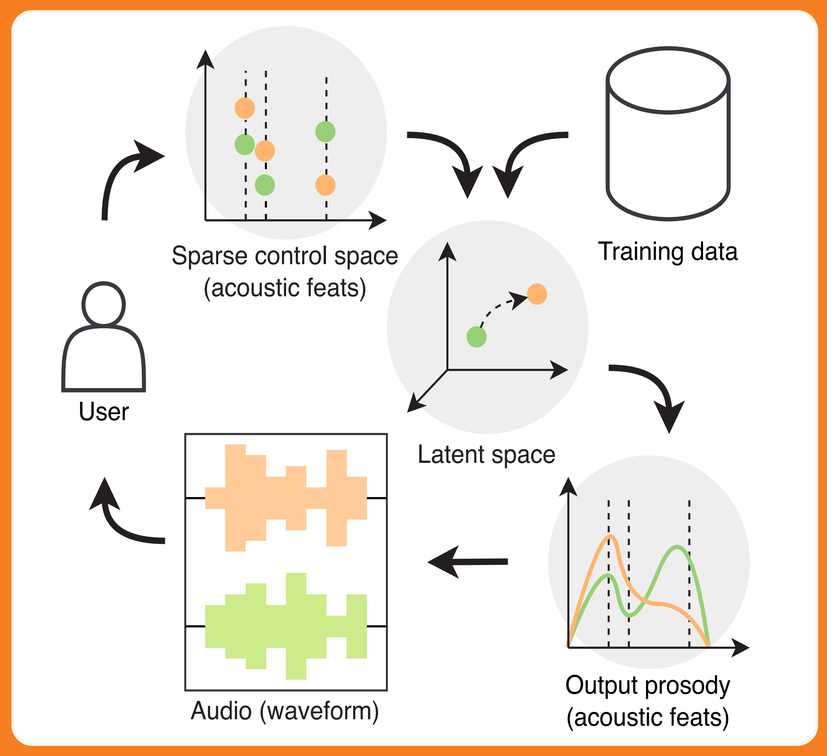
We enable Efficient and Flexible Prosody Control for Human-in-the-Loop Speech Synthesis



Controllable Prosody Generation with Partial Inputs

Dan Andrei Iliescu, Devang S Ram Mohan, Tian Huey Teh, **Zack Hodari**

Introduction

Appropriate prosodic choices depend on the context. One approach is for a human-in-the-loop (HitL) to pick the best prosody.

Often there are specific nuanced prosodic choices that convey the intended meaning in a given context.

We propose a system where HitL users can provide any number of prosodic controls. This allows for flexibility and removes the need for redundant (inefficient) work defining the entire prosodic specification.

MICVAE

Multiple **I**nstance **C**onditional **VAE**

Encoder design enables user control with sparse "bag-of-feature" inputs

- Positional embeddings
- Feature embeddings
- Self-attention for multiple instance aggregation

Predictions passed to controllable TTS model.

