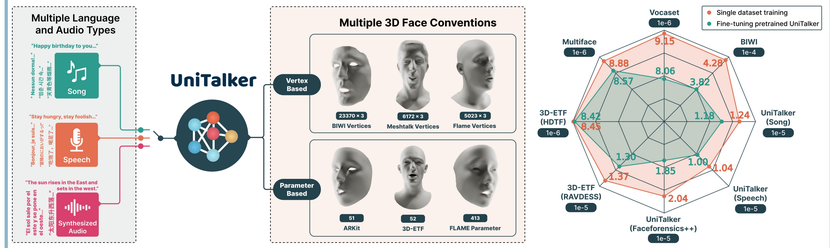
UniTalker: Scaling up Audio-Driven 3D Facial Animation through A Unified Model

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, Lei Yang

SenseTime Research, China

Introduction



- **Problem:** Inconsistent data annotation. Insufficient data variety.
- **Objective:** Employ multiple datasets. Eliminate data pre-processing.
- **Key components**: Training Strategy, Model Design.

Results

Quantitative results on BIWI-Test-A and VOCA-Test.

Dataset	Method	LVE ↓ $\times 10^{-4}$	MVE ↓ $\times 10^{-3}$	UFVE ↓ $\times 10^{-3}$	FDD ↓ $\times 10^{-5}$	Params M	Time s
BIWI	FaceFormer	4.9836	7.2750	6.9081	4.0062	109	0.705
	CodeTalker	4.7914	7.3784	7.0050	4.2147	561	4.4
	SelfTalk	4.2485	6.9152	6.5428	**3.5851**	539	0.071
	FaceDiffuser	4.2985	6.8088	6.6220	3.9101	189	16.50
	UniTalker-B-[D0]	4.3681	6.8948	6.6277	4.6789	92	0.024
	UniTalker-B-[D0-D7]	4.0804	6.6458	6.3774	5.0438	92	0.024
	UniTalker-L-[D0-D7]	**3.8587**	**6.4166**	**6.1483**	5.2307	313	0.054

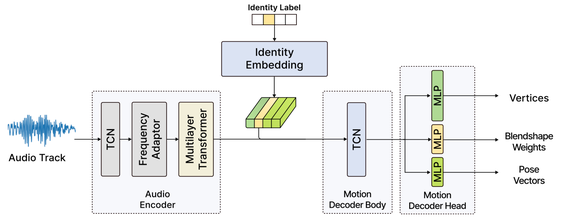
Dataset	Method	LVE ↓ $\times 10^{-5}\ m^2$	MVE ↓ $\times 10^{-3} m$	UFVE ↓ $\times 10^{-3} m$	FDD ↓ $\times 10^{-7} m^2$	Params M	Time s
Vocaset	FaceFormer	1.1696	0.6364	0.4972	2.4812	92	0.624
	CodeTalker	1.1182	0.5750	0.4708	1.2594	315	3.464
	SelfTalk	0.9626	0.5665	0.4805	**1.0511**	450	0.053
	FaceDiffuser	0.9684	0.5768	0.4772	1.7335	89	13.08
	UniTalker-B-[D1]	0.9381	0.5695	0.4829	1.2115	92	0.022
	UniTalker-B-[D0-D7]	**0.8136**	**0.5338**	**0.4494**	1.3962	92	0.022
	UniTalker-L-[D0-D7]	0.8303	0.5524	0.4756	1.5206	313	0.053

Quantitative comparison between single dataset training and mixed dataset training. The metric is LVE. L-[D*] denotes the eight individual models trained on each dataset. L-[D0-D7] denotes UniTalker-Large trained on A2F-Bench. L-FT denotes the eight models finetuned from L-[D0-D7]

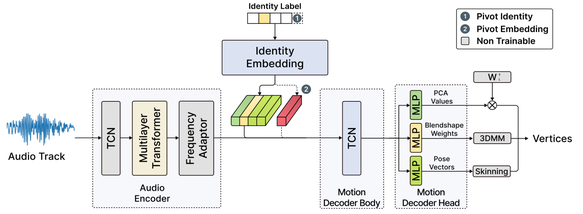
Method	D0 0.33h	D1 0.56h	D2 0.67h	D3 5.49h	D4 1.48h	D5 3.65h	D6 1.24h	D7 5.11h
L-[D*]	4.279	9.153	8.881	8.445	1.370	2.040	1.043	1.235
L-[D0-D7]	3.859↓9.8%	8.303↓9.3%	8.648↓2.6%	8.991↑6.5%	1.326↓3.2%	2.056↑0.8%	1.145↑9.7%	1.211↓1.9%
L-FT	3.816↓11%	8.060↓12%	8.56↓3.5%	8.417↓0.3%	1.30↓5.2%	1.848↓9.4%	0.998↓4.3%	1.178↓4.6%

