SFU SIMON FRASER UNIVERSITY

INTERSPEECH 2025

Generating Consistent Prosodic Patterns from Open-Source TTS Systems

Ha Eun Shim[1*], Olivia Yung[1*], Paige Tuttösí[2], Boey Kwan[1], Angelica Lim[2], Yue Wang[1], H. Henny Yeung[1]

Department of Linguistics[1] and School of Computing Science[2], Simon Fraser University, Canada

Presenters: Ha Eun Shim, Olivia Yung

Department of Linguistics, Simon Fraser University

INTERSPEECH 2025 @ Rotterdam, Netherlands

*Equal Contributions

Simon Fraser University
Language and Brain Lab

SFU LANGUAGE LEARNING AND DEVELOPMENT LAB

ROSIE LAB

PI: Dr. Yue Wang
Boey Kwan
Ivan Fong
Grace Zhang

PI: Dr. H. Henny Yeung
Ha Eun Shim
Olivia Yung
Abigail Agyeiwaa

PI: Dr. Angelica Lim
Dr. Paige Tuttösí

We thank Abigail Agyeiwaa, Ivan Fong, Grace Zhang, and all lab members for their valuable contributions and discussions, as well as the Rajan Family for their support.

We thank the anonymous reviewers for INTERSPEECH 2025.

This work was supported by the Simon Fraser University FASS Breaking Barriers Interdisciplinary Incentive Grant, the Social Sciences and Humanities Research Council of Canada Grant (SSHRC Insight Grant 435–2019–1065), and the NSERC Discovery Grant (RGPIN-2024-06519).

Introduction

Clifton et al. (2006)

Pat or Jay and Lee convinced the bank president to extend the mortgage.

Introduction

Clifton et al. (2006)

Pat or Jay and Lee convinced the bank president to extend the mortgage.

Which meaning?

(1) (Pat) or (Jay and Lee)

(2) (Pat or Jay) and (Lee)

(1) Pat // or Jay and Lee convinced the bank president to extend the mortgage.

(2) Pat or Jay // and Lee convinced the bank president to extend the mortgage.

Introduction

Ambiguous sentences!

One way to resolve the ambiguity:

Prosodic breaks!

Introduction

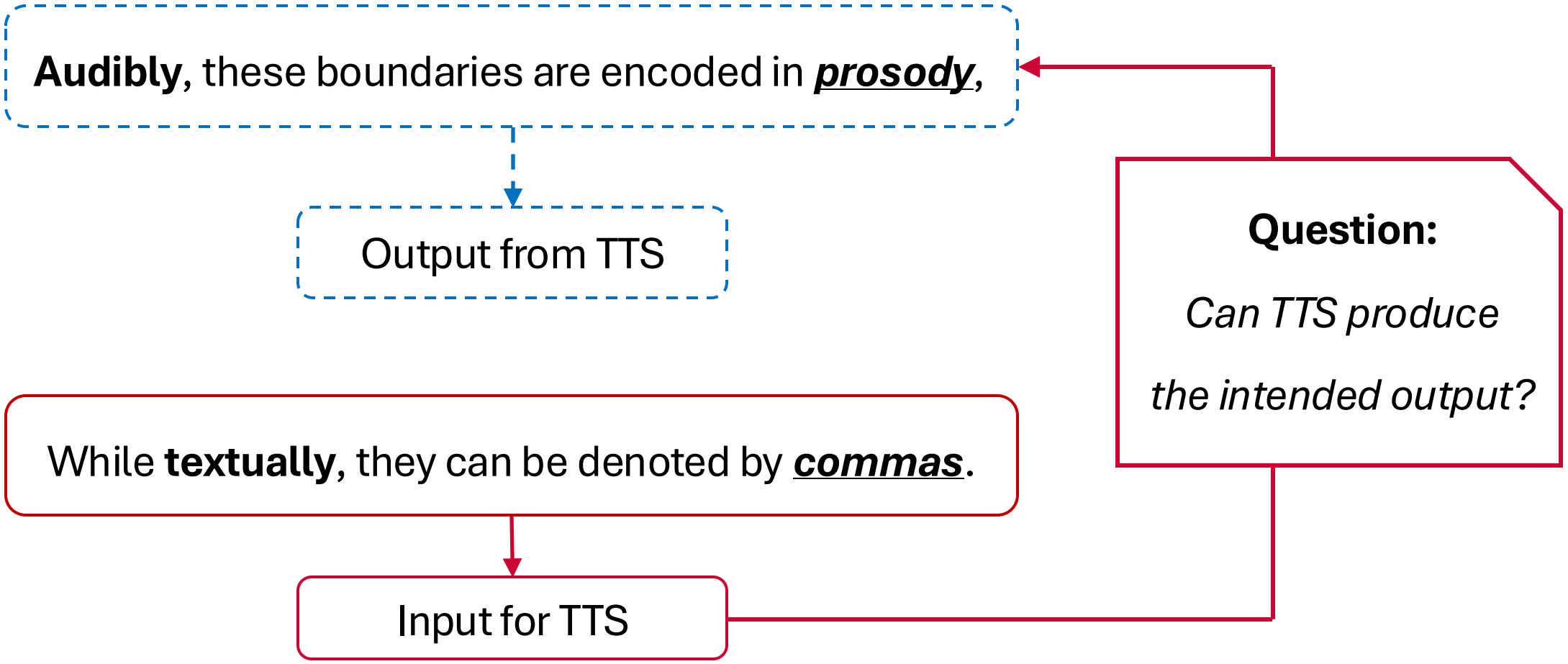
Audibly, these boundaries are encoded in ***prosody***,

while **textually**, they can be denoted by ***commas***.