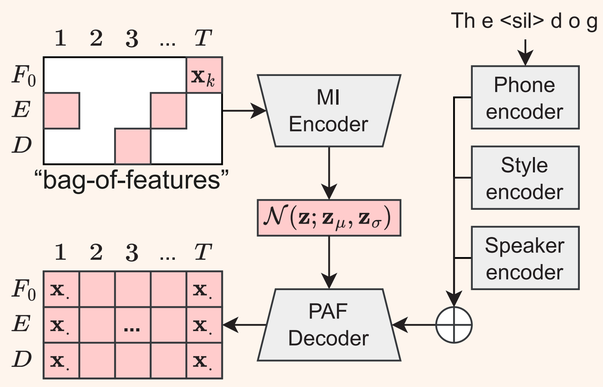


Figure 1: MICVAE prosody model

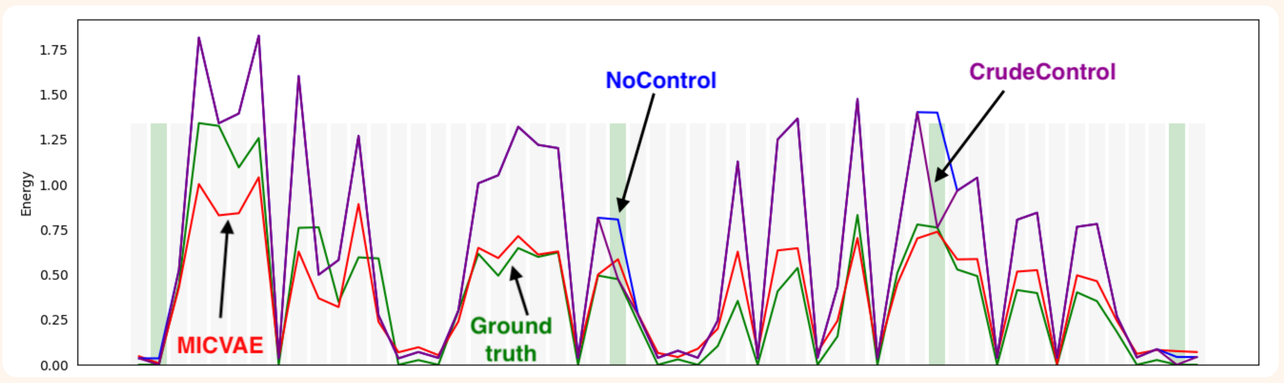


Figure 2: HitL control example. Green bars indicate phones controlled by human user

1. Efficiency

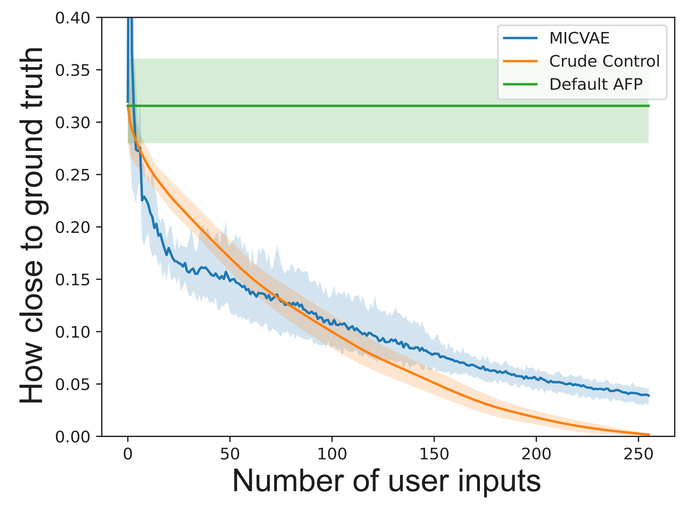


Figure 3: MICVAE can get closer to the goal with much fewer user inputs

2. Robustness

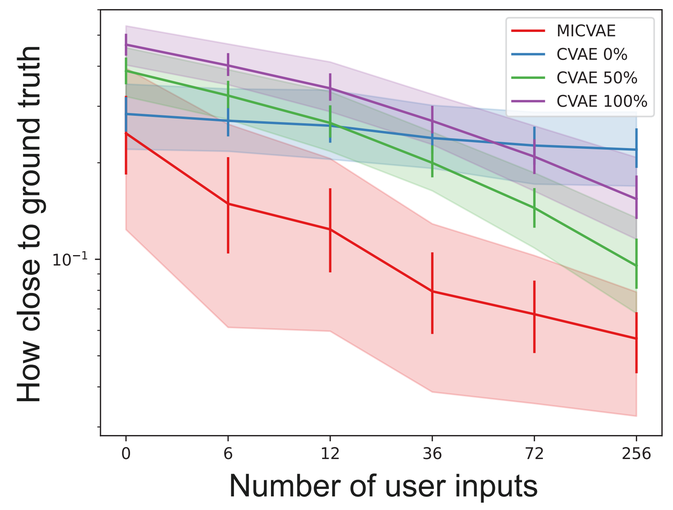


Figure 4: MICVAE works even if the pattern of user-provided inputs is different at train and test time

3. Faithfulness

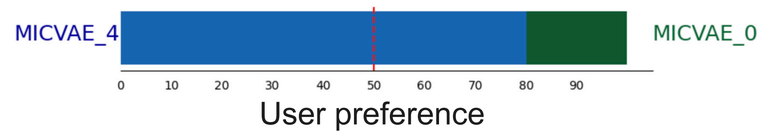


Figure 5: MICVAE improves the prosody according to the user goal

PAPERCUP®

UNIVERSITY OF CAMBRIDGE

View paper

See samples

Conclusion:

We designed a prosody prediction model that can be conditioned on sparse user inputs. This led to more efficient, robust, and faithful prosody control.