icassp 2022 Singapore

Unsupervised Word-Level Prosody Tagging for Controllable Speech Synthesis

Poster ID: 3004

Yiwei Guo, Chenpeng Du, Kai Yu
MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Motivation

In previous researches:

- Modeling of the **diversity** and **controllability** of speech prosody in neural text-to-speech systems already exists.
- But, it's still hard to control prosody **by words without reference speech**. Most of them use fine-grained prosody or explicit prosody features.
- In other words, difficult **human-friendly** control.

Our goal includes:

- **Word-level** prosody modeling as basic prosody units.
- **Prosody diversity**, as in GMM-MDN (Du, 2021)
- **Interpretability**, for humans to understand
- **Easy controllability**: only need to provide an additional abstract signal

Difficulty & Idea

▸ *Words' prosody may vary much, e.g. "congratulations" & "cat" should not be treated the same.*

▸ *Prosody may be disentangled with phonetic content.*

✦ To generate a **prosody tag** for every word in each utterance.

✦ **Words should be grouped** by phonetics at first.

✦ Different groups of words should have different tagging sets.

✦ Each prosody tag should specify one kind of natural prosody in that word group.

Methods

A) Two-stage word-level **prosody tagging** by clustering
- **Stage 1**: word-level prosody extraction and decision tree clustering
- **Stage 2**: Gaussian mixture clustering

B) **Prosody controlling** with prosody tags

A) Two-stage prosody tagging

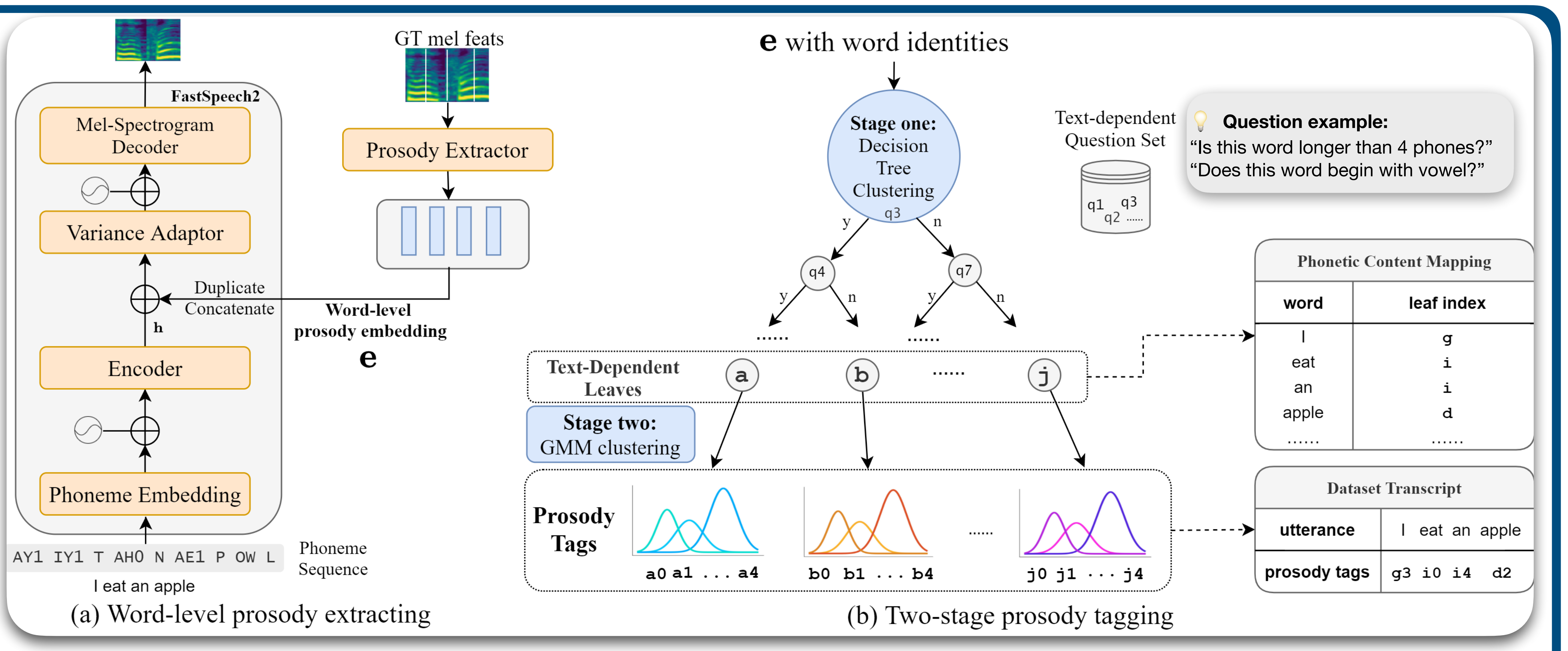
Stage 1:

- Word-level prosody embedding $\mathbf{e}$ extraction: jointly train a "prosody extractor" with FastSpeech2.
- Decision tree clustering: iteratively select a **phonetic question** that best split data according to Gaussian log-likelihood:

$$\text{Loop}\begin{cases} LL^{(i)} = \sum_{\mathbf{e}\in\mathscr{E}^{(i)}} \log \mathcal{N}\left(\mathbf{e} \mid \boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}^{(i)}\right) \\ \Delta_q LL^{(i)} = LL^{(\text{left child by } q)} + LL^{(\text{right child by } q)} - LL^{(i)} \\ j = \arg\max_{i\in\text{leaf nodes}} \left(\max_{q\in Q} \Delta_q LL^{(i)}\right) \end{cases}$$

Stage 2:

GMM clustering: applied to every leaf node.
Prosody tag is defined as Gaussian index.



Phonetic Content Mapping

word	leaf index
I	g
eat	i
an	i
apple	d
......	......

Dataset Transcript

utterance	I eat an apple
prosody tags	g3 i0 i4 d2

(a) Word-level prosody extracting

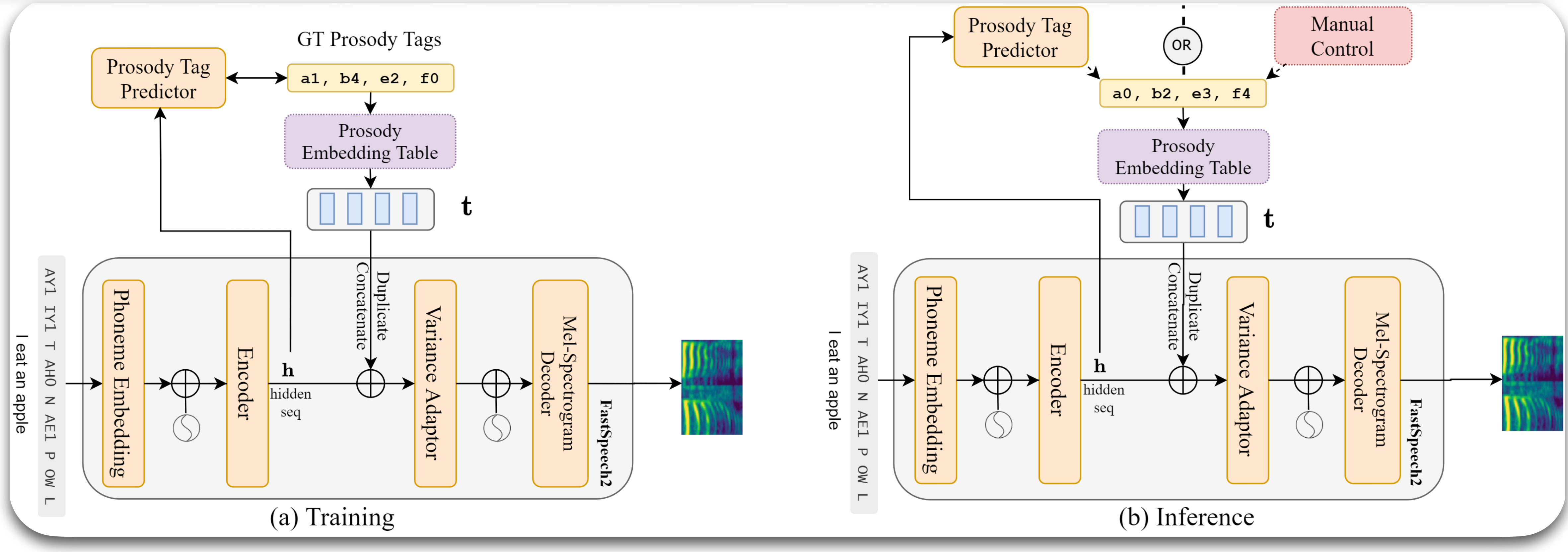
(b) Two-stage prosody tagging

Decision tree: disentangles prosodic & phonetic information; **GMM:** discovers diversified prosody structure

B) Prosody control

(a)Training: Use a **prosody predictor** to predict prosody tags from text input. This is modeled as a classification task, and cross entropy loss $\mathscr{L}_{PP}$ is added to the FastSpeech2 loss $\mathscr{L}_{\text{FastSpeech}}$.

(b)Inference: For every word, we can either specify the prosody tag, or let the prosody predictor find the most suitable tag. In other words, **manual control** is easy and natural with prosody tags.