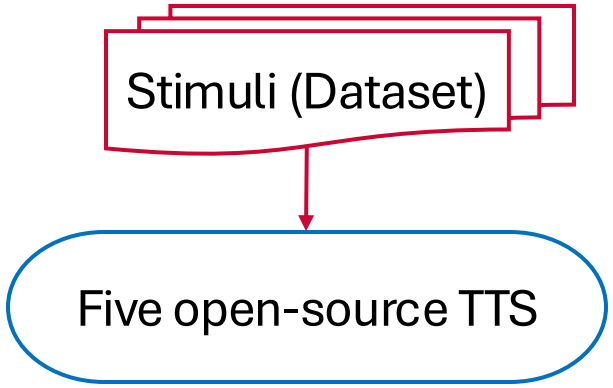
He wants to hit, Gabe

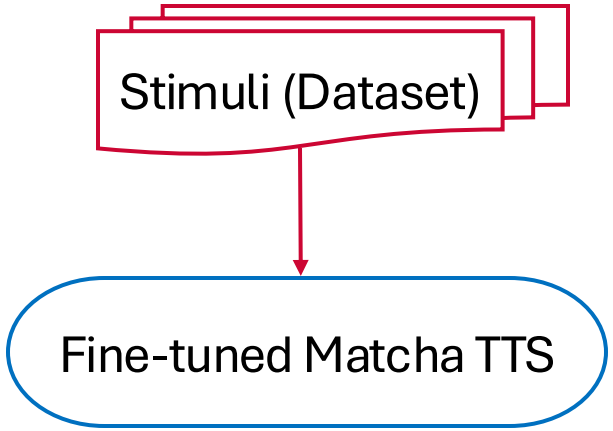
He wants to hit Gabe

Introduction



Introduction





Study 1- Evaluation of TTS systems:

How effectively do open-source TTS systems convey measurable prosodic boundaries in response to comma contrasts?

Study 2- Fine-tuning a TTS system:

How can we enhance TTS output to produce predictable prosodic cues using a customized dataset?

In which textual contexts does this succeed or fail?

Stimuli Examples

Conditions:

1. Condition A (with commas)
2. Condition B (without commas)

Sentence Type I - Direct Address (DA):

Comma signaling the vocative use of a proper noun ***vs.*** Objective case

(DA-A) *He* ***wants*** *to* ***hit, Gabe***

(DA-B) *He* ***wants*** *to hit* ***Gabe***

* Stress pattern: **Bolded** = stressed; *Italics* = unstressed.

Stimuli Examples

Sentence Type II - List:

Commas signaling the listing of multiple people or objects ***vs.*** Restrictive noun phrases

(List-A) *The **mom** drove her **kids**, Eu**gene**, and A**dele***

(List-B) *The **mom** drove her kids Eu**gene** and A**dele***

Commas signaling the listing of multiple people or objects ***vs.*** Accusative cases

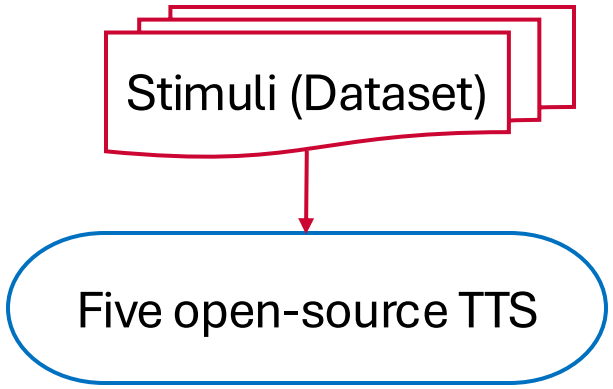
(List-A) *The **man** got a **child**, a **ball**, and a **game***

(List-B) *The **man** got a **child** a **ball** and a **game***

* Stress pattern: **Bolded** = stressed; *Italics* = unstressed.

Study 1: Evaluation of TTS Systems

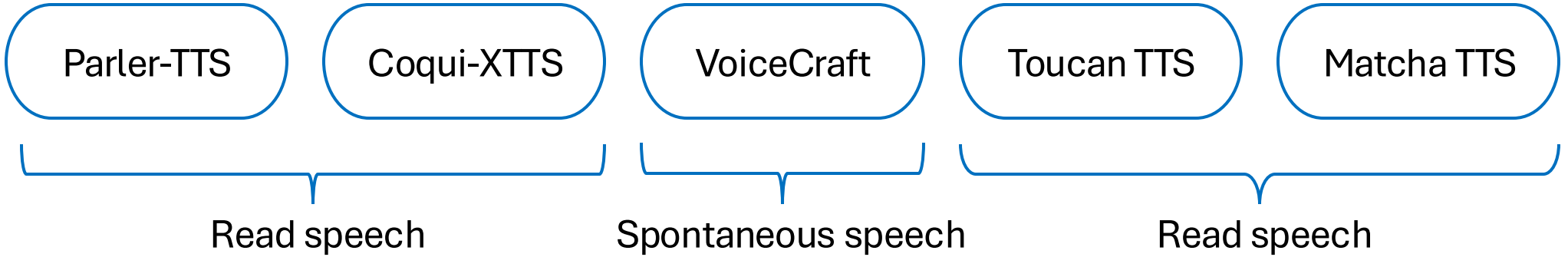
TTS systems



Five Open-source TTS systems

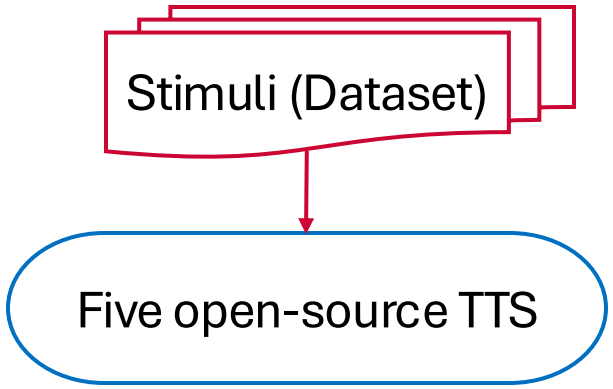
(1) Free** open-source with **online demos

