

Prosody-Synchronous Trust-Preserving Translation (PSTT): A Contract for Flow-First Conversational Mediation

Anonymous Author(s)

January 7, 2026

Abstract

We propose Prosody-Synchronous Trust-Preserving Translation (PSTT), a *flow-first* speech-to-speech mediation framework for informal multilingual conversation. Unlike conventional translation systems optimized primarily for semantic fidelity, PSTT treats conversation as a composite of (i) propositional content, (ii) prosodic timing and turn-taking structure, and (iii) paralinguistic signals such as laughter, fillers, and hesitation. We formalize a *conversational mediation contract*: a single governing principle, a failure taxonomy that distinguishes human-like noise from trust-breaking errors, and guardrails that prohibit predictive commitments in high-risk semantic regions (negation, quantities, names, commitments). We further provide a feasibility map that decomposes the contract into known components (streaming ASR, incremental/simultaneous MT, prosody-conditioned synthesis, paralinguistic detection, confidence estimation) and outline evaluation protocols (including Wizard-of-Oz studies) focused on *trust* and *prosodic synchrony* rather than verbatim adequacy alone.

1 Introduction

Speech translation has rapidly advanced from cascaded pipelines to end-to-end models, and from offline translation to low-latency *simultaneous* settings [Sperber and Paulik, 2020, Ma et al., 2019, Ren et al., 2020]. Yet everyday conversation is not merely a sequence of semantic propositions. Humans manage turn-taking with fine temporal precision [Sacks et al., 1974], use disfluencies and self-repair as normal interactional tools [Shriberg, 2001, Schegloff et al., 1977], and rely on paralinguistic cues (laughter, fillers, hesitation) to regulate rapport and intent [Truong and van Leeuwen, 2007, Kaushik et al., 2015].

This paper frames a distinct target: *trusted, informal, low-stakes multilingual conversation* where maintaining perceived human presence and conversational rhythm is often more important than verbatim completeness. In this regime, the most damaging failures are not minor paraphrases, but *trust-breaking* errors: polarity flips (negation), fabricated specifics, false commitments, and speaker/intent misattribution.

Contributions. We contribute:

- A **conversational mediation contract** for flow-first real-time speech-to-speech mediation, including a signal model, a governing objective, and a failure taxonomy (acceptable vs. forbidden failures).
- A set of **predictive guardrails** specifying when early generation is permitted and when it must be delayed, hedged, or de-specified.

- A **feasibility map** connecting contract requirements to known technical methods in streaming ASR/MT/TTS, paralinguistic detection, and confidence estimation [Graves, 2012, Raffel et al., 2017, Ma et al., 2019, Li et al., 2021].
- An **evaluation framework** emphasizing prosodic synchrony and trust, including Wizard-of-Oz protocols [Dahlbäck et al., 1993, Kelley, 1984].

2 Problem Setting and Scope

2.1 What PSTTis (and is not)

PSTTis not a general-purpose translator. It is a *personality-preserving conversational mediator* for **trusted**, **informal** interactions where participants can tolerate small drift but not trust violations. We explicitly exclude high-stakes domains (legal, medical, financial, contractual) where even “human-like” uncertainty is unacceptable.

2.2 Interaction and channel model

We assume a dyadic (or small-group) setting where each participant speaks naturally in their own language, and receives mediated speech in that same language from the other side(s). In typical usage, each participant need not hear the target-language audio produced for their interlocutor; the system is therefore free to optimize the outgoing channel for the listener while keeping the speaker experience natural (e.g., via local audio routing or earbud playback). Any “traceability” UI (e.g., confidence indicators, flags) should be off-channel and optional, to avoid breaking conversational immersion.

2.3 Explicit non-goals

PSTTdoes *not* guarantee verbatim translation, legal or factual precision, or suitability for official contexts. Conversational self-repair and mild vagueness are expected behaviors under the contract, not failures.

2.4 Operating constraints

Real time means prosodic synchrony. We treat “real time” as maintaining turn-taking rhythm and overlap behavior, not merely minimizing end-to-end latency. The system should enter the channel at turn start, avoid unnatural silence, and preserve the timing of transition-relevance places (TRPs) [Sacks et al., 1974].

