Contraction Control in Prosody-Driven Human-Robot Interaction

River Adkins
2.165 Term Paper
Massachusetts Institute of Technology

December 10, 2025

Abstract

Spoken communication introduces variability into human-robot interaction, as prosodic cues such as hesitancy can cause large language models (LLMs) to produce inconsistent interpretations of a user's intent. This paper investigates how contraction theory can provide stability guarantees for robots operating under prosody-driven command. Using a hesitancy signal extracted from real speech with a pretrained WavLM-based classifier, we construct reference trajectories whose lateral component oscillates in proportion to the user's uncertainty. We then compare a contraction-based tracking controller to a proportional baseline on a planar robot model. The contracting controller remains stable and accurately tracks the fluctuating reference, while the baseline exhibits lag and error spikes during high-hesitancy intervals. These findings show that contraction theory offers a principled means of stabilizing robot behavior when high-level commands derived from speech and LLM parsing are noisy or inconsistent, and they offer a new foundation for incorporating prosody-aware intent estimation into future human–robot interaction systems.

1 Introduction

My research this semester in the Laboratory for Information and Decision Systems (LIDS) under Sertac Karaman and Ifueko Igbinedion focused on developing prosody-aware pipelines for speech-based robot control. The broader aim is to better understand how prosody, specifically hesitancy, can enhance human-robot collaboration.

Modern speech-driven robotic systems increasingly rely on large language models (LLMs) to translate spoken input into high-level commands. However, speech-derived signals are often inconsistent or emotionally variable, for users may hesitate, restart instructions, express uncertainty, or change their mind mid-utterance. While currently, most systems rely almost exclusively on automatic speech recognizer (ASR) transcripts for command inputs, there is increasing evidence that prosody should be a factor to enhance human-robot collaboration. However, even in the current standard system, there are still inconsistencies from phrasing differences and ASR errors. Because LLMs are vulnerable to hallucinations and model-dependent interpretation differences, their outputs naturally vary from moment to moment. Prosodic hesitancy in speech as well as ambiguous textual phrasing can magnify these effects by producing uncertain inputs. This interaction leads to inconsistent or rapidly shifting commands that pose significant challenges for control.

Robots operating under such conditions require controllers that are robust to inconsistent, time-varying references. Traditional stability analysis tools, such as Lyapunov methods, linearization, and frequency-domain techniques, typically assume smooth, slowly varying command inputs. These assumptions break down in speech-driven control, where natural prosody and LLM interpretation

produce inconsistent commands. Thus exposing the need for contraction theory, which examines incremental stability by ensuring that all system trajectories converge toward each other regardless of initial conditions or disturbances.

In this paper, we show how contraction theory can provide formal stability guarantees for robots responding to prosody-driven and LLM-mediated commands. We model the speech-to-action pipeline as a hierarchical dynamical system, in which human prosody influences the structure and variability of the reference trajectory provided to the robot. We then design a contracting controller and analyze how it ensures stable robot behavior even when the high-level command signal is noisy. To validate our approach, we conduct simulation experiments in which reference trajectories are intentionally made inconsistent to mimic patterns commonly observed in natural speech. We compare contracting control to a non-contracting baseline and demonstrate that contraction yields smoother tracking, reduced sensitivity to fluctuations, and predictable performance under prosody-like variability.

By integrating prosody-aware modeling with nonlinear stability tools, this work aims to advance the foundations of speech-driven HRI and contribute a mathematically grounded framework for stable robot behavior under uncertain, emotionally expressive human communication.

2 Background and Related Work

2.1 Prosody as a Signal of Cognitive State

Prosody, the rhythm, pitch, timing, and energy patterns of speech, plays a crucial role in conveying information about a speaker's cognitive and emotional state. Decades of psycholinguistic research show that filled pauses, elongated syllables, disfluencies, and pitch modulations reliably correlate with retrieval difficulty, uncertainty, and increased cognitive load [1, 2, 3]. For example, Clark and Fox Tree demonstrated that speakers use "uh" and "um" systematically to signal delays in planning or problems in recalling information [2]. These findings establish hesitancy as a natural, interpretable cue that emerges when speakers experience uncertainty or incomplete knowledge.

Within HRI, prosodic cues influence user perceptions of robot intention, confidence, and reliability. A robot's voice quality and prosodic expressiveness have been shown to modulate trust and engagement [4]. Additionally, robot hesitation behaviors themselves can improve human trust calibration, suggesting that agents, biological or artificial, leverage temporal cues to manage expectations in joint tasks [5]. However, despite increasing interest in prosody-driven adaptation, relatively little work examines how robots should respond when humans exhibit hesitancy during natural communication. This gap motivates the development of systems that can interpret prosodic variability and adjust behavior accordingly.

