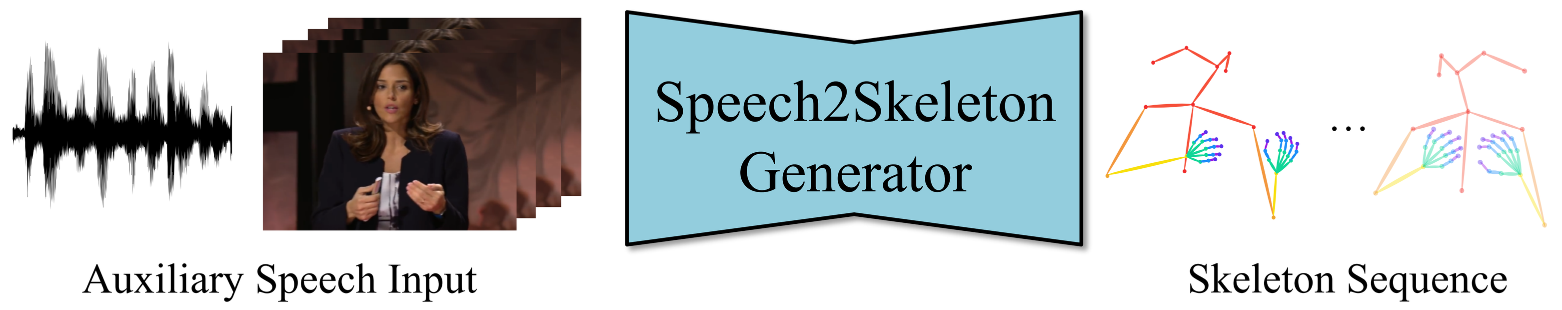
Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, Bolei Zhou

香港中文大學 The Chinese University of Hong Kong · MONASH University · Zhejiang University · sensetime

Co-Speech Gesture Motivation

- **Task Definition**: Given a clip of speech information as input, we predict the 3D skeleton sequence that is aligned with the speech.

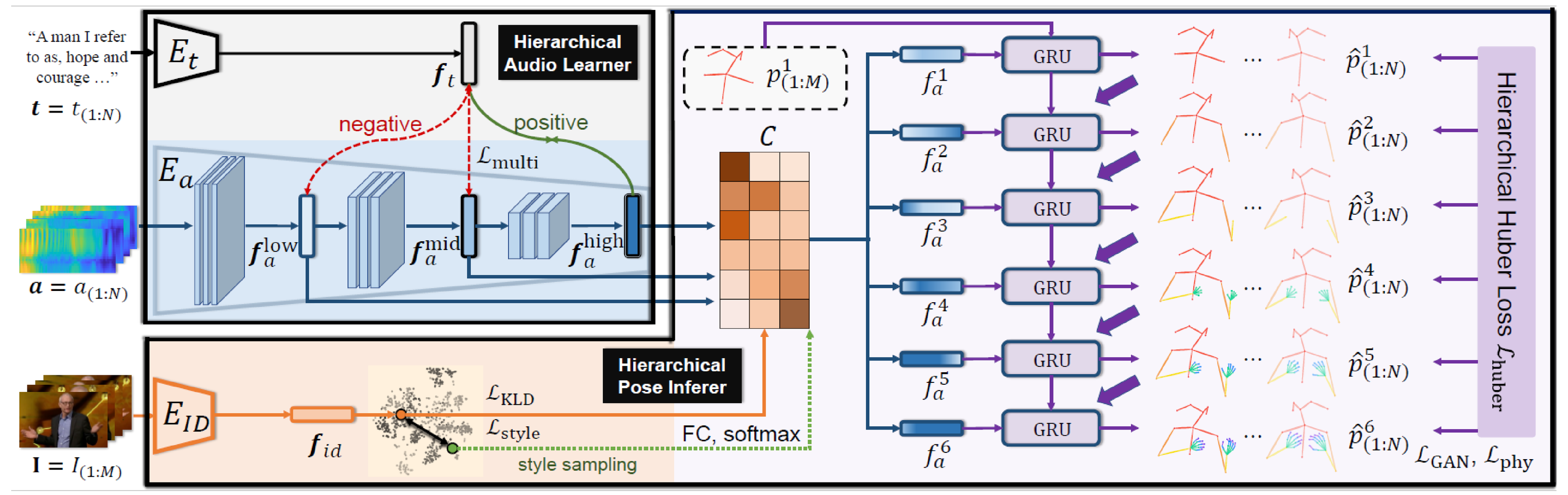


- **Key Observations**: 1) Different co-speech gestures are related to *distinct levels* of audio. For example, the metaphorical gestures are associated with the high-level speech semantics (e.g., when depicting a ravine, one would moving two outstretched hands apart and saying gap"), while the low-level audio features of beat and volume lead to the rhythmic gestures. 2) The dynamic patterns of different body parts are not the same, such as the flexible fingers and relatively still upper arms. Instead of holistically generating the whole skeleton, we should *treat each part differently*.