Selective slowdown is allowed. Delays are acceptable *locally* in high-risk semantic zones (negation, numbers, names, commitments), provided the interaction remains natural (e.g., through hedging, fillers, or partial framing).

3 Related Work

3.1 Simultaneous and incremental translation

Simultaneous MT formalizes the quality–latency trade-off and motivates policies that interleave reading and writing [Gu et al., 2017, Ma et al., 2019]. Streaming speech translation extends these

ideas to speech, often using segmentation and wait- k -style constraints [Ren et al., 2020]. Monotonic attention provides a differentiable mechanism for online alignment [Raffel et al., 2017].

3.2 Speech-to-speech translation

Direct speech-to-speech translation has been demonstrated in end-to-end settings [Jia et al., 2019, 2021]. These systems highlight feasibility, but are typically evaluated with text-centric metrics (e.g., BLEU) and often do not explicitly optimize for conversational trust and turn-taking naturalness.

3.3 Conversation structure: turn-taking and repair

Turn-taking is a core organizational principle of conversation [Sacks et al., 1974]. Self-repair is not an error mode to eliminate but a structured interactional mechanism [Schegloff et al., 1977]; disfluencies can be informative and context-sensitive [Shriberg, 2001]. Modern dialogue systems model turn timing via continuous predictors and transformer-based models [Skantze, 2017, Ekstedt and Skantze, 2020].

3.4 Paralinguistic events and confidence

Automatic discrimination of laughter from speech is well studied [Truong and van Leeuwen, 2007], and joint laughter/filler detection has been explored in naturalistic audio [Kaushik et al., 2015]. For guardrails, confidence estimation in ASR is crucial to prevent premature commitments [Li et al., 2021].

3.5 Interpreting theory and Wizard-of-Oz prototyping

Conference interpreting research emphasizes cognitive load and trade-offs between accuracy and timing [Gile, 2009]. Wizard-of-Oz methods are a pragmatic way to study interactional demands before full automation [Dahlbäck et al., 1993, Kelley, 1984].

4 A Conversational Mediation Contract

We formalize PSTT as a contract that constrains behavior in real time. The contract is intended to be implementable and falsifiable: every observed failure must be classifiable under the taxonomy below.

4.1 Governing principle

When forced to choose, PSTT must favor conversational flow and perceived human presence over semantic completeness, provided no hard semantic violations occur.

This principle shifts optimization from verbatim adequacy to *safe mediation*: preserve rhythm and affect, but never fabricate commitments or polarity.

4.2 Signal model

We treat live speech as a layered signal:

1. **Linguistic content:** propositional meaning, requests, claims.

Class	Description / examples
A (acceptable)	Natural conversational noise: mild paraphrase drift, vagueness, self-repair markers, brief rephrasing.
B (tolerable)	Managed disruption: local slowdown in detail-heavy segments, hedging, partial framing that recovers naturally.
C (forbidden)	Trust-breaking errors: negation/polarity flips, fabricated specifics, false commitments, identity or speaker misattribution, emotional polarity inversion in sensitive contexts.

Table 1: Failure taxonomy for PSTT. The contract permits A, minimizes B, and forbids C.

2. **Prosodic structure:** timing, rhythm, intonation, overlap.
3. **Paralinguistics:** laughter, sighs, fillers, hesitations, backchannels.
4. **Meta-conversational events:** self-repair, rephrasing, clarification.

Translation rule. Only linguistic content is translated; prosody and paralinguistics are preserved or mirrored; meta-conversational events are rendered naturally in the target language (not explained).

4.3 Prosodic lock-in

“Real time” is operationalized as *prosodic synchrony*: the mediated voice should begin promptly, track rhythm, and respect TRPs. The system is allowed to use non-lexical vocalizations (e.g., target-language-appropriate fillers or backchannels) to maintain presence when semantic content is unstable.

4.4 Failure taxonomy

We define three failure classes (Table 1).

4.5 Prediction contract (guardrails)

Simultaneous systems often benefit from prediction/anticipation [Ma et al., 2019]. PSTT permits prediction only when error impact is low.

Never predict (hard guardrails).