2.2 Prosody in Human-Robot Collaboration

Beyond conveying a speaker's cognitive state, prosody plays a central role in shaping how humans express their expectations during joint tasks. Research in human–human communication shows that prosodic cues such as timing, pitch modulation, and hesitancy guide turn-taking and disambiguate intent [6, 7]. These cues promote smoother coordination by enabling listeners to anticipate upcoming actions and better understand their partner.

Similar benefits have been observed in human–robot interaction. Robots that generate expressive prosody, such as slowing down, pausing, or modulating pitch, are perceived as easier to collaborate with [8, 9]. Prosody improves a robot's ability to signal confidence, intention, and internal state [10], helping users form accurate mental models of the robot's behavior. Conversely,

robots that ignore prosodic information in human speech often misinterpret instructive nuances, leading to brittle interaction and reduced trust [11].

A growing body of HRI work argues that incorporating prosodic cues into robot perception could substantially improve collaborative fluency. Hesitancy, for instance, has been linked to increased user uncertainty, hesitation in decision making, and the need for confirmation or clarification [12]. Detecting such cues allows a robot to adjust its assistance level, slow or pause its motion, or request further instruction when the human appears unsure. These behaviors have been shown to improve joint task performance, user satisfaction, and safety in shared autonomy settings [13, 14].

Despite these findings, the integration of prosody into robot control pipelines remains limited. Most speech-driven systems rely solely on transcript-level ASR, leaving prosodic cues unused and forcing robots to infer intent from text alone. This gap motivates the development of models and control strategies that can make principled use of prosody, particularly hesitancy, to improve robustness, interpretability, and responsiveness in human–robot collaboration. The present work contributes to this emerging direction by examining how prosody-induced variability affects downstream control and how contraction-based methods can stabilize robot behavior under such fluctuations.

2.3 Speech-Driven Command Generation and LLM-Based Interpretation

Modern human-robot systems increasingly rely on large language models (LLMs) to translate spoken instructions into high-level action plans. LLMs enable remarkable flexibility in interpreting natural language [15], yet they also introduce a vulnerability. Small differences in phrasing can yield different semantic parses or action sequences [16, 17].

In many existing HRI and spoken-dialogue systems, speech-driven commands are treated as discrete symbolic inputs rather than continuous, time-varying signals [18]. However, when prosody is incorporated into the interpretation pipeline, real-time LLM parsing produces a continuously evolving command trajectory whose fluctuations mirror uncertainty or revisions in the speaker's intent [19, 20]. This creates a new class of control challenges. Robots must track reference signals shaped not only by linguistic content but also by the speaker's cognitive and emotional state, both of which can fluctuate unpredictably.

2.4 Contraction Theory for Robust, Hierarchical Robot Control

Contraction theory provides a mathematical foundation for understanding stability in nonlinear, time-varying systems. Unlike classical Lyapunov analysis, which typically studies convergence to a fixed equilibrium point, contraction theory establishes incremental stability. This means the system will forget disturbances or fluctuations as it evolves, making it naturally robust to noise, switching inputs, or inconsistent references.

These properties make contraction theory particularly well-suited for hierarchical control architectures, where low-level controllers must remain stable despite noisy or uncertain high-level commands. When applied to robot tracking problems, contraction analysis yields controllers that maintain smooth, predictable motion even when the target trajectory exhibits discontinuities or rapid variations. This robustness is especially valuable in speech-driven HRI, where human hesitancy and LLM variability can cause rapid changes in the inferred reference trajectory.

2.5 Gap in Existing Work

While prior research has examined prosody as an indicator of uncertainty and has explored trust enhancement through robot hesitation, there is limited work on using human hesitancy as an

input to robot control systems. Furthermore, existing prosody-driven HRI approaches generally focus on interaction and perception rather than providing formal stability guarantees for robot motion. Similarly, although contraction theory has been applied to robot tracking, coordination, and adaptive control, it has not yet been studied in the context of speech-derived, emotionally modulated command signals.

This combination of prosody-driven reference trajectories interpreted through LLMs and contraction-based control remains largely unexplored. The work in this paper builds on established foundations in psycholinguistics, HRI, and nonlinear control to address this gap and develop a stability-centered framework for prosody-influenced human–robot collaboration.

3 System Model

3.1 Overall Speech-to-Action Pipeline

In a complete prosody-driven human–robot collaboration system, the pipeline would begin by converting a user's spoken command into both linguistic content and prosodic information. An ASR would provide the transcript, while a prosody model would extract acoustic features. A large language model (LLM) would then take both streams as input and produce a high-level plan expressed in terms of defined robot actions. This full stack forms a hierarchical perception–reasoning–control pipeline in which prosody can influence not only what command the LLM produces, but also how confidently or consistently that command is expressed over time.

Rather than modeling the entire ASR–prosody–LLM stack, we adopt an abstraction suitable for control analysis. In our research environment, a voice-quality classifier processes raw audio and outputs estimates of attributes related to hesitancy. These outputs are converted into a scalar hesitancy signal $p(t)$, which captures moment-to-moment variability in the speaker's stability and certainty.