***(2) User modifications** for customizing speech prosody (e.g., gender, speech rate, pitch)*



Study 1: Evaluation of TTS Systems

TTS systems



Five Open-source TTS systems

(1) Free** open-source with **online demos

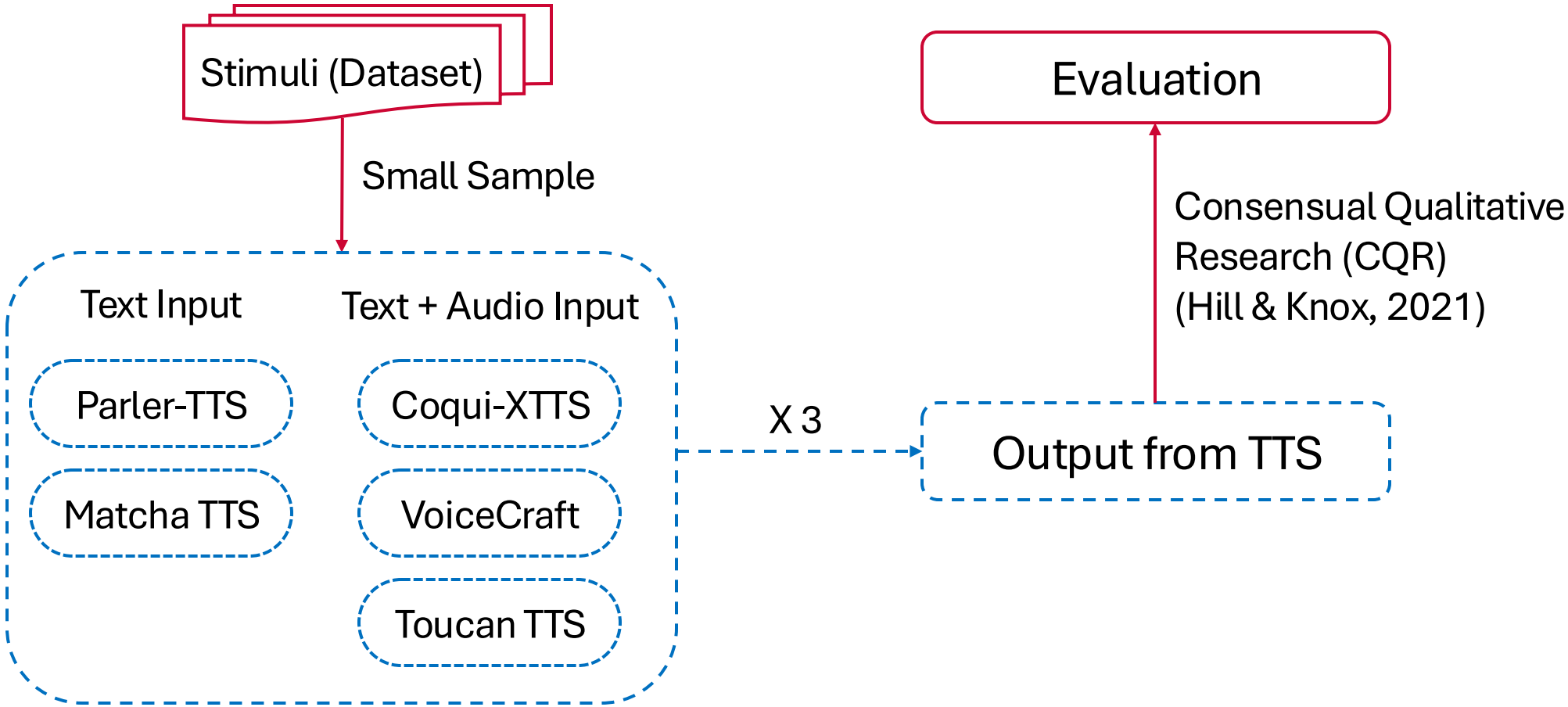
***(2) User modifications** for customizing speech prosody (e.g., gender, speech rate, pitch)*

Parler-TTS | Coqui-XTTS | VoiceCraft | Toucan TTS | Matcha TTS

Zero shot voice cloning

Study 1: Evaluation of TTS Systems

Procedure



Study 1: Evaluation of TTS Systems

Evaluation: Consensual Qualitative Research (CQR) Method

Qualitative questionnaire - five criteria

Table 3: *Results of the qualitative criteria questionnaire for Study 1*

Question	VoiceCraft	Parler	Toucan	Coqui	Matcha
1. Is the quality of the generated voice inconsistent each time?	X	X	X		
2. Is the sound quality noisy?	X			X	
3. Is the RTF* $>$ 1?	X	X	X	X	
4. Are there any inserted speech disfluencies?	X				X
5. Is the difference in meaning between Condition A and Condition B unclear?	X	X	X		

*Measured on Hugging Face Spaces, which may use different GPUs and may be affected by resource queuing.

* Real Time Factor (RTF)

Study 1: Evaluation of TTS Systems

Evaluation Results

Table 3: *Results of the qualitative criteria questionnaire for Study 1*

Question	VoiceCraft	Parler	Toucan	Coqui	Matcha
1. Is the quality of the generated voice inconsistent each time?	X	X	X		
2. Is the sound quality noisy?	X			X	
3. Is the RTF* $>$ 1?	X	X	X	X	
4. Are there any inserted speech disfluencies?	X				X
5. Is the difference in meaning between Condition A and Condition B unclear?	X	X	X		

*Measured on Hugging Face Spaces, which may use different GPUs and may be affected by resource queuing.

	VoiceCraft	Matcha-TTS
List_A[1]		
List_B[2]		

[1] The dad bought his kid, a car, and a plane.
[2] The dad bought his kid a car and a plane.

Study 1: Evaluation of TTS Systems

Evaluation Results

Table 3: *Results of the qualitative criteria questionnaire for Study 1*

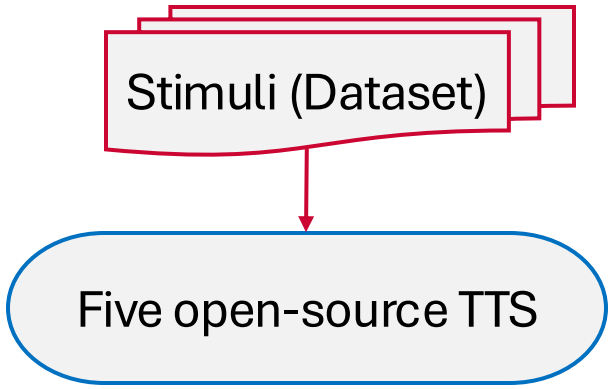
Question	VoiceCraft	Parler	Toucan	Coqui	Matcha
1. Is the quality of the generated voice inconsistent each time?	X	X	X		
2. Is the sound quality noisy?	X			X	
3. Is the RTF* $>$ 1?	X	X	X	X	
4. Are there any inserted speech disfluencies?	X				X
5. Is the difference in meaning between Condition A and Condition B unclear?	X	X	X		

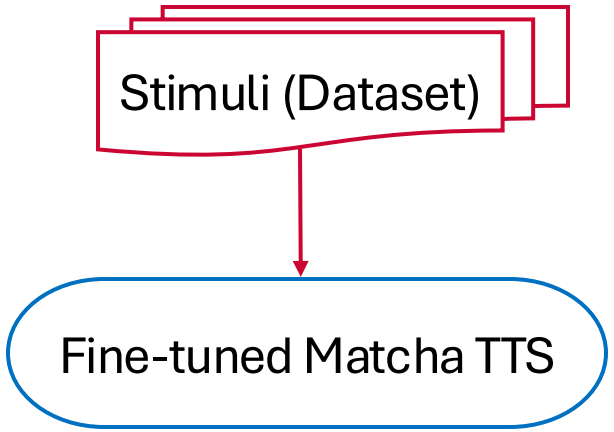
*Measured on Hugging Face Spaces, which may use different GPUs and may be affected by resource queuing.