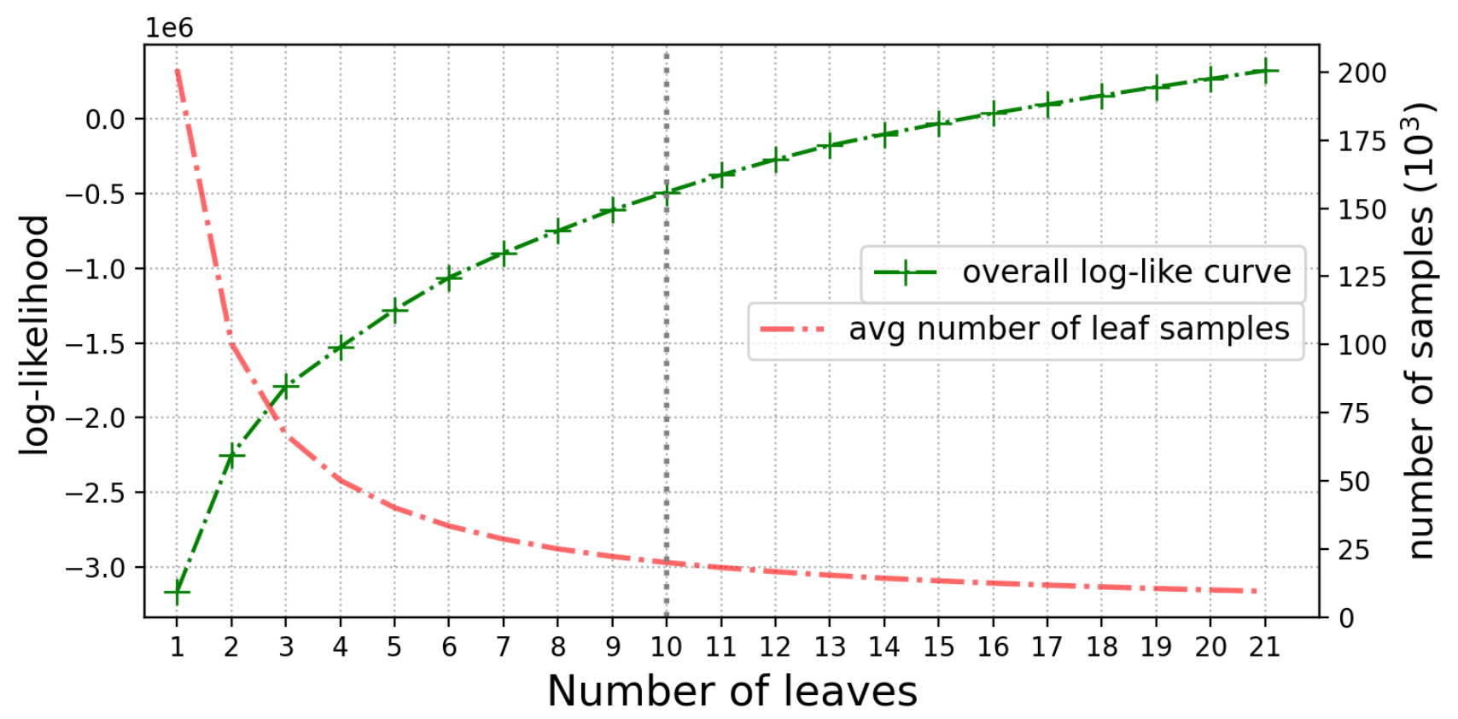
(a) Training

(b) Inference

Experiments

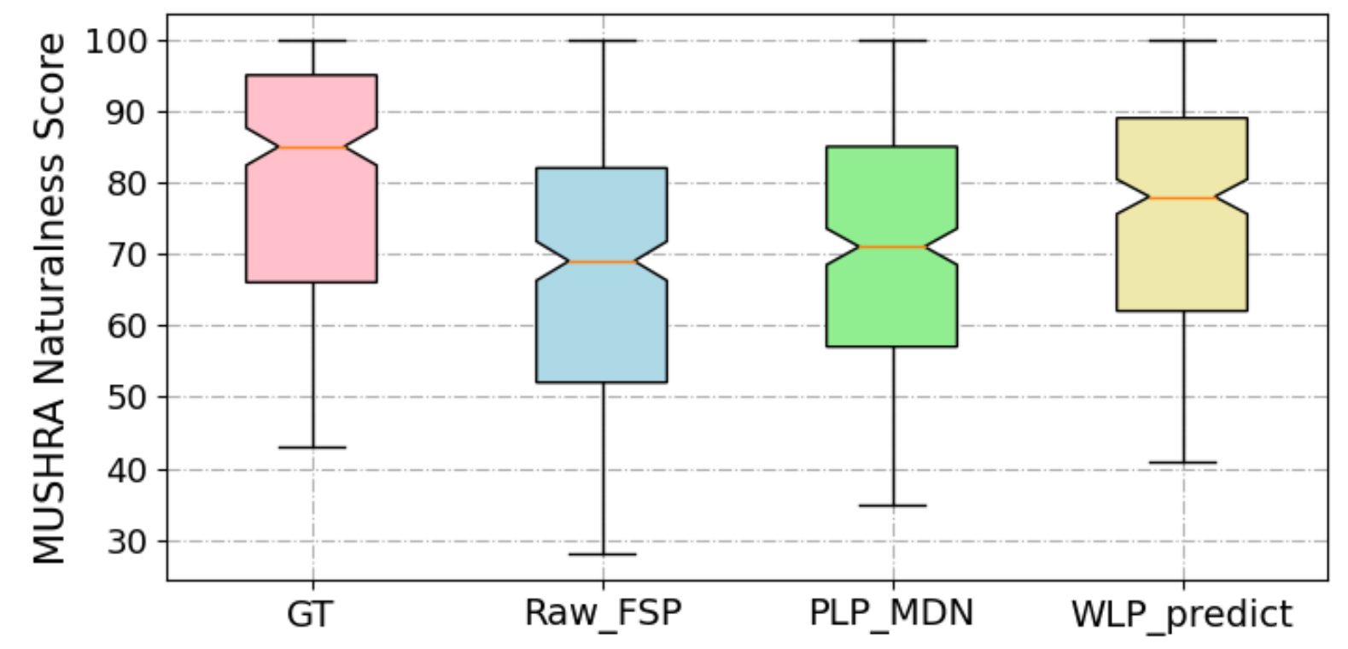
A) Decision Tree

Experiments are held on **LJspeech**. With the tree growing, the overall likelihood of the data increases substantially. This means a lot of phonetic information are split out of the prosody embeddings.
We choose 10 leaves and apply GMM with 5 components for each leaf. This generates 50 prosody tags in total.