We treat this hesitancy signal as a disturbance acting on an otherwise smooth reference trajectory $r_{\text{base}}(t)$ representing the user's intended command. When the speaker hesitates, the LLM in a full system would likely produce more unstable or inconsistent interpretations of the instruction. In our model, this effect is represented by using $p(t)$ to modulate the trajectory, yielding a perturbed reference $r(t)$. Thus, $r(t)$ stands in for the fluctuating, prosody-influenced high-level commands that arise from hesitant speech.

This abstraction allows us to focus on the central question of this paper: **Can contraction-based control guarantee stable robot behavior even when the high-level commands derived from speech are inconsistent?** By injecting prosody-derived fluctuations into $r(t)$ and evaluating different controllers, we analyze how robust the closed-loop robot system remains when exposed to human prosodic variability in the speech-to-action pipeline.

3.2 Hesitancy Signal Derived From Real Speech Data

We derive a time-varying hesitancy signal $p(t)$ using a pretrained WavLM-based voice-quality classifier,[1] which was trained on the ParaSpeechCaps dataset.[2] This model maps short audio segments to continuous-valued predictions across several vocal-quality dimensions.

To obtain a temporally resolved hesitancy curve, we process the user's recorded spoken instructions using a sliding-window procedure:

[1] `tiantiaf/wavlm-large-voice-quality`, available on HuggingFace.

[2] `ajd12342/paraspeechcaps`, a prosody-rich corpus annotated for voice-quality attributes such as breathiness, roughness, instability, and disfluency.

1. The audio waveform is segmented into overlapping windows of a fixed 3-second duration.
2. Each window is fed into the WavLM classifier, producing a vector of voice-quality scores.
3. We extract the classifier dimensions associated with hesitancy-related traits such as stammering, hushed, singsong, and hesistant.[3]
4. These scores are normalized into a scalar value $p(t_i) \in [0, 1]$, representing the hesitancy level at time t_i.

After extracting the discrete hesitancy scores $p(t_i)$, we filter out small, rapid fluctuations in the classifier output so that the hesitancy curve is smooth enough for control analysis, but still preserves the real shape of how the user's hesitancy changes over time. An example hesitancy trajectory obtained from real speech using this method is shown in Fig. 1.

[3]In the WavLM voice-quality classifier, the label *hesistant* refers specifically to rhythmic irregularity or timing instability in speech. In this paper, we use the term more broadly to refer to cognitive or semantic uncertainty reflected in prosody.

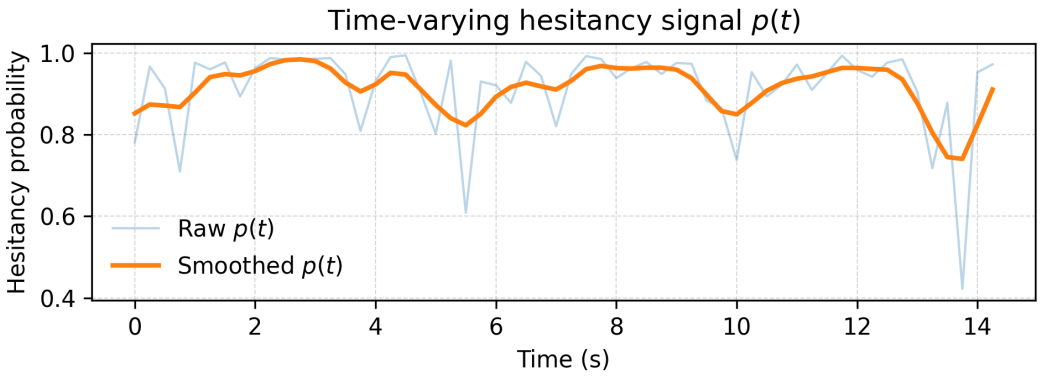

Figure 1: Example hesitancy signal $p(t)$ extracted from spoken command audio using a WavLM-based voice-quality classifier. Higher values indicate greater hesitancy and correspond to moments when the user's instructions were uncertain.

The hesitancy signal plays a central role in our system model. Rather than assuming user commands are smooth or consistent, we explicitly incorporate prosodically induced variability into the reference trajectory that the robot is asked to follow.

3.3 Reference Trajectory Generation

Let $r(t) \in \mathbb{R}^2$ denote the desired robot position inferred from the semantic content of the user's spoken command. If the user speaks fluently, the LLM-derived intent produces a relatively smooth, stable reference trajectory $r_{\text{base}}(t)$. However, when the user hesitates the inferred command becomes less stable. This instability is reflected in the hesitancy signal $p(t)$, which modulates the variability of the target trajectory.

