconstruction using generative adversarial networks," *arXiv preprint arXiv:1906.06301*, 2019.

[14] Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W Schuller, and Maja Pantic, "End-to-end video-to-speech synthesis using generative adversarial networks," *arXiv preprint arXiv:2104.13332*, 2021.

[15] Ravindra Yadav, Ashish Sardana, Vinay P Namboodiri, and Rajesh M Hegde, "Speech prediction in silent videos using variational autoencoders," in *ICASSP*. IEEE, 2021, pp. 7048–7052.

[16] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *CVPR*, 2020, pp. 13796–13805.

[17] Dan Oneata, Adriana Stan, and Horia Cucu, "Speaker disentanglement in video-to-speech conversion," *arXiv preprint arXiv:2105.09652*, 2021.

[18] 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.

[19] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[20] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 87–103.

[21] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng, "VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion," in *INTERSPEECH*, 2021, pp. 1344–1348.

[22] Benjamin van Niekerk, Leanne Nortje, and Herman Kamper, "Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge," in *INTERSPEECH*, 2020, pp. 4836–4840.

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[24] Laurenz Wiskott and Terrence J Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, no. 4, pp. 715–770, 2002.

[25] Benedikt Gierlichs, Lejla Batina, Pim Tuyls, and Bart Preneel, "Mutual information analysis," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2008, pp. 426–442.

[26] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1779–1788.

[27] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[28] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[29] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Lipreading using temporal convolutional networks," in *ICASSP*. IEEE, 2020, pp. 6319–6323.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[32] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[33] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *ICASSP*, vol. 32, no. 2, pp. 236–243, 1984.

[34] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*. IEEE, 2020, pp. 6199–6203.

[35] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[36] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.

[37] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*. IEEE, 2010, pp. 4214–4217.

[38] Jesper Jensen and Cees H Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[39] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.