ellis unit | TURIN Talk

Politecnico di Torino

AiH ARTIFICIAL INTELLIGENCE HUB POLITECNICO DI TORINO

VANDAL

http://vandal.polito.it

www.ellis.eu

Patrick Pérez

CEO @ Kyutai, Paris, France

May 12th, 2025
starting at 10:00 AM cet
Link:
https://tinyurl.com/3wwvj7ks

A multistream multimodal foundation model for real-time voice-based applications

A unique way for humans to seamlessly exchange information and emotion, speech should be a key means for us to communicate with and through machines. This is not yet the case. In an effort to progress toward this goal, we introduce a versatile speech-text decoder-only model that can serve a number of voice-based applications. It has in particular allowed us to build Moshi, the first-ever full-duplex spoken-dialogue system (with no latency and no imposed speaker turns) as well as Hibiki, the first simultaneous voice-to-voice translation model with voice preservation to run on a mobile phone. This multistream multimodal model can also be turned into a visual-speech model (VSM) via cross-attention with visual information, which allows Moshi to freely discuss about an image while maintaining its natural conversation style and low latency. This talk will provide an illustrated tour of this research.

Patrick Pérez is CEO at Kyutai, a non-profit open-science AI lab, based in Paris. Prior to this, Patrick was at Valeo as VP of AI and Scientific Director of valeo.ai (2018-2023), and with Technicolor (2009-2018), Inria (1993-2000, 2004-2009) and Microsoft Research Cambridge (2000-2004) as research scientist. His research interests lie in reliable multimodal AI for the benefit of all.