To capture this effect, we model the commanded reference trajectory as

$$r(t) = r_{\text{base}}(t) + \alpha\, f(p(t)),$$

where $\alpha > 0$ is a scaling factor and $f(p(t))$ generates hesitancy-induced perturbations. Intuitively, when $p(t)$ is small, the perturbation is negligible and the commanded trajectory remains smooth; when $p(t)$ is large, the user appears uncertain, and the trajectory fluctuates accordingly.

A simple and effective choice for f is an oscillatory perturbation,

$$f(p(t)) = \begin{bmatrix} 0 \\ p(t)\,\sin(\omega t) \end{bmatrix},$$

where ω controls the frequency of the fluctuations. This formulation reflects the observation that hesitancy often manifests as inconsistent or rapidly changing instructions. The resulting reference trajectory may contain reversals, jitter, or transient oscillations when the user is unsure, and remains smooth when the user is confident.

This model allows us to focus on how hesitancy influences the stability properties of the tracking controller. In the following section, we analyze how a contraction-based controller ensures stable tracking of $r(t)$ despite these hesitancy-induced variations.

3.4 Robot Dynamics

To study the effect of hesitancy-induced variation in the commanded reference trajectory, we adopt a simple planar point-mass model for the robot. Let the robot state

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \in \mathbb{R}^2$$

denote its Cartesian position in the plane, and let the control input

$$u(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} \in \mathbb{R}^2$$

specify the commanded velocity. The dynamics are given by the first-order system

$$\dot{x}(t) = u(t),$$

which provides an analytically tractable model for examining contraction-based tracking of a time-varying reference trajectory.

Although real mobile robots exhibit more advanced dynamics, the point-mass approximation captures the essential structure of planar trajectory tracking while simplifying the stability analysis. In practice, our long-term goal is to deploy this framework on a real mobile robot platform that I designed (Fig. 2). The platform consists of an omnidirectional mecanum-wheel base, onboard NVIDIA compute hardware, a ZED stereo camera, and a multi-DoF manipulator arm. The simplified model used here can be viewed as a lower-dimensional approximation of our robot's base motion. Therefore the contraction-based controller developed in the following sections serves as a foundational analysis step toward future deployment on the full system.

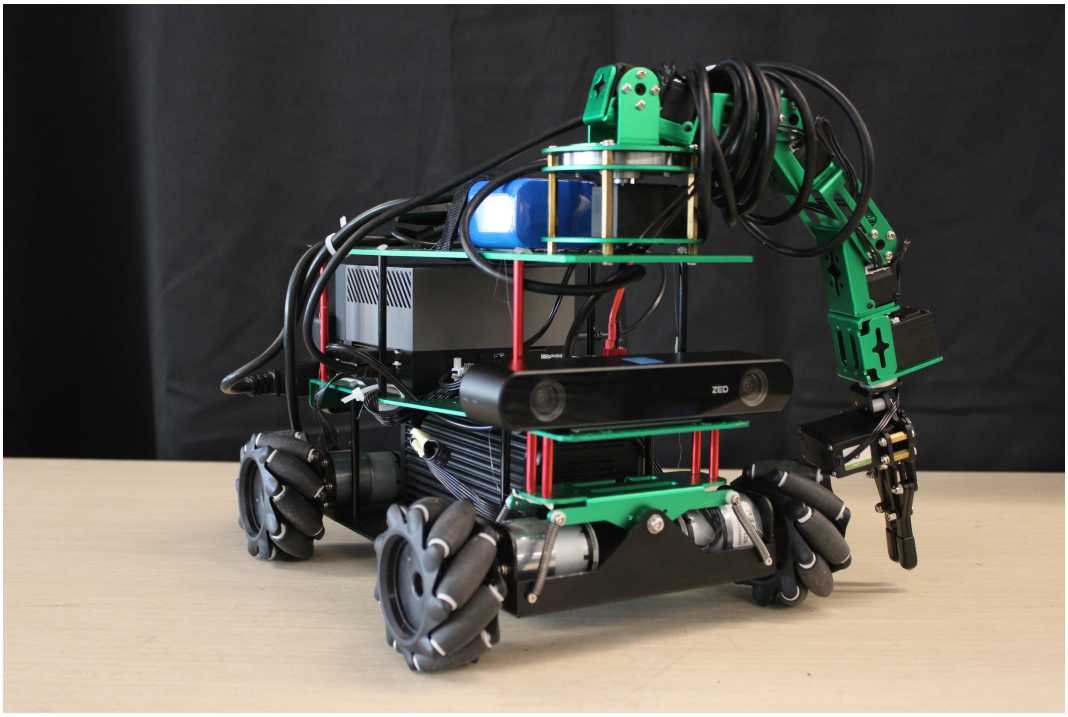

Figure 2: Mobile robot that I built for future deployment of the proposed prosody-aware human-robot interaction framework.