- **Matcha-TTS:** while not perfect, it was **the most viable** at synthesizing punctuation-induced prosodic contrasts in A/B conditions.
- **Limitation:** **Still failed** to sufficiently convey the intended prosodic contrasts in full stimulus set.

Continuation to Study 2

Study 2 – Fine-tuning Matcha TTS





Study 1- Evaluation of TTS systems:

They all struggle to disambiguate the stimulus using prosodic cues, even when commas were included in the input.

Study 2- Fine-tuning a TTS system:

How can we **enhance** TTS output to produce predictable prosodic cues using a customized dataset?

In which textual contexts does this succeed or fail?

Study 2 – Fine-tuning Matcha TTS

Study Design

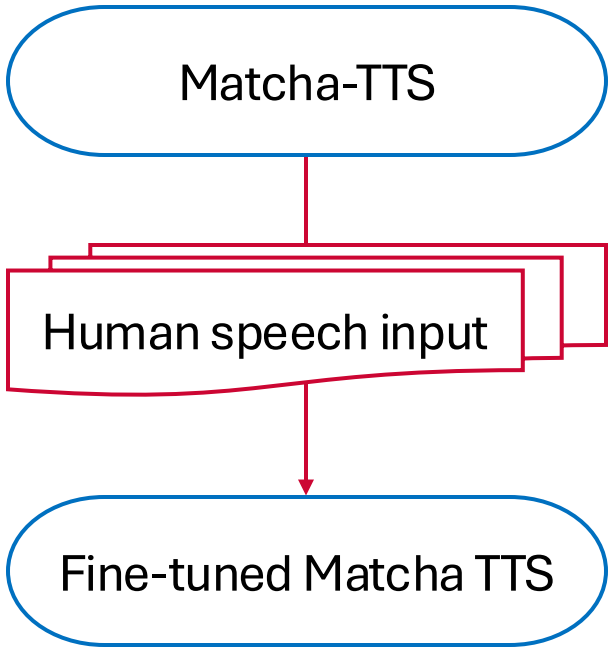
Fine-tuning with a tailored speech input

- To replicate human prosodic cues
- To produce consistent and measurable punctuation induced-prosodic parsing

Matcha-TTS as an example (Mehta et al., 2024)

Study 2 – Fine-tuning Matcha TTS

Study Design: Measurement



Focus: *Pause duration*

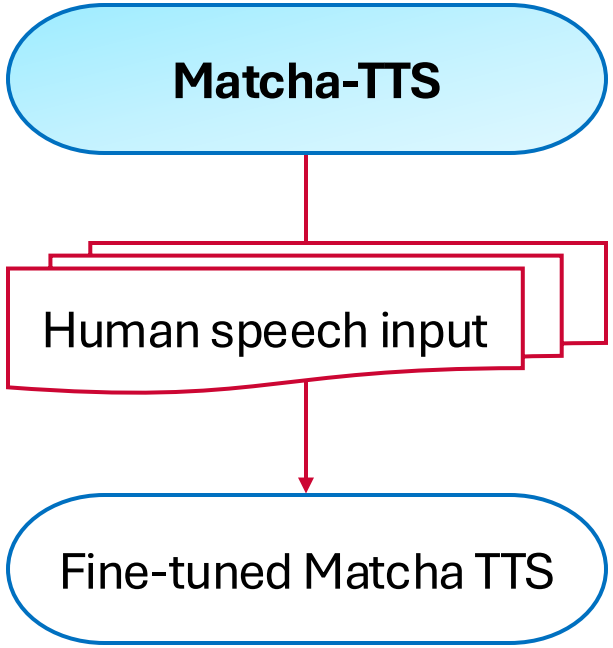
- Critical prosodic cue (Wagner & Watson, 2010)
- Denoted textually as a comma

Customized speech input with distinct pause contrasts

- *Longer pauses* at commas in *Condition A*
- *Shorter pauses* without commas in *Condition B*

Study 2 – Fine-tuning Matcha TTS

Study Design: Procedure

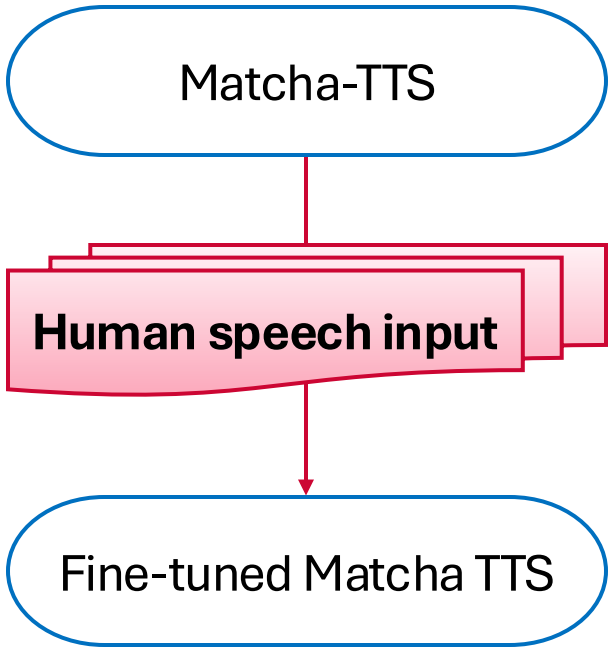


Before Fine-tuning

- Default values: Ordinary Differential Equation (ODE) solver and sampling temperature
- Speaking rate: reduced to 0.85