Our solution: Capture hierarchical audio-pose associations!

Framework

Overview



Key Components

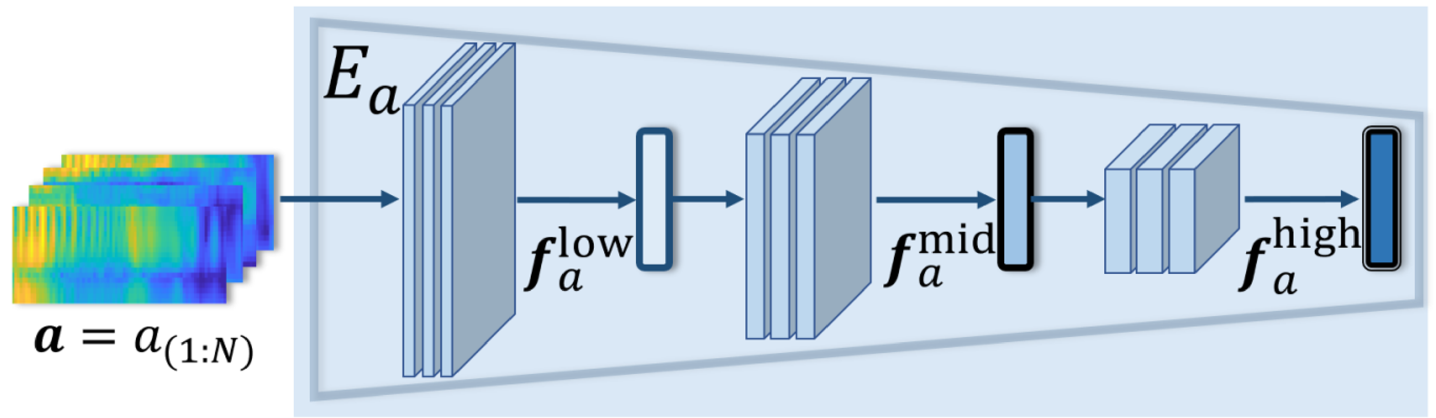
★ Hierarchical Audio Learner: Encode multi-level audio feature to extract both rhythmic and semantic information.

★ Hierarchical Pose Inferer: Infer gestures hierarchically and capture associations between multi-level audios and poses.

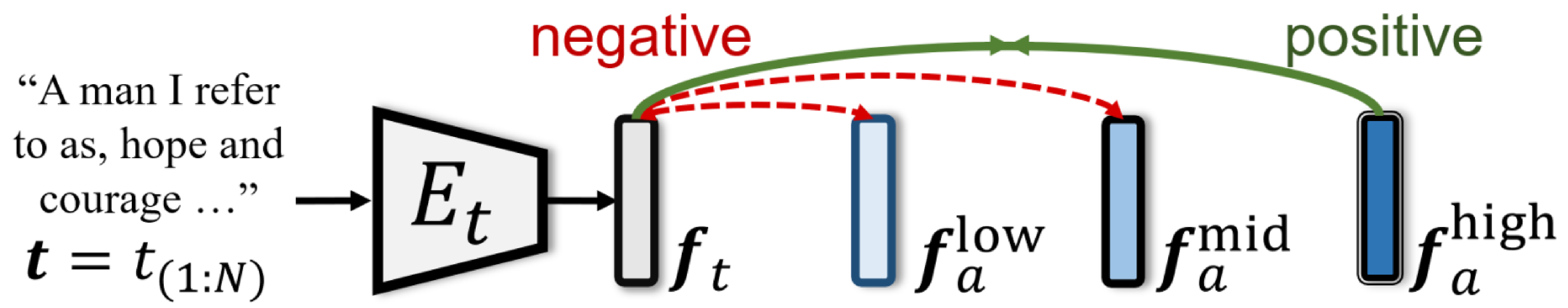
Our Approach

Hierarchical Audio Learner

- Hierarchical Audio Feature Extraction: We take output of shallow, middle, deep encoder layers as low, middle, high level features f_a^{low}, f_a^{mid}, f_a^{high}:



- Contrastive Learning Strategy: We take text feature f_t with high-level feature f_a^{high} as positive pairs; with low/mid level f_a^{low}, f_a^{mid} as negative pairs:

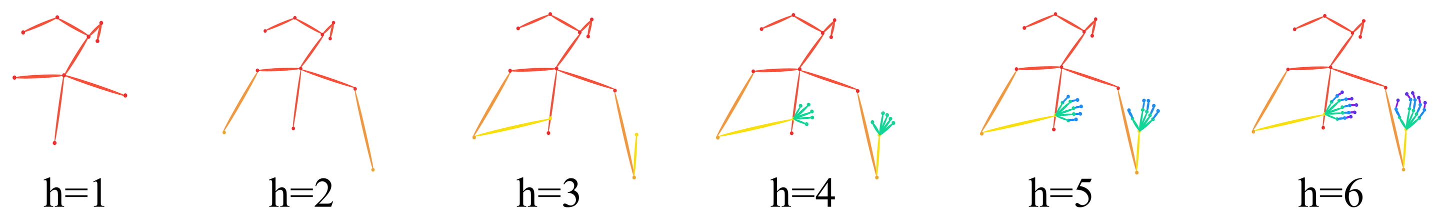


Hierarchical Pose Inferer

- Multi-Level Feature Blending: Style coordinator $C \in \mathbb{R}^{3\times H}$ controls ratio between hierarchical audio features and each level of motion hierarchy.

$$f_a^h = C[1,h] * f_a^{low} + C[2,h] * f_a^{mid} + C[3,h] * f_a^{high}$$

- Coarse-to-Fine Pose Generation: We design a $H = 6$ level body hierarchy and predict from previous level's inferred pose and current level's audio. In this way, fine-grained gesture is learned in a coarse-to-fine manner.



$$\hat{p}_i^h = \left[h_i; \hat{p}_i^{h-1}; f_{a(i)}^h\right] * W^h + b^h, h_i = \mathrm{GRU}(h_{i-1}, \hat{p}_{i-1}^h)$$

Experiments

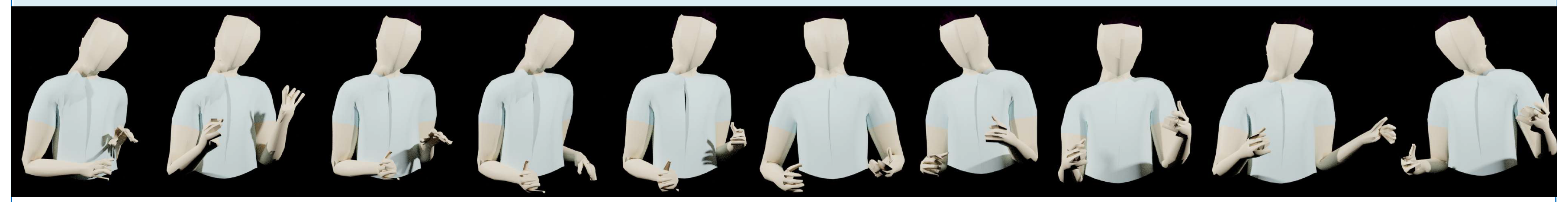
Quantitative Comparisons

Dataset	TED Gesture			TED Expressive		
Method	FGD	BC	Div.	FGD	BC	Div.
GT	0	0.795	110.821	0	0.723	175.231
Atten seq2seq	18.154	0.186	92.176	54.920	0.155	122.693
S2G	19.254	0.764	98.095	54.650	0.714	142.489
Joint Embed.	22.083	0.177	91.223	64.555	0.131	120.627
Trimodal	3.729	0.688	102.539	12.613	0.592	154.088
HA2G (Ours)	**3.072**	**0.769**	**108.086**	**5.306**	**0.715**	**173.899**

Ablation Study

Settings	f_a^{low}	f_a^{mid}	f_a^{high}	w/o text	w/o f_a^{high}	w/o f_a^l, f_a^m
FGD	6.588	7.212	7.421	9.228	7.982	6.998
BC	0.704	0.682	0.661	0.619	0.652	0.701
Diversity	171.482	168.223	165.741	158.236	163.649	169.021
Settings	holistic	w/o hand	w/o body	same f_a^h	ASR	**HA2G**
FGD	11.989	10.832	5.882	6.801	5.319	**5.306**
BC	0.594	0.606	0.709	0.701	**0.716**	**0.715**
Diversity	156.079	158.823	173.066	170.085	173.058	**173.899**

Qualitative Results



Conclusion with Github, Project Page

- In this paper, we propose a novel framework **HA2G** with Hierarchical Audio Learner and Hierarchical Pose Inferer for fine-grained co-speech gesture generation.

Github

Project Page