4 Contracting Controller Design

4.1 Tracking Error Dynamics

Our goal is to design a controller that makes the robot track the hesitancy-modulated reference trajectory $r(t)$ in a stable way. We begin by defining the tracking error

$$e(t) = x(t) - r(t),$$

which measures the difference between the robot's position and the desired reference at time t. Differentiating both sides with respect to time gives

$$\dot{e}(t) = \dot{x}(t) - \dot{r}(t).$$

Using the point-mass robot dynamics

$$\dot{x}(t) = u(t),$$

we obtain the error dynamics

$$\dot{e}(t) = u(t) - \dot{r}(t).$$

The tracking behavior is therefore determined entirely by the choice of control input $u(t)$ and the time derivative of the reference trajectory $\dot{r}(t)$. In the next subsection, we specify a control law $u(t)$ that renders the error dynamics contracting, ensuring that $e(t)$ converges exponentially toward zero even when $r(t)$ varies due to hesitancy in the user's speech.

4.2 Contraction-Based Controller Design

Given the error dynamics

$$\dot{e}(t) = u(t) - \dot{r}(t),$$

our objective is to choose a control law $u(t)$ such that the error system is contracting, meaning all trajectories of $e(t)$ converge exponentially toward one another and in particular toward the origin.

Control law. We consider a linear tracking controller with feedforward compensation of the reference velocity,

$$u(t) = \dot{r}(t) - Ke(t),$$

where $K \in \mathbb{R}^{2\times 2}$ is a constant, positive definite gain matrix. Substituting this into the error dynamics yields

$$\dot{e}(t) = \big(\dot{r}(t) - Ke(t)\big) - \dot{r}(t) = -Ke(t).$$

Thus the error evolves according to a linear, time-invariant system with dynamics

$$\dot{e}(t) = -Ke(t).$$

Contraction analysis. We write the error system in the form

$$\dot{e} = f(e) = -Ke,$$

whose Jacobian with respect to e is simply

$$\frac{\partial f}{\partial e} = -K.$$

Choosing the standard Euclidean metric, the symmetric part of the Jacobian is

$$\frac{1}{2}\left(\frac{\partial f}{\partial e} + \frac{\partial f}{\partial e}^{\top}\right) = -\frac{1}{2}(K + K^{\top}).$$

If K is positive definite (e.g., $K = kI$ with $k > 0$), then $K + K^{\top}$ is also positive definite and the symmetric part of the Jacobian is negative definite. Consequently, all eigenvalues of the symmetric Jacobian are strictly negative, and the error system is globally contracting in the Euclidean metric with contraction rate at least $\lambda_{\min}(K)$.

Intuitively, the controller uses $\dot{r}(t)$ as a feedforward term to follow the motion of the reference and applies a proportional feedback term $-Ke(t)$ that pulls the robot state toward the reference. Because the error dynamics are contracting, any two error trajectories converge exponentially toward each other, and in particular $e(t)$ converges exponentially toward zero regardless of the initial condition. This guarantees stable tracking of the hesitancy-modulated reference $r(t)$ even when it varies rapidly in response to hesistations in the user's speech.

4.3 Baseline Controller

To evaluate the benefits of the contraction-based controller, we compare it against a simple baseline that does not guarantee exponential convergence of the tracking error. We choose a pure proportional controller without feedforward compensation,

$$u_{\text{base}}(t) = -K_{\text{b}}\, e(t),$$

where K_{b} is a constant gain matrix. Substituting this control law into the error dynamics,

$$\dot{e}(t) = u_{\mathrm{base}}(t) - \dot{r}(t),$$

yields

$$\dot{e}(t) = -K_{\mathrm{b}} e(t) - \dot{r}(t).$$

Unlike the contraction-based controller, which cancels the reference velocity through a feedforward term, the baseline controller treats rapid changes in $r(t)$ and $\dot{r}(t)$ as disturbances. When the hesitancy signal $p(t)$ is small and $r(t)$ varies smoothly, this controller can track the reference adequately. However, when $p(t)$ grows and the reference trajectory exhibits oscillations or abrupt transitions, the term $-\dot{r}(t)$ acts as a strong external input that the baseline controller cannot reject. As a result, the tracking error may grow or fluctuate significantly rather than converging.

This baseline demonstrates how hesitation-induced variability in the reference trajectory can degrade performance when the controller does not explicitly ensure contraction of the tracking error dynamics.

5 Simulation Setup

To evaluate the effect of prosody-driven hesitancy on robot trajectory tracking, we simulate a simplified speech–to–action pipeline in which prosodic variability disturbs the desired motion of a planar robot. Disturbances are applied only to the lateral (y) component of the reference trajectory. This isolates the effect of hesitancy and allows us to examine how each controller responds to prosody-induced variability without confounding changes in forward motion.