Study 2 – Fine-tuning Matcha TTS

Study Design: Procedure



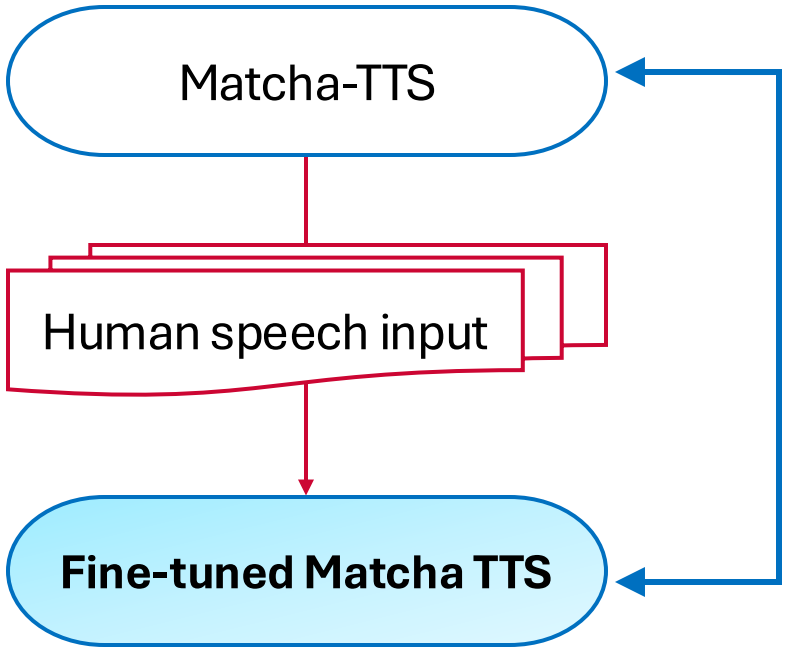
Fine-tuning dataset was recorded & fed

Seen set - 100 training sentences

- 10 A/B pairs per type (40, retained from Study 1)
- 10 tense/lax vowel contrast fillers per type (40)
- 20 generic fillers
- 10% held out for validation

Study 2 – Fine-tuning Matcha TTS

Study Design: Procedure



Performance Testing

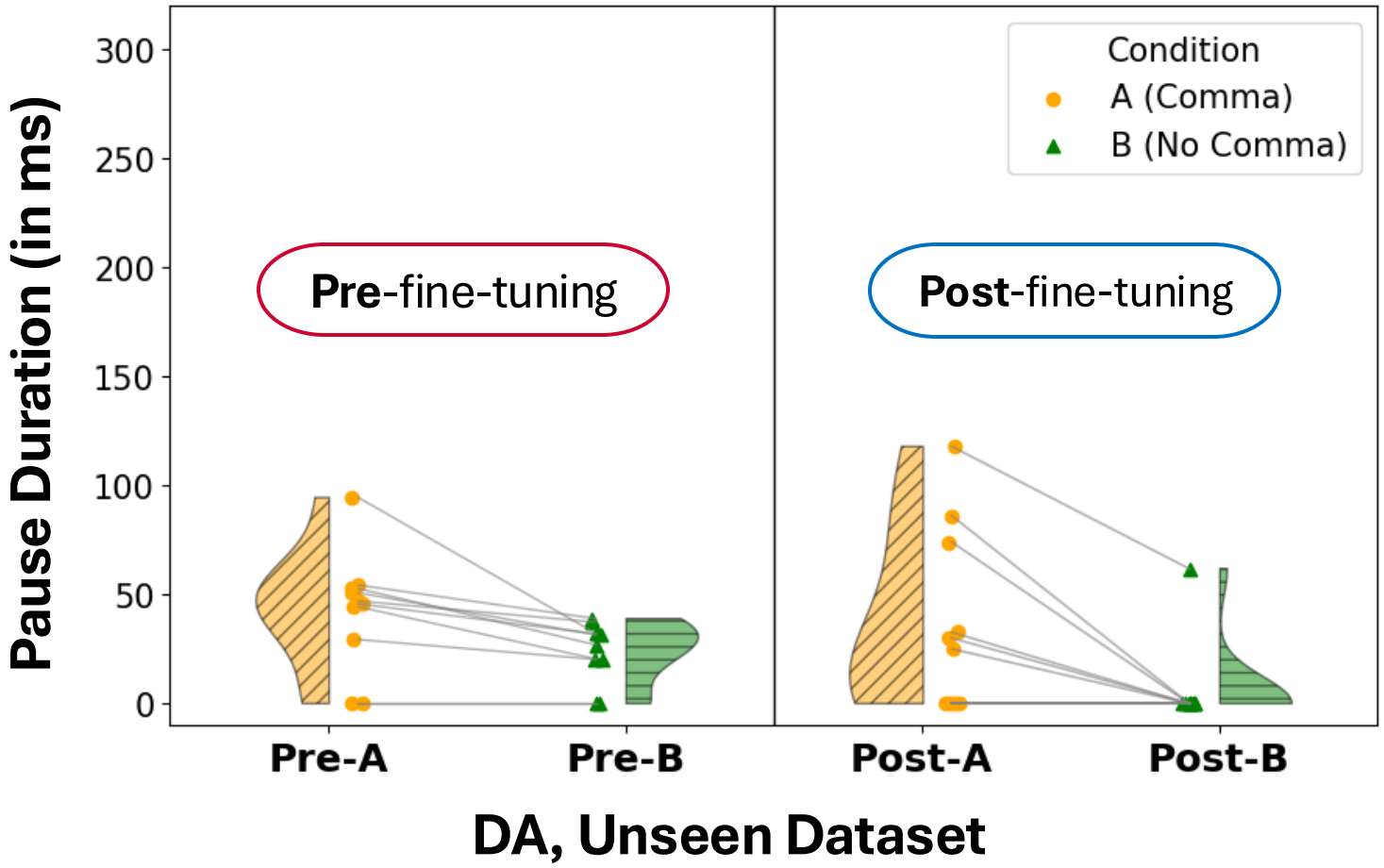
- 10 ***seen*** A/B pairs for each DA and List type (40, from training)

To test for generalizability

- 10 ***unseen*** A/B pairs for each DA and List type (40, new)

