The Hong Kong Polytechnic University

Department of Electrical and Electronic Engineering Innovative

AI Applications (2025 – 26)

Proposal

Project Title: Multimodal Virtual Communication System

Candidate Information: Ma Kai Lun Donovan (24024192d) (Leader)

Abstract

This project proposes the design of a Multimodal Virtual Communication System integrating facial expression recognition (FER), acoustic-phonological modeling, expressive speech generation, and personality reconstruction. The system aims to support naturalistic, emotionally meaningful human–machine interaction. Methodologically, the project employs knowledge distillation for FER robustness, generative phonology models for prosody, zero-/few-shot text-to-speech for expressive synthesis, and persona-conditioned dialogue for interaction style. Expected outcomes include validated FER, a prosody module, expressive TTS, a persona prototype, and an integrated multimodal system. Applications span remembrance, healthcare, and self-reflection, contributing to affective computing and human-centered AI.

Introduction

Human communication is inherently multimodal, combining speech, facial expression, prosody, and subtle paralinguistic cues. While unimodal systems have advanced, integration remains limited. This project envisions a Virtual Human Communication System capable of simulating human-like interaction across voice, dialogue, expression, and personality. Applications include remembrance, therapy, and personal reflection. The research roadmap begins with FER, extends into acoustic-phonological modeling, and culminates in multimodal integration. The central problem is how to design robust unimodal modules and progressively integrate them into a trustworthy, emotionally meaningful system.

Related Work

FER has evolved from handcrafted features to CNNs and Vision Transformers, with knowledge distillation improving efficiency but robustness under stress still limited. Acoustic-phonological modeling has applied GANs and transformer models to vowel length contrasts and prosody, yet individuality and

cross-linguistic variation remain challenging. Speech generation has advanced through zero-/few-shot TTS systems such as FastSpeech2, VITS, and flow/diffusion models. Persona reconstruction has been explored via datasets like PersonaChat, with conditioning on Big Five traits and coherence models. Multimodal affective computing employs fusion strategies but remains task-specific, lacking holistic integration.

Methodology

- **FER Module:** Teacher–student distillation (ResNet, EfficientNet, ConvNeXt → ViT-Tiny), ArcFace loss, stress testing.
- **Acoustic Module:** x-vector/ECAPA embeddings, prosody encoder, GAN phonology, few-shot adaptation.
- **Speech Generation:** FastSpeech2/VITS baseline, flow/diffusion models, prosody factorization.
- **Persona Module:** LLM backbone, persona embeddings, Big Five traits, variational inference, coherence.
- **System Integration:** Transformer-based fusion, LLM dialogue agent, streaming inference.

Evaluation includes accuracy, calibration, MOS, similarity, latency, and user studies.

Significance

Academically, the project advances FER robustness, expressive phonology modeling, and multimodal integration. Practically, it enables remembrance technologies, healthcare support, and self-reflection tools. Ethically, it addresses risks of misuse, authenticity, and trust, embedding safeguards into design.

Project Plan

Objectives: robust FER, acoustic modeling, expressive TTS, persona reconstruction, integration.

Timeline: staged development from FER to integration.

Risks: data scarcity, ethical misuse, deployment drift.

Expected outcomes: validated FER, prosody module, expressive TTS, persona prototype, integrated system.

Estimated Budget: Unknow