Indian Institute of Technology Jodhpur

Computer Science Department

M.TECH, ARTIFICIAL INTELLIGENCE (Exe.)

NAME – TORSHA CHATTERJEE
ROLL NO. - M23CSA536

Speech Understanding

Assignment - 3

Paper Review

"LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures"

1. Title of the Paper

LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures

2. Summary of the Paper

The paper titled “LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures” introduces GesTran, a novel LLM-driven framework that translates spoken language into meaningful human gestures. Unlike traditional co-speech gesture generation methods that rely on regression-based or rule-based models, this approach reframes gesture generation as a multimodal translation problem, treating gestures as a form of "body language" that can be tokenized, processed, and decoded using Large Language Models (LLMs).

The pipeline begins with a Vector Quantized Variational Autoencoder (VQ-VAE) that converts continuous gesture sequences into discrete tokens, effectively forming a gesture vocabulary. These tokens, similar to language words, are then used as inputs/outputs for a pre-trained LLM. The LLM receives the speech signal (both audio and transcribed text) and is fine-tuned using LoRA (Low-Rank Adaptation) to predict gesture token sequences autoregressively, akin to machine translation.

Through this formulation, the model not only learns the semantic and rhythmic alignment between speech and gestures but also benefits from the zero-shot generalization ability of LLMs. It can generate diverse and expressive gestures even for previously unseen speech patterns. The proposed method significantly outperforms existing baselines on multiple metrics (FGD, BC, Diversity) across two benchmark datasets (TED Gesture and TED Expressive), and also shows robust performance on zero-shot test sets. Additionally, user studies confirm that the generated gestures are more natural, smooth, and better synchronized with speech than prior methods.

3. Figure Representing the Main Architecture

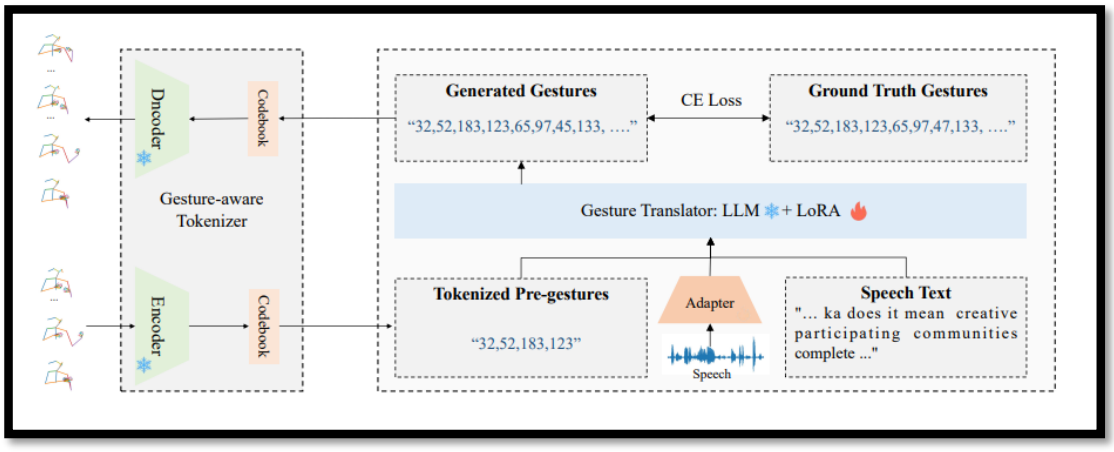

Figure 1The overall framework of GesTran (extracted from the paper)

4. Technical Strengths of the Paper

- **Innovative Paradigm**: Reframes gesture generation as a translation task using LLMs, introducing a linguistic abstraction of gestures.
- **Discrete Gesture Representation**: The use of **VQ-VAE** for tokenizing gesture sequences allows LLMs to treat gestures as "words", enabling efficient translation.
- **Multimodal Integration**: Combines audio, text, and gesture tokens within a single LLM-based pipeline.
- **Zero-shot Generalization**: Demonstrates superior performance on unseen data through language model generalization.
- **State-of-the-art Results**: Outperforms all baselines in key metrics (FGD, BC, Diversity) on multiple datasets.
- **User Study Validation**: Human evaluators corroborate the effectiveness in terms of naturalness, synchrony, and smoothness.

5. Technical Weaknesses of the Paper

- **Language Dependency**: The model is trained exclusively on English, limiting its multilingual applicability.
- **No Real-Time Performance Analysis**: Lacks benchmark on inference speed and resource efficiency, which is important for practical deployment.
- **Fine-Grained Gesture Accuracy**: No explicit analysis of how well fine motor details (e.g., finger articulation) are captured.
- **Over-reliance on Pretraining**: Performance significantly drops without pretraining (as seen in ablation studies), suggesting strong dependency on LLM's pretrained knowledge.

6. Minor Questions / Minor Weakness

- How robust is the model to noisy audio or speech disfluencies?
- Is the gesture lexicon learned by VQ-VAE interpretable or meaningful semantically?
- Would combining additional modalities (e.g., facial expressions) further enhance realism?

7. Reviewer's Suggestions and Rating

a.Suggestions

The authors should explore training the model on multilingual or cross-lingual data to expand its applicability. Introducing real-time performance metrics would also enhance its practical value, especially for deployment in interactive systems or robotics. Further, integrating fine-grained

gesture controls (e.g., hand or finger gestures) or multi-modal cues (like facial expressions) could boost expressive richness. Additional ablations on gesture token interpretability might help understand the learned gesture vocabulary. Lastly, evaluating robustness to noise or spontaneous speech could improve the model's real-world readiness.

b. Rating with Justification

Rating: 4.5 / 5

The paper presents a novel and technically sound approach to gesture generation, achieving state-of-the-art results. Its strength lies in the effective use of LLMs, but improvements in generalization, language diversity, and real-time feasibility could make it even stronger.

Github Link:

https://github.com/M23CSA536-Codes/Speech_Understanding_Assignment_PA3

--- THANK YOU ---