Study 2 – Fine-tuning Matcha TTS

Results – DA type



DA type sentences &

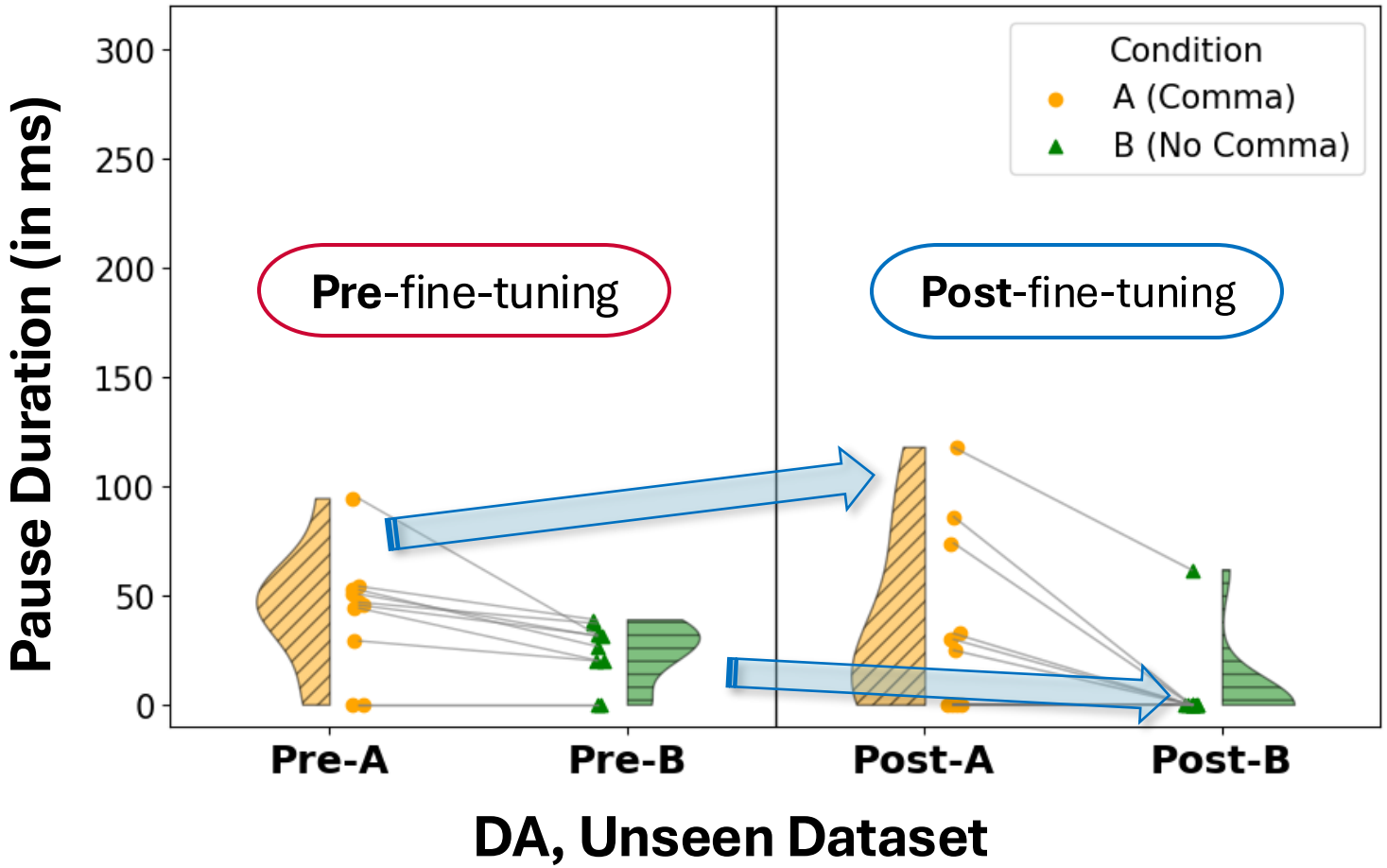
Conditions A and B

(A) *He* ***wants*** *to* ***hit,*** ***Gabe***

(B) *He* ***wants*** *to hit* ***Gabe***

Study 2 – Fine-tuning Matcha TTS

Results – DA type



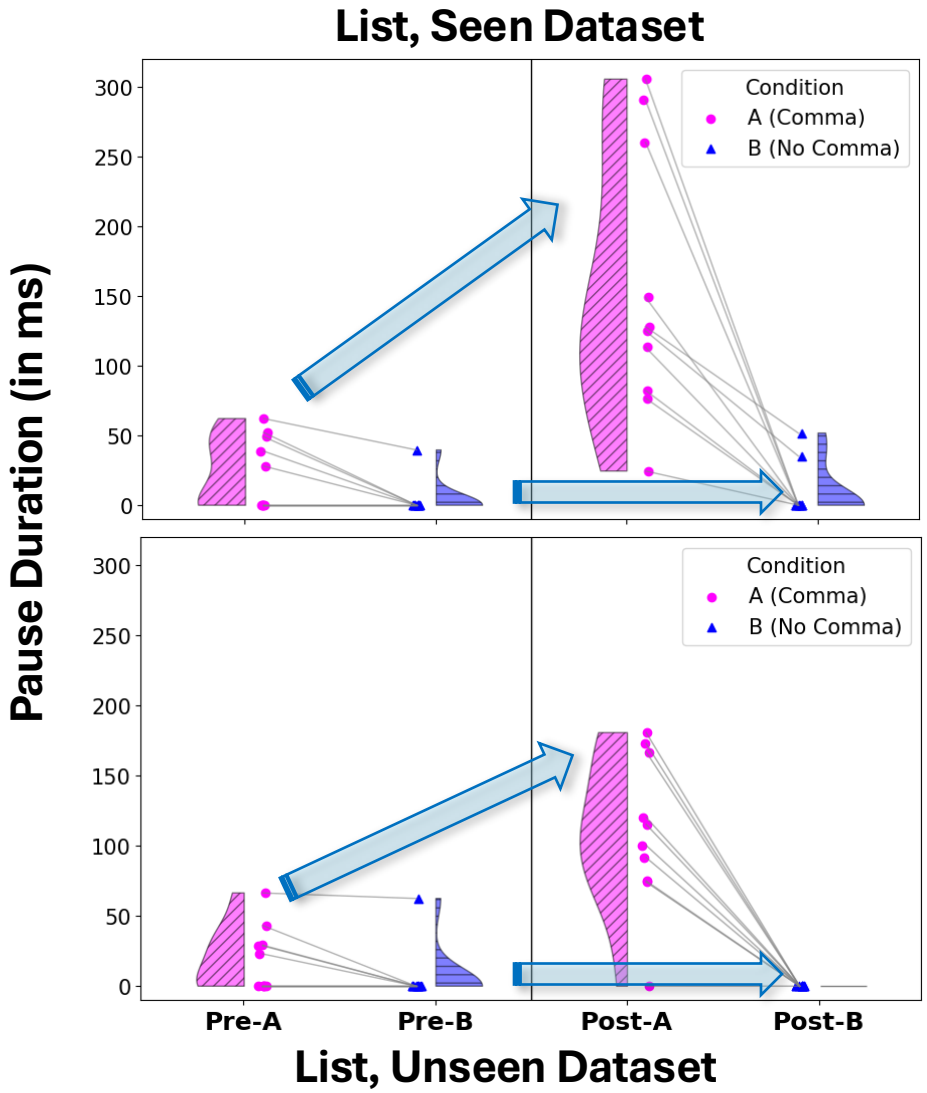
Despite some improvements…

Pause duration contrasts between pre- and post-fine-tuning: **insignificant** in both datasets (Seen: $p = 0.89$, Unseen: $p = 0.59$)