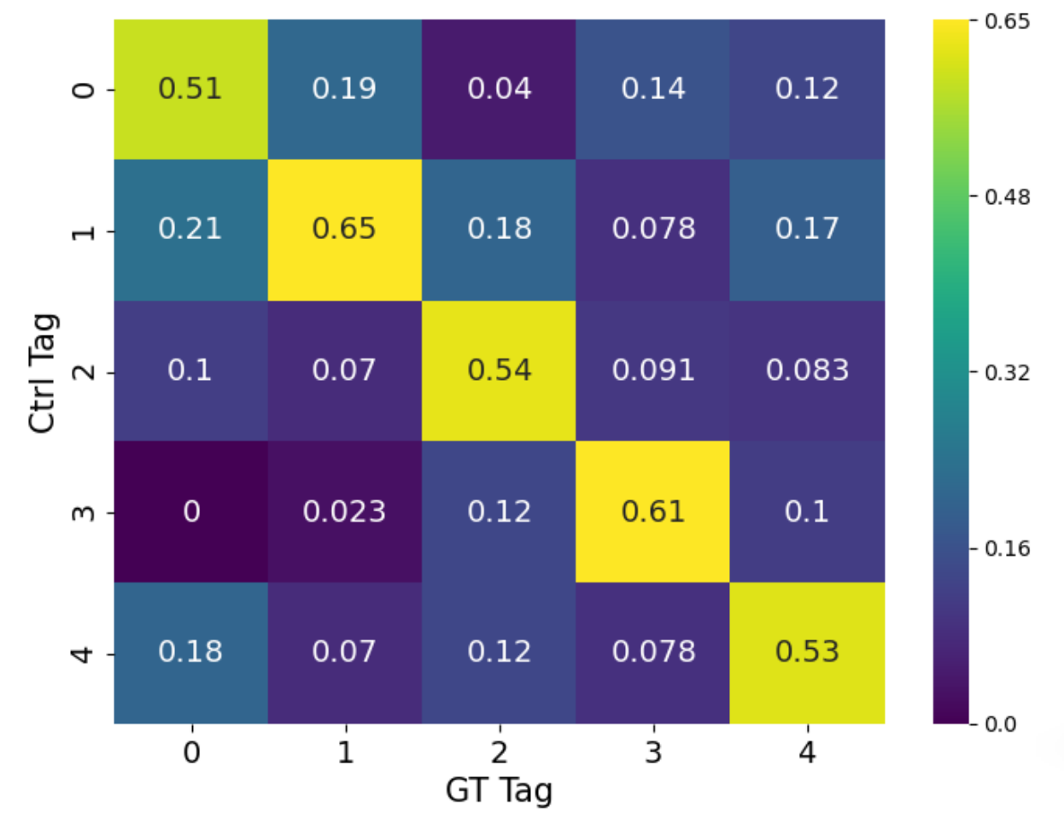
B) Naturalness

We let the prosody predictor freely predict the prosody tags in inference stage. Our word-level prosodic model is denoted as WLP_predict.
This results in better naturalness compared with raw FastSpeech2 and PLP_MDN, a phone-level prosody modeling approach (Du, 2021), as words are more appropriate prosody modeling units.



C) Prosody Controllability

Here we control the words by each of the five prosody tags. With ground truth recordings provided, listeners are asked to select which of the controlled synthetic word is most similar to the recording, in terms of prosody.
In most cases it's easy to identify the true prosody tags. This means real word-level prosody can be represented and reconstructed with our prosody tags. Also, our method owns good prosody diversity, as other tags sound differently.



Check out the examples here!

Audio demo

https://cantabile-kwok.github.io/word-level-prosody-tagging-control/