5.1 Reference Trajectory Under Hesitancy Modulation

Given a smoothed hesitancy signal $p(t)$ extracted from real speech (Sec. 3.2), we construct a time-varying reference trajectory

$$r(t) = \begin{bmatrix} r_x(t) \\ r_y(t) \end{bmatrix},$$

where

$$r_x(t) = v_0 t, \qquad r_y(t) = \alpha\, p(t) \sin(\omega t).$$

The forward motion $r_x(t)$ represents a human issuing a consistent high-level instruction such as "move forward," which typically remains stable even when the speaker is uncertain. In contrast, prosodic hesitancy tends to affect how a user refines or corrects a command rather than the overall intent. To model this, the lateral component $r_y(t)$ is scaled by the hesitancy signal $p(t)$. When $p(t)$ is small, the instruction is clear and the desired path is nearly straight. When $p(t)$ increases, an LLM interpreting the speech would receive a more ambiguous or inconsistent prompt, producing small fluctuations in the inferred lateral motion.

By adding hesitancy-dependent oscillations only to $r_y(t)$, the model captures how prosody can introduce variability in the interpreted command, allowing us to isolate its effect on tracking stability without altering the underlying task of moving forward.

5.2 Simulation Parameters

The simulation uses the same time base as the hesitancy signal extracted from speech. The classifier outputs a sequence of hesitancy values $p(t_i)$ sampled at approximately uniform intervals ($\Delta t \approx$

mean$(t_{i+1} - t_i)$). To maintain alignment between the prosody-derived variations and the robot controller, the simulation runs over the exact timestamp sequence provided by the classifier.

The timestep for integration is therefore

$$\Delta t = \text{mean}(t_{i+1} - t_i),$$

and the total simulation duration equals the length of the hesitancy recording.

All robot states and control inputs are updated synchronously with this classifier-derived timeline.

5.3 Evaluation Metrics

Trajectory tracking performance is assessed using:

- The lateral tracking curves $y(t)$ compared against $r_y(t)$, highlighting how each controller responds to hesitancy-induced oscillations.
- The Euclidean tracking error
$$\|e(t)\| = \|x(t) - r(t)\|,$$
which reflects overall deviation from the reference. Since hesitancy affects only the y direction, variations in $\|e(t)\|$ primarily correspond to lateral tracking performance.

These metrics allow us to directly examine how prosody-driven disturbances propagate through the reference trajectory and how effectively each controller rejects them.

6 Results

We compare the tracking performance of the contraction-based controller and a proportional baseline controller under prosody-induced variability in the reference trajectory. The hesitancy signal $p(t)$ extracted from speech determines the amplitude of lateral oscillations in $r_y(t)$, while the forward component $r_x(t)$ remains smooth. This isolates how each controller responds to fluctuations in high-level intent.

6.1 Lateral Tracking Performance

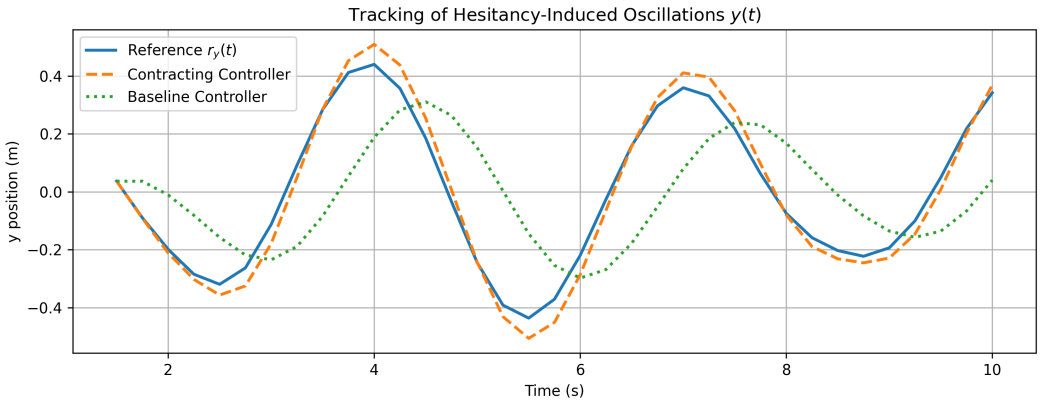

Figure 3: Tracking of hesitancy-induced oscillations in the lateral direction. The contraction-based controller (dashed line) follows the perturbed reference closely, while the baseline controller (dotted line) lags and fails to capture the oscillatory structure produced by prosodic hesitancy.

During periods of low hesitancy, both controllers remain close to the straight-line path. When $p(t)$ increases, the reference oscillates, reflecting uncertainty in the spoken instruction. The contraction-based controller closely follows these oscillations, maintaining small deviations from the reference. In contrast, the baseline controller exhibits noticeable lag and attenuation. This is consistent with the error dynamics because the baseline lacks the $\dot{r}$ feedforward term, so variations in the reference appear as external disturbances that the controller cannot fully reject.