For the DA types, fine-tuning did not produce significant contrasts.

Study 2 – Fine-tuning Matcha TTS

Results – List type



List type sentences - <u>first</u> prosodic boundary

(A) *The **mom** drove her **kids**, Eu**gene**, and A**dele***

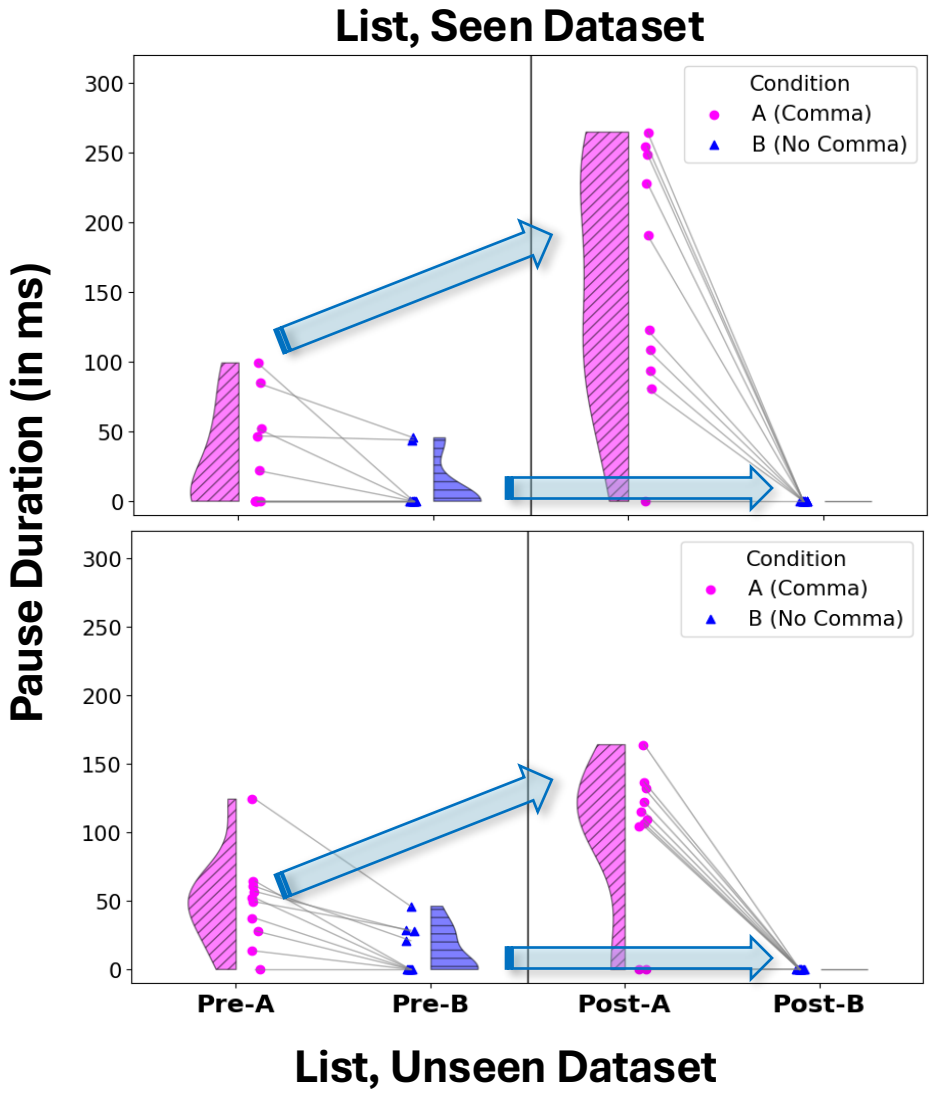
(B) *The **mom** drove her kids Eu**gene** and A**dele***

Pause duration contrasts between pre- and post-fine-tuning: **significant** in both datasets.

(Seen: $p < .01$, Unseen: $p < .001$)

