Warmblood: A Text-Guided Full-Duplex Spoken Dialogue System

Author: Mgr. Eliáš Cizl Supervisor: Mgr. et Mgr. Ondřej Dušek, Ph.D.

FACULTY OF MATHEMATICS AND PHYSICS Charles University

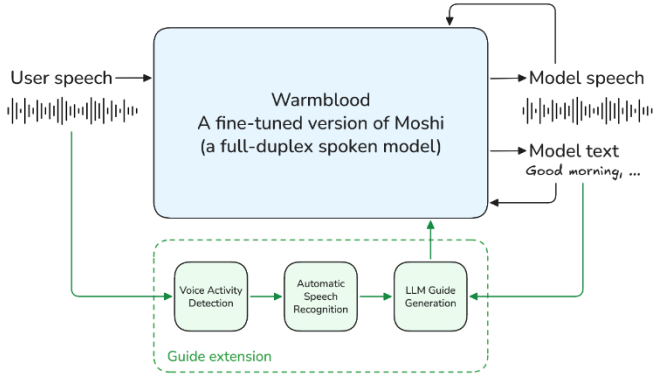
Motivation

Talking to a computer should feel like talking to a person – but voice assistants have long forced an uncomfortable trade-off. Text-based LLMs are genuinely intelligent, reasoning well and answering with high factual quality, and cascaded voice assistants built around them keep much of that intelligence – but they remain strictly turn-based, with the stiff "walkie-talkie" feel of waiting for your turn to speak. **Fully duplex models** go the other way: able to genuinely listen and talk at the same time, interrupt naturally, and backchannel like a real conversation partner, with *Moshi* as one of the most influential open examples – but that naturalness comes at the cost of intelligence, since everything happens inside one audio model with no access to reliable facts. These systems can easily say something factually wrong, or even drift into incoherence mid-conversation – repeating a word or a phrase over and over with no way out short of restarting the whole session. This thesis asks: can we combine the two strengths – the natural, interruptible flow of a full-duplex assistant with the reasoning and factual reliability of a text-based LLM?

Approach

The answer is **Warmblood**, a fine-tuned version of Moshi, extended with a new **"guide"** channel. Alongside the audio streams Moshi already processes for the user and for itself, Warmblood adds a **dedicated stream of short text hints** – facts, numbers, or instructions – generated on the fly by a separate, fast text-based LLM watching a live transcript of the conversation. This guide text is injected the moment the user finishes speaking, without ever interrupting Warmblood's ability to listen and talk simultaneously; think of it as a rider gently steering a spirited horse – the animal still moves on its own, but a firm hand keeps it on course (hence the name, a nod to the Czech Warmblood horse breed). Since no dataset combining natural two-party dialogue with guide annotations existed, a data pipeline was built to synthesize conversational audio directly from text dialogues paired with matching guide annotations, complemented by existing natural recordings used to help the model retain its natural conversational abilities.

Architecture



Experiments & Results

Warmblood was tested in three stages. Internal experiments confirmed the **model genuinely learns to exploit the guide**: with an immediate hint, answer correctness and coherence consistently beat scenarios with a delayed or missing guide, across every dataset tried. It was then evaluated on *VoiceAssistant-Eval*, a public benchmark for general voice-assistant intelligence – covering instruction following, assistance, reasoning, and everyday question answering – where Warmblood outperformed the original Moshi across the board, including robustness and safety. Results were more nuanced on *Full-Duplex-Bench*, which probes live interaction quality: pause handling and backchanneling improved, but smooth turn-taking and response speed slipped slightly, since the model now has to weigh its own instinct against incoming guidance. Finally, a 100-participant **user study** had real people talk to the system: it scored well on truthfulness and coherence but landed only around the middle on naturalness and overall impression, revealing a familiar gap between what benchmarks measure and how a system actually feels to talk to.

Example

Without guide

User: How many hearts does an octopus have?
Warmblood: I think it has more than one, though I'm not sure exactly how many.

With guide

User: How many hearts does an octopus have?
Warmblood: I think it's
(guide, invisible to the user): three, two gills, one body
Warmblood: three. Two pump blood to the gills, and one to the rest of the body.

User Interface

Why it matters

The everyday promise here is a **voice assistant you can actually trust** with facts – a receptionist, customer-support agent, or in-car assistant that answers naturally and quickly, but grounds its answers in real information rather than making things up mid-sentence, all without breaking the **fluid, interruption-friendly feel of natural speech**. The guide mechanism is also modular: it doesn't have to come from a text LLM alone – it could plug into a database, a company knowledge base, or a task-specific script, making the same architecture adaptable to many real deployment scenarios. Tellingly, since this thesis was written, full-duplex systems paired with some form of external information retrieval have begun appearing as deployed products – a sign that the direction explored here, combining natural duplex conversation with grounded external knowledge, struck on a **genuinely important idea**. Beyond the concrete system, the thesis's honest reporting of where things still fall short – the **gap between lab metrics and live user experience** – is itself a contribution, underscoring that spoken dialogue systems can only really be judged by putting them in front of people.