- Negation and polarity markers (e.g., “not”, “never”)
- Numbers, quantities, dates, times
- Names and proper nouns (entities)
- Commitments/approvals (“yes”, “no”, “I agree”, “I will”)
- Sensitive-domain content (health, money, legal commitments)

If these are unstable, the system must hold, hedge, or ask for clarification. Guardrails can be implemented via token- and segment-level confidence, NER, and risk classifiers [Lample et al., 2016, Li et al., 2021, Morante and Sporleder, 2012].

Conditionally predict. Discourse markers, politeness strategies, and clause framing can often be produced in a way that remains compatible with multiple future completions.

Freely predict. Prosodic timing, backchannels, and non-lexical “presence” cues can be produced early, since they rarely create Class C failures.

4.6 Translation gating for non-verbals

When laughter, dominant fillers, or hesitation are detected, translation output should be gated: no propositional content is emitted, but paralinguistic mirroring and timing continuity continue [Truong and van Leeuwen, 2007, Kaushik et al., 2015]. This prevents “over-translating humanity” while preserving affect and conversational flow.

4.7 Context-sensitive conservatism

PSTT adapts conservatism based on conversational context:

- **Flow zones** (small talk, storytelling, low-specificity opinions): aggressive flow preservation, optimistic mediation.
- **Detail zones** (names, numbers, logistics): selective slowdown and near-zero guessing.
- **Sensitive zones** (health, money, family conflict, politics): expanded guardrails and tone neutrality.

This can be implemented as a risk-aware decoding policy conditioned on topic and confidence signals.

4.8 Repair and recovery semantics

Repair. Repairs must be additive or reframing, consistent with the conversational preference for self-correction [Schegloff et al., 1977]. No apologies or system explanations occur in the mediated speech channel.

Recovery. After a rare hard correction, the system should revert to the last stable semantic state and re-enter with a clean utterance, prioritizing re-immersion over transparency.

5 Feasibility Map

The contract is only useful if each requirement maps to at least one plausible implementation technique. Table 2 summarizes a non-exhaustive mapping; PSTT’s novelty lies primarily in *coordinating* these components under the failure taxonomy, rather than inventing new primitives.

6 Architecture Sketch

We outline a minimal architecture consistent with the contract (Figure 1).

Contract requirement	Plausible methods (examples)
Prosodic lock-in (early turn entry, TRP timing)	Streaming ASR with partial hypotheses [Graves, 2012]; turn-taking predictors [Skantze, 2017, Ekstedt and Skantze, 2020]; low-latency synthesis.
Signal separation (speech vs laughter/fillers)	Paralinguistic classifiers and laughter/filler detection [Truong and van Leeuwen, 2007, Kaushik et al., 2015].
Never-predict guardrails (negation, entities, commitments)	Confidence estimation [Li et al., 2021]; NER [Lample et al., 2016]; negation/modality modeling [Morante and Sporleder, 2012].
Local holding / hedging while speaking	SimulMT policies and incremental decoding [Gu et al., 2017, Ma et al., 2019].
Human-like repair (additive, no system meta-talk)	Conversation-analytic repair structure [Schegloff et al., 1977]; incremental regeneration of recent clauses.

Table 2: Feasibility map: each PSTRequirement maps to known methods.

Streaming ASR → Incremental MT → Commit / Guardrail Manager → Prosody & Paralinguistic Planner → Low-latency TTS Paralinguistic / VAD classifiers → gating signals

Figure 1: A contract-aligned architecture sketch. Each block can be implemented with existing techniques; novelty lies in the policy logic that enforces the failure taxonomy.

6.1 Streaming ASR

Streaming ASR provides partial hypotheses with token-level uncertainty. RNN-Transducers [Graves, 2012] and related streaming architectures enable low-latency decoding, which is a prerequisite for prosodic lock-in.

6.2 Incremental MT / SimulMT

Incremental MT can be implemented via fixed-latency policies such as wait- k [Ma et al., 2019] or adaptive read/write policies [Gu et al., 2017]. Online alignment mechanisms, including monotonic attention, can reduce recomputation and support incremental commitment [Raffel et al., 2017].

6.3 Commit manager and guardrails

The commit manager decides which target tokens are safe to emit now versus which must be delayed, hedged, or de-specified. Inputs include:

- token/segment confidence (ASR and MT),
- entity and number detection,
- negation risk markers,
- domain/sensitivity classifiers,
- dialogue state (last stable semantic state).