Study 2 – Fine-tuning Matcha TTS

Results – List type



List type sentences – <u>second</u> prosodic boundary

*(A) The **mom** drove her **kids**, Eu**gene**, and A**dele***

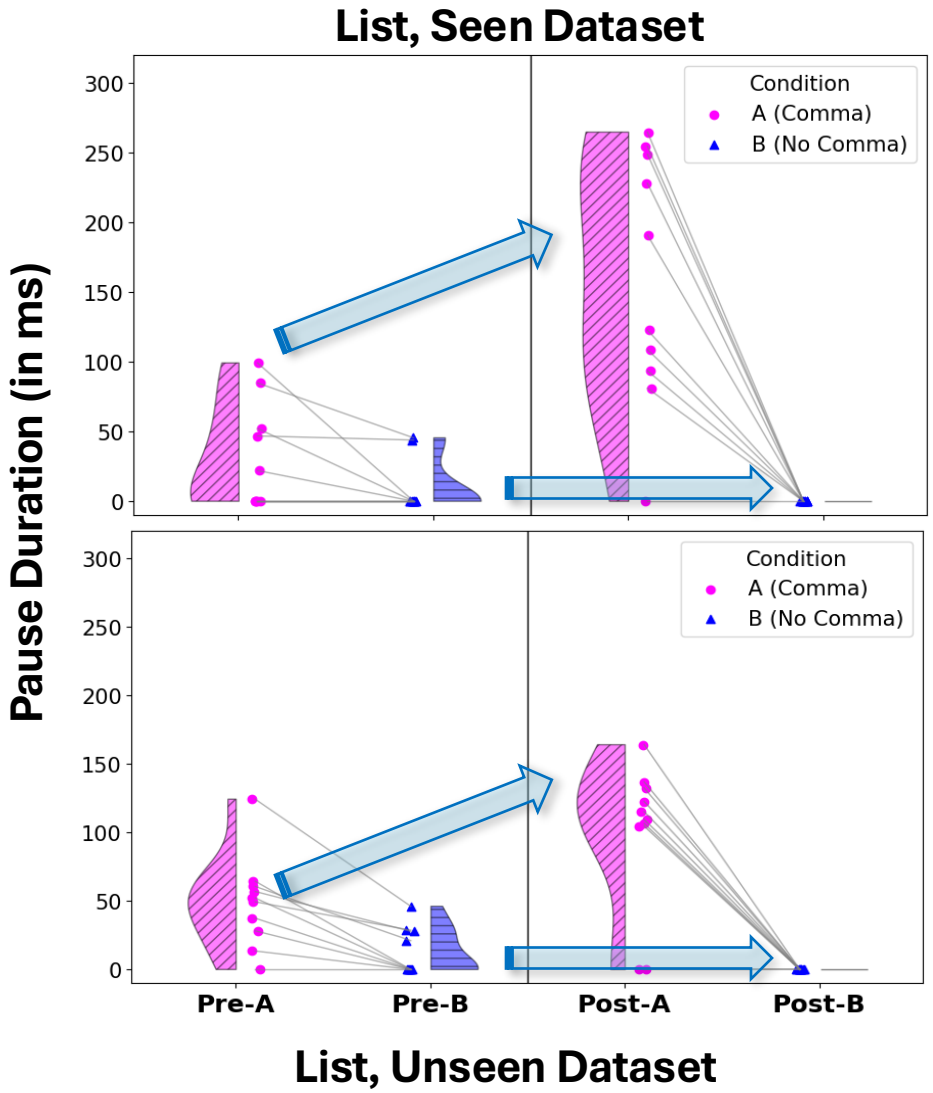
*(B) The **mom** drove her kids Eu**gene** and A**dele***

Pause duration contrasts between pre- and post-finetuning: **significant** in both datasets.

(Seen: $p < .01$, Unseen: $p < .05$)

Study 2 – Fine-tuning Matcha TTS

Results – List type



For the List types, fine-tuning effectively reproduced significant pause duration contrasts.

Discussion and limitations

Study 1- Evaluation of TTS systems:

Five TTS systems in their default settings **fail** to accurately replicate human prosody cues, **even when commas were included in the input.**

Our dataset can be included in large TTS corpora to better train the models.

Study 2- Fine-tuning a TTS system:

We demonstrated **a pipeline** to synthesize speech with clear prosodic cues by (1) **customizing** minimal amount of training input and (2) **fine-tuning** a TTS system.

- **Successful for List types**

Our approach can be used with other TTS models to improve the accuracy to the human speech subtleties.

28

Discussion and limitations

Study 1- Evaluation of TTS systems:

- Collecting **large-scale** subjective ratings for the questionnaire

Study 2- Fine-tuning a TTS system:

- **DA types:** an issue in Matcha-TTS adhering to limited prosodic patterns
- Limited sample size and sole focus on pauses being insufficient for DA types
- **Future work: multiple acoustic measurements along with pause duration**

Conclusion

- We showed that modern TTS systems **struggle with prosodically ambiguous sentences**, particularly in distinguishing **comma contrasts**.

- **In study 1,** we showed an evaluation that can be easily replicated by non-computer scientists and non-experts using online demos.

- **In study 2,** we demonstrated how to fine-tune a TTS system with a bespoke dataset of prosodic contrasts to resolve the ambiguities.

SFU SIMON FRASER UNIVERSITY

INTERSPEECH 2025

Thank You

The complete datasets and audio files are available in the supplementary materials.

Paper & Data

ha_eun_shim@sfu.ca olivia_yung@sfu.ca

Generating Consistent Prosodic Patterns from Open-Source TTS Systems

Ha Eun Shim[1], Olivia Yung[1], Paige Tuttösí[2], Boey Kwan[1], Angelica Lim[2], Yue Wang[1], H. Henny Yeung[1]
Department of Linguistics[1] and School of Computing Science[2], Simon Fraser University, Canada
INTERSPEECH 2025 @ Rotterdam, Netherlands

Selected References

C. Clifton, K. Carlson, and L. Frazier, “Tracking the what and why of speakers’ choices: Prosodic boundaries and the length of constituents,” *Psychonomic bulletin & review*, vol. 13, no. 5, pp. 854–861, 2006.

C. E. Hill and S. Knox, *Essentials of consensual qualitative research*. American Psychological Association, 2021.

M. Wagner and D. G. Watson, “Experimental and theoretical advances in prosody: A review,” *Language and cognitive processes,* vol. 25, no. 7-9, pp. 905–945, 2010.

S. Mehta, R. Tu, J. Beskow, Székely, and G. E. Henter, “Matcha TTS: A fast TTS architecture with conditional flow matching,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 341–11 345.

**Complete references are available in our paper.*

SFU SIMON FRASER UNIVERSITY

Generating Consistent Prosodic Patterns from Open-Source TTS Systems

Ha Eun Shim[1], Olivia Yung[1], Paige Tuttösí[2], Boey Kwan[1], Angelica Lim[2], Yue Wang[1], H. Henny Yeung[1]

Department of Linguistics[1] and School of Computing Science[2], Simon Fraser University, Canada

Presenters: Ha Eun Shim, Olivia Yung
Department of Linguistics, Simon Fraser University
INTESPEECH 2025 @ Rotterdam, Netherlands

***Equal Contributions**