6.2 Tracking Error Over Time

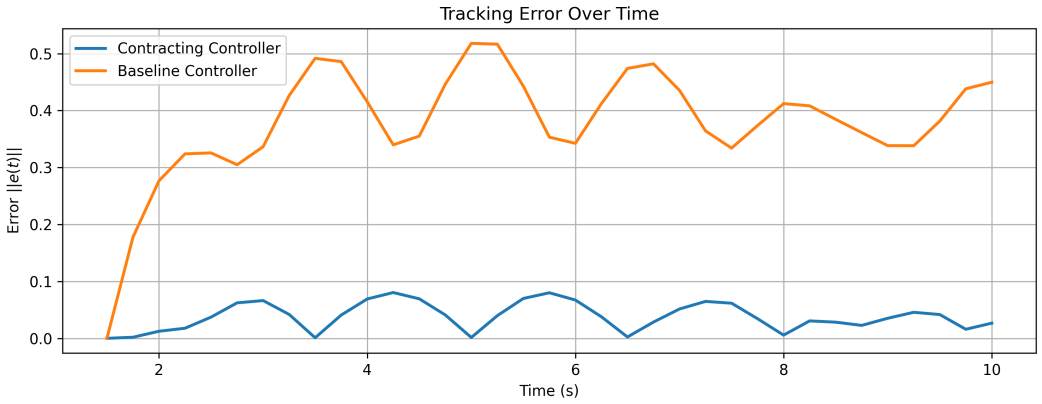

Figure 4: Tracking error over time. The contraction-based controller maintains low error throughout the trajectory, whereas the baseline controller exhibits error spikes during high-hesitancy intervals.

The contraction-based controller converges rapidly and remains stable throughout the trajectory. Even when hesitancy increases, the error while increases remains minimal. The baseline controller fluctuates significantly and peaks during high-hesitancy intervals in which $r_y(t)$ varies most rapidly.

Because the disturbance affects only the lateral component, differences in $\|e(t)\|$ primarily reflect each controller's ability to handle variability in $r_y(t)$. The results demonstrate that the contraction-based controller is significantly more robust to prosody-induced fluctuations in high-level commands.

6.3 Summary of Findings

The contraction-based controller significantly achieves lower tracking error. These results support the claim that contraction theory provides a principled mechanism for stabilizing robot behavior when the high-level command is influenced by speech hesitancy or other forms of human-driven variability.

7 Discussion

7.1 Interpretation of Results

The simulation results highlight the difference between a controller that explicitly incorporates the structure of the reference trajectory and one that treats all variability as external disturbance. Because the contraction-based controller includes the feedforward term $\dot{r}(t)$, it is able to track rapid changes in the lateral reference caused by hesitancy without degrading stability. The baseline controller, by contrast, cannot compensate for high-frequency variations in $r_y(t)$ and therefore exhibits persistent tracking error during hesitant segments. These observations are consistent with the theoretical error dynamics derived earlier and demonstrate the advantage of contraction-based designs when the reference contains structured but unpredictable fluctuations.

7.2 Limitations and Scope of the Abstraction

The disturbance model used in this work deliberately restricts hesitancy-induced variability to the lateral dimension of the reference trajectory. This reflects the intuition that uncertainty in spoken navigation commands typically affects fine-grained steering rather than the overall forward direction. However, real LLM outputs may exhibit more complex forms of variability, including non-sinusoidal fluctuations, abrupt reinterpretations of user intent, or context-dependent corrections. Additionally, the simplified point-mass dynamics abstract away many aspects of physical robot behavior, such as nonholonomic constraints or actuator limitations.

Despite these simplifications, the simulation provides a controlled environment in which the effect of prosody-induced reference variability can be isolated and analyzed. The qualitative conclusion that contraction-based controllers maintain stability and low tracking error under human-driven variability are likely to extend to more complex robot models.

7.3 Implications for Prosody-Aware Robot Control

The simplified simulation presented in this work abstracts hesitancy as a modulation of the lateral reference trajectory. While this captures only a narrow slice of the real speech-to-action pipeline, the resulting behavior reflects a broader challenge facing next-generation human–robot interaction systems. As prosody-aware interfaces develop, the outputs of an LLM interpreting a user's command would become inherently time-varying, reflecting both linguistic content and the speaker's cognitive state.

In such a system, prosody would not directly alter the robot's motion. Instead, a prosody encoder would supply features such as a hesitancy score $p(t)$ to an LLM, which would adjust its interpretation of the instruction accordingly. Hesitant phrasing might lead the LLM to return a conservative or unstable command, while confident phrasing would yield a more decisive trajectory. Combined with the LLM's own sources of variability from wording changes, inconsistent parsing, or momentary hallucinations the system would generate a fluctuating stream of high-level motion targets. Our abstraction captures this phenomenon by treating $p(t)$ as a perturbation of $r_{\text{base}}(t)$, producing a reference $r(t)$ that behaves like the inconsistent output an LLM might generate under uncertain speech.

