Evaluating Interruption Handling in Voice Agents for European Portuguese and English

Valerii Kondrashov
GoIPpro, LLC
valerii@goippro.com

Technical report GP-TR-2026-01 · Protocol version 0.1 · 27 October 2026

> Study protocol for discussion. This document reports a proposed evaluation method, not completed experiments. It has not undergone peer review. No empirical performance, participant recruitment, external endorsement or collaboration is claimed.

Abstract

Conversational voice systems must distinguish requests to yield from acknowledgements that allow speech to continue. This protocol proposes an evaluation of interruption handling in European Portuguese (pt-PT) and English, with attention to hesitation in language-learning conversations. It separates explicit stop requests, corrections, backchannels, within-turn pauses and third-party speech. The method combines synchronized two-channel audio, listener-side stop latency, response appropriateness and false-yield rates. It specifies a small exploratory pilot, reporting rules and the evidence required for a later empirical report. No new benchmark results are presented.

1. Motivation and scope

Moshi demonstrates a framework that models user and system audio as parallel streams [1]. Full-Duplex-Bench evaluates pause handling, backchanneling, turn-taking and interruption management [2]. These works motivate assessing several conversational behaviours together rather than treating the ability to stop on any detected speech as sufficient.

Our proposed contribution is a practical evaluation procedure for pt-PT and English at the complete voice-system level. An agent may detect an interruption internally while queued audio continues to reach the listener. Conversely, an agent that stops at every sound may respond quickly but mishandle acknowledgements and background speech. We therefore evaluate both yielding and continuing when appropriate.

This protocol does not establish novelty over all prior work, claim improved language learning or demonstrate that any model supports pt-PT. Eligibility and actual language behaviour must be documented for each tested system.

2. Research questions

1. How often and how quickly does the complete system yield to a deliberate stop request or correction?
2. How often does it yield unnecessarily to an acknowledgement or unrelated speech?
3. Does it take the floor during a planned hesitation before the speaker finishes?
4. After an interruption, does it address the correction or request without returning to obsolete queued content?

3. Proposed pilot

The exploratory pilot will use 12 scripted scenarios per language: three explicit stop requests, three corrections, two backchannels, two within-turn hesitation cases and two third-party-speech cases. Five runs per scenario produce 60 trials per language and system, or 120 for a system tested in both languages. These are planned counts, not completed measurements. Repeated trials with the same speaker are not independent evidence about a population.

Use a fresh session for each trial with a fixed preceding dialogue. Record the scenario order and random seed; randomize order. Intended-interruption audio should start while system playback is active, at a predefined point relative to detected playback onset. Record actual onset and any deviation. No-output trials are reported as setup failures, not discarded silently.

Scenario	Draft example	Expected behaviour
Stop request	“Stop, please.” / “Para, por favor.”	Yield and acknowledge the request as appropriate.
Correction	“Actually, I meant Porto.” / “Afinal, queria dizer Porto.”	Yield and use the corrected detail.
Backchannel	“Uh-huh.” / “Pois.”	Continue if context makes this an acknowledgement.
Hesitation	“I would like to… [pause] practise.” / “Eu gostava de… [pause] praticar.”	Allow completion of the intended turn.
Third-party speech	A separate voice addresses someone else.	Do not treat it automatically as an instruction.

Examples require review by a native pt-PT speaker before data collection. The expected action is assigned from the full context, including prosody and speaker intent; it must not be inferred solely from a word such as “pois”. Planned hesitation lengths are 0.5, 1 and 2 seconds, assigned to scenarios in the frozen trial manifest.

4. Recording and system configuration

Capture user input and audible system output as separate synchronized channels on the same clock. Prefer a digital listener-output loopback; an acoustic recording requires additional echo and leakage controls. Record sample rate, channel layout, clock synchronization method and the location of measurement.

For each run, record the model/provider and exact version, prompt and configuration hashes, VAD settings, playback buffer, codec, device, operating system, transport, region and test date. Keep model detection timestamps separate from listener-output timestamps. Compare systems on the same transport first; browser, SIP and WhatsApp results must be separate conditions.

Start with a quiet baseline. Any added noise condition must specify the source, signal-to-noise ratio and mixing method. Network effects must be measured or declared unknown. Model-native duplex capability and transport-level cancellation are separate properties.

5. Measurements and scoring

Yield latency: time from the first audible sample of the intended interruption at the captured input to the end of the system utterance being interrupted at listener output. A gap counts as an end only when at least 300 milliseconds of silence follow it. If the system resumes stale content within two seconds, mark a resumed-output failure. These windows are protocol choices, not validated universal thresholds.

Successful yield: a requested interruption causes the current utterance to end within a predefined two-second observation window without a resumed-output failure. Report the numerator and denominator. Runs without a successful yield are failures/right-censored latency observations at the window boundary; never convert them into two-second successful responses or omit them from success statistics.

False-yield rate: fraction of backchannel or unrelated-speech trials in which the system unnecessarily abandons its turn. Human annotation must distinguish a conversational pause from abandonment.

Premature takeover: fraction of hesitation trials in which the system starts a substantive response before the speaker completes the scripted turn. A brief appropriate backchannel is annotated separately.

Recovery correctness: fraction of correction trials in which the next substantive response uses the corrected detail. Score semantic appropriateness independently of stop latency.

Publish per-trial outcomes, counts, median and p95 latency among successful yields, and the number of censored observations. With a small pilot, p95 is descriptive and unstable. Report language and scenario breakdowns; do not pool repeated runs into a claim about all speakers. Two annotators review ambiguous trials and document disagreements and adjudication.

6. Evidence and data handling

The trial manifest must include trial ID, system version, language variety, scenario, speaker pseudonym, intended action, prompt hash, seed, onset times, outcome, failure reason and recording hash. Store scripts, annotation guidance and analysis code with a version tag. Keep recordings only from consenting participants; remove customer identifiers, credentials and private operational details before any release. Publication of recordings requires separate permission from participants.

Release a redacted per-trial manifest and aggregate analysis when permissions permit. If audio cannot be released, describe the resulting reproducibility limit explicitly. Do not publish private caller recordings as research examples by default.

7. Status, limitations and next steps

Version 0.1 specifies a proposed method. No test recordings were analysed for this document, no participants were recruited for this pilot and no numerical outcomes are available. Existing deployment-journal narratives are context, not a benchmark dataset.

The first milestone is native-language review of scripts, then a calibration of audio capture and annotation on a small non-scored run. Freeze the protocol and manifest before scored collection. Record deviations in the subsequent report. A larger study must include more speakers and a sample-size justification.

Current limitations include small planned trial counts, subjective pragmatic labels, sensitivity to transport and buffering, unvalidated timing windows and limited speaker diversity. Speech fluency, dialect quality and educational benefit require separate evaluations.

8. Collaboration and authorship

Contact valerii@goippro.com for script review, annotation and reproducible system evaluations. Contributions and final authorship will be reviewed for later versions. This protocol was prepared with AI assistance for drafting and technical editing; its proposed methods and future empirical claims require human review. No relationship with Kyutai, Block or the benchmark authors is implied.

References

1. Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave and Neil Zeghidour. 2024. *Moshi: a speech-text foundation model for real-time dialogue.* arXiv:2410.00037v2. https://arxiv.org/abs/2410.00037.

2. Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu and Hung-yi Lee. 2025. *Full-Duplex-Bench: A Benchmark to Evaluate Full-duplex Spoken Dialogue Models on Turn-taking Capabilities*. arXiv:2503.04721v3. https://arxiv.org/abs/2503.04721.