Methodology

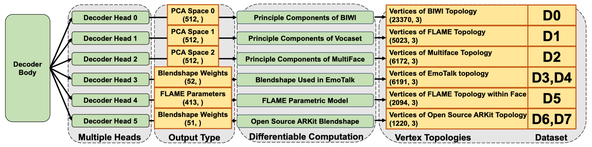
Model Architecture



(a) Vanilla Multi-Head Model

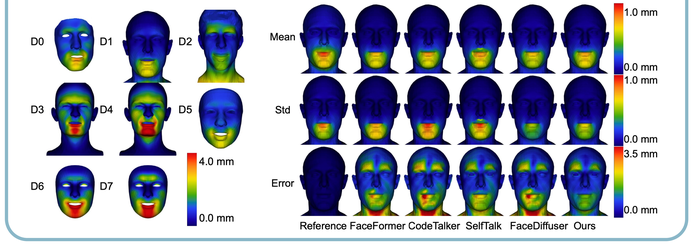


(b) UniTalker Model



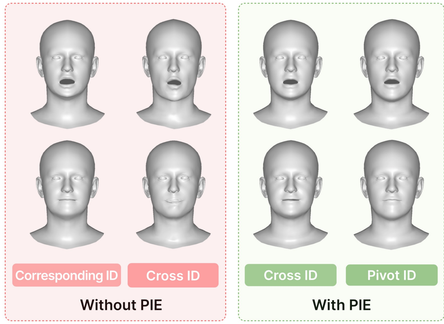
(c) Zoomed-in View of UniTalker Decoder

(a) The standard deviation of facial motion within each training set. The upper face of D1(Vocaset) shows little motion variation and is close to static. (b) The temporal statistics (mean and standard deviation) of adjacent-frame motion variation and the mean of per-frame predicted-to-GT Euclidean distance within a sequence.

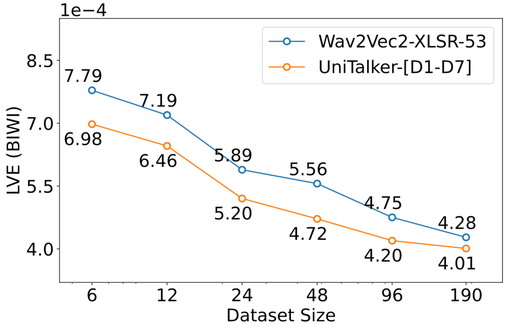


Results

Effect of PIE. Without PIE, the model generates unnatural face motion when input identity and out- put annotation mismatch.



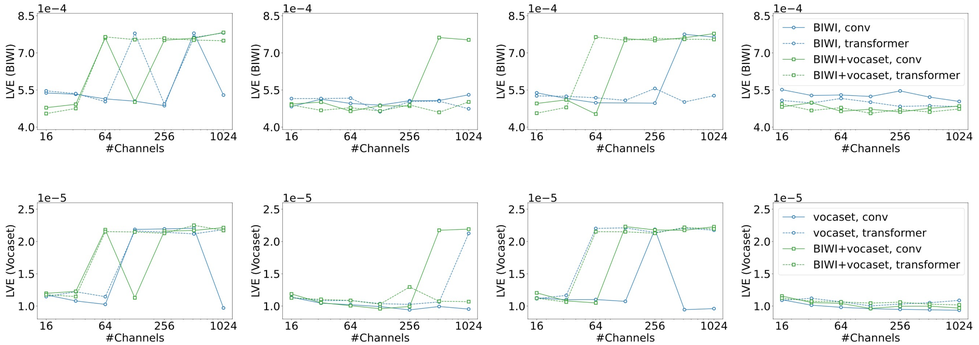
Comparison between finetuning Wav2vec2-xlsr-53 and UniTalker-L- [D1-D7] on D0. The x-axis is in log-scale.



The effect of pre-trained audio encoders.

Audio Encoder	D0	D1	D2	D3	D4	D5	D6	D7
Wav2Vec2-Base-960h [3]	4.491	9.916	9.887	9.812	1.585	2.217	1.351	1.409
WavLM-Base [11]	4.033	8.269	9.253	9.117	1.417	2.044	1.184	1.340
WavLM-Base-Plus [10]	4.080	**8.136**	9.776	9.053	1.392	**1.975**	1.158	1.264
Wav2Vec-XLSR-53 [12]	**3.859**	8.303	**8.648**	**8.991**	**1.326**	2.056	**1.145**	**1.211**

The effect of PCA and DW. LVE values are evaluated on test set at 100th epoch. Training with both PCA and DW ensures training stability. Removing either strategy harms training robustness.



(a) w.o. PCA, w.o. DW (b) w.o. DW (c) w.o. PCA (d) w. PCA, w. DW