Let's Chorus: Partner-aware Hybrid Song-Driven 3D Head Animation —Supplementary Materials—

Xiumei Xie[1] Zikai Huang[1] Wenhao Xu[1] Peng Xiao[1] Xuemiao Xu[1,2†] Huaidong Zhang[1,2†]
[1]South China University of Technology
[2] Guangdong Engineering Center for Large Model and GenAI Technology

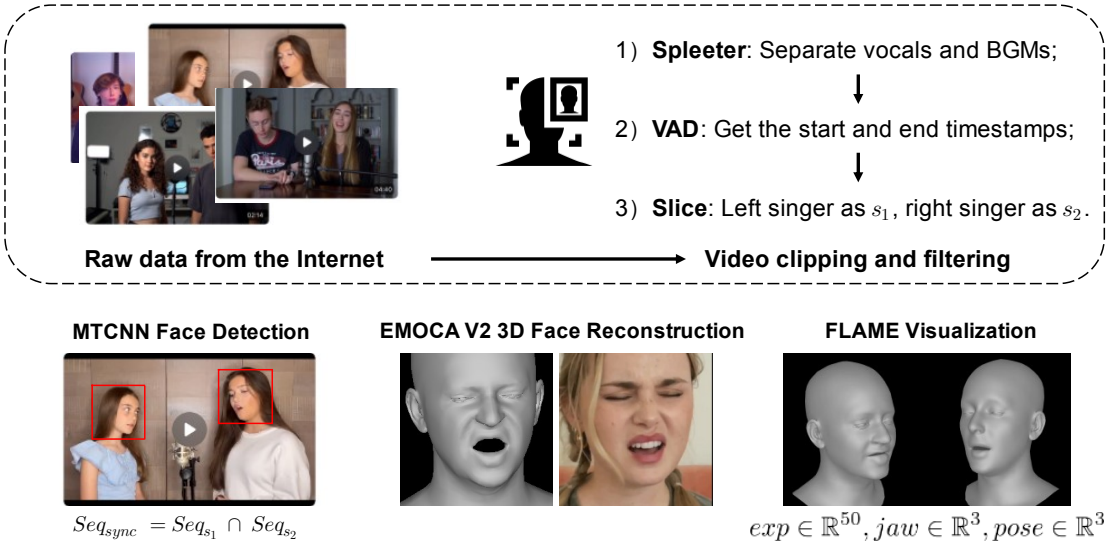

Figure 1. Overview of ChorusHead Dataset Collection. An efficient workflow includes two stage: data preparation and preprocessing.

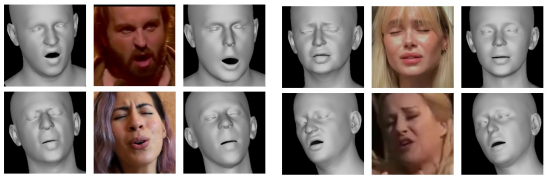

Figure 2. EMOCA(leftmost) vs. EMICA(rightmost).

1. ChorusHead Dataset

Figure 1 illustrates the overview of ChorusHead Data Collection, which consists of two main stages: data preparation and preprocessing. Voice activity detection(VAD) [1] is used to detect and filter valid segments from the raw data. In the initial phase, to select an appropriate facial reconstruction model, we conducted extensive comparisons and finally adopted EMOCA [2], as shown in Fig. 2. The left panel represents EMOCA, while the right panel illustrates EMICA, a combination of DECA [3], EMOCA, SPECTRE [4] and MICA [5]. We observed that EMOCA achieves a balance between facial expression and mouth reconstruction accuracy, which is particularly crucial in singing scenarios where facial expressions play a vital role. The reconstructed FLAME [6] facial coefficients is decomposed into facial expression coefficients $\alpha \in \mathbb{R}^{50}$, jaw coefficients $\theta \in \mathbb{R}^3$ and pose coefficients $\beta \in \mathbb{R}^3$. In methodlogy section, we classify jaw coefficients as part of the expression coefficients, thus representing the expression coefficients as $\alpha \in \mathbb{R}^{50+3}$. Therefore, we denote each singer's motion as $f^i = \{\alpha, \beta\} \in \mathbb{R}^{D=53+3=56}$.

2. More Results

For more visualized results generated by PaChorus, please refer to https://xxiexm.github.io/PaChorus/.

†Corresponding author (*xuemx@scut.edu.cn, huaidongz@scut.edu.cn*).

References

[1] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 1

[2] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *CVPR*, pages 20311–20322, 2022. 1

[3] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 1

[4] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022. 1

[5] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 1

[6] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 1