6.4 Prosody and paralinguistic planner

Prosody and paralinguistics are first-class outputs. When propositional content is gated, the system can still produce appropriate presence cues (e.g., acknowledgments, hesitation markers) to preserve synchrony without risking Class C failures.

7 Evaluation Framework

7.1 What to measure

We argue for evaluation targets beyond adequacy/fluency:

- **Class C rate:** frequency of forbidden failures (polarity flips, false commitments, fabricated specifics, misattribution).
- **Prosodic synchrony:** deviation in onset timing and TRP alignment; overlap/gap statistics relative to human baselines [Sacks et al., 1974].
- **Conversational naturalness:** subjective ratings of “human presence” and “flow”.
- **Recovery quality:** how quickly interaction returns to smooth flow after disruptions.

7.2 Protocols

Wizard-of-Oz (WoZ). WoZ studies can validate the contract and collect interaction data before full automation [Dahlbäck et al., 1993, Kelley, 1984]. A human mediator can follow the contract rules (including guardrails) to establish human tolerance thresholds for hedging, delays, and repairs.

Incremental technical MVP. After WoZ, a narrow technical MVP (single language pair, limited domains) can test whether automatic components can satisfy the Class C constraints under realistic latency.

8 Limitations and Ethical Considerations

The contract explicitly excludes high-stakes domains. Even in low-stakes contexts, voice preservation in direct speech-to-speech translation introduces misuse risks, and prior work has proposed mitigations [Jia et al., 2021]. We emphasize that the goal is trusted informal communication, not authoritative translation.

9 Conclusion

We presented PSTT, a contract-driven framing for flow-first conversational mediation. By separating acceptable conversational noise from forbidden trust-breaking failures and by formalizing guardrails for prediction, PSTT offers a practical target for building real-time multilingual mediation that feels human without pretending to be perfect.

References

- Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of Oz studies: Why and how. *Knowledge-Based Systems*, 6(4):258–266, 1993.
- Erik Ekstedt and Gabriel Skantze. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.268. URL <https://aclanthology.org/2020.findings-emnlp.268>.
- Daniel Gile. *Basic Concepts and Models for Interpreter and Translator Training*. John Benjamins Publishing Company, Amsterdam, revised edition, 2009.
- Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012. URL <https://arxiv.org/abs/1211.3711>.
- Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062. Association for Computational Linguistics, 2017. URL <https://aclanthology.org/E17-1099/>.
- Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. In *Proc. Interspeech 2019*, 2019. URL https://www.isca-archive.org/interspeech_2019/jia19_interspeech.html.
- Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. *arXiv preprint arXiv:2107.08661*, 2021. URL <https://arxiv.org/abs/2107.08661>.
- Lakshminish Kaushik, Abhijeet Sangwan, and John H. L. Hansen. Laughter and filler detection in naturalistic audio. In *INTERSPEECH 2015*, pages 2509–2513, 2015. URL https://www.isca-archive.org/interspeech_2015/kaushik15_interspeech.html.
- John F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Office Information Systems*, 2(1):26–41, 1984.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1030. URL <https://aclanthology.org/N16-1030/>.
- Qiuji Li, David Qiu, Yu Zhang, Bo Li, Yanzhang He, Philip C. Woodland, Liangliang Cao, and Trevor Strohman. Confidence estimation for attention-based sequence-to-sequence models for speech recognition. In *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6388–6392. IEEE, 2021.
- Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1289. URL <https://aclanthology.org/P19-1289>.

Roser Morante and Caroline Sporleder. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260, 2012. doi: 10.1162/COLI_a_00095. URL <https://aclanthology.org/J12-2001/>.

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. Online and linear-time attention by enforcing monotonic alignments. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2837–2846. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/raffel17a.html>.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. SimulSpeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.350. URL <https://aclanthology.org/2020.acl-main.350>.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. URL <http://www.jstor.org/stable/412243>.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.

Elizabeth Shriberg. To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169, 2001. doi: 10.1017/S0025100301001128.

Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5527. URL <https://aclanthology.org/W17-5527/>.

Matthias Sperber and Matthias Paulik. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.661. URL <https://aclanthology.org/2020.acl-main.661/>.

Khiet P. Truong and David A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007. doi: 10.1016/j.specom.2007.01.001.