The results here suggest that contraction-based controllers provide a principled foundation for stabilizing robot behavior in this emerging class of speech-driven, prosody-sensitive systems. Because contraction guarantees exponential convergence of the tracking error for any admissible reference signal, the robot remains stable even when the interpreted intent shifts rapidly or becomes partially unreliable. This robustness stands in contrast to conventional controllers that degrade under rapid command fluctuations. More broadly, these findings point toward a novel general HRI framework in which:

1. Prosody can quantify human uncertainty.
2. An LLM integrates linguistic and prosodic cues to produce a time-varying semantic interpretation.
3. A contraction-based controller stabilizes the robot's response to these fluctuations.

Such an architecture would enable robots to operate safely and predictably in the presence of human hesitation. The present work represents an initial step toward this direction, demonstrating

how prosody-induced variability can be modeled at the control level and highlighting opportunities for deeper integration of prosody, language models, and nonlinear stability tools in future human–robot collaboration.

8 Conclusion

This paper examined how prosodic hesitancy in human speech can influence the behavior of a speech-driven, LLM-mediated robot control pipeline. By treating hesitancy as a source of variability in the high-level reference trajectory, we isolated the effect of prosody-induced uncertainty on the robot's tracking performance. Using a real hesitancy signal derived from spoken audio, we showed that even simple navigation tasks can be disrupted by momentary inconsistencies in user intent as interpreted by a language model.

The proposed contraction-based controller demonstrated strong robustness to these fluctuations, maintaining low tracking error and accurately following the perturbed reference. In contrast, a baseline proportional controller exhibited lag and error spikes during high-hesitancy intervals, reflecting its inability to compensate for rapid variations in the desired trajectory. These results illustrate how contraction theory can stabilize robot behavior in contexts where there is human-driven variability.

Although the simulation uses simplified robot dynamics and a structured model of hesitancy-induced disturbance, the qualitative conclusions extend to more realistic speech-to-action systems. As LLMs become increasingly central to natural-language robot control and prosody becomes increasingly necessary to advance human-robot collaboration, understanding how controllers can absorb human-input variability will be essential. The framework developed here provides a foundation for integrating prosody-aware intent estimation with contraction-based control in future human–robot interaction research.

References

[1] S. E. Brennan and M. Williams, "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers," *Journal of Memory and Language*, 1995.

[2] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, 2002.

[3] V. L. Smith and H. H. Clark, "On the course of answering questions," *Journal of Memory and Language*, 1993.

[4] B. Mutlu, A. Steinfeld, J. Forlizzi, J. Hodgins, and S. Kiesler, "The role of voice in human–robot interaction," *Human–Computer Interaction*, 2012.

[5] M. Traeger, F. Ferland, A. Tapus, *et al.*, "Robot hesitation behaviors improve human trust calibration," in *International Conference on Social Robotics*, 2014.

[6] J. P. De Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, 2006.

[7] R. Levitan, A. Gravano, and J. Hirschberg, "Acoustic–prosodic entrainment and rapport in peer tutoring," *Speech Communication*, 2015.

[8] H. Admoni, A. Dragan, S. S. Srinivasa, and B. Scassellati, "Deliberate delays during robot-to-human handovers improve compliance," in *International Conference on Human–Robot Interaction (HRI)*, 2014.

[9] D. Szafir and B. Mutlu, "Pay attention! designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2015.

[10] C. Stanton, V. Bogdanovic, and M. Kapadia, "Expressive robot motion timing improves transparency and collaboration," *ACM Transactions on Human–Robot Interaction*, 2022.

[11] K. Fischer, "Why it is interesting to investigate how humans and robots talk to each other," *Dialogue and Discourse*, 2016.

[12] S. A. Borrie, "Considerations for measuring speech disfluency in interaction," *Journal of Speech, Language, and Hearing Research*, 2017.

[13] S. Tellex, R. A. Knepper, A. Li, and D. Rus, "Asking for help using inverse semantics," in *Robotics: Science and Systems (RSS)*, 2014.

[14] S. Nikolaidis, D. Hsu, and S. Srinivasa, "Human–robot mutual adaptation in shared autonomy," in *International Conference on Human–Robot Interaction (HRI)*, 2017.

[15] T. Brown *et al.*, "Language models are few-shot learners," in *NeurIPS*, 2020.

[16] B. Schuller and A. Batliner, "Speech emotion recognition: Two decades in a nutshell," *IEEE Transactions on Affective Computing*, 2018.

[17] J. Huang *et al.*, "What makes llms parse instructions inconsistently?," in *ACL Findings*, 2023.

[18] S. Tellex and J. Thomason, "Robots that use language," in *Annual Review of Control, Robotics, and Autonomous Systems*, 2020.

[19] J. Li *et al.*, "Modeling uncertainty and prosody in speech interfaces," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[20] E. Sheng *et al.*, "Llm uncertainty and instability in interactive agents," in *ICLR*, 2023.