Using a Large Language Model to Control Speaking Style for Expressive TTS

Atli Thor Sigurgeirsson, Simon King

The Centre for Speech Technology Research, University of Edinburgh, UK

s2063518@ed.ac.uk, simon.king@ed.ac.uk

Abstract

Large generative language models have been used to solve various language-related tasks. We explore whether such models can suggest appropriate prosody for expressive TTS. We train a TTS model and then prompt the language model to suggest appropriate changes to pitch, energy and duration. The prompt can be designed for any task and we prompt the model to make suggestions based on target speaking style and dialogue context. The proposed method is rated most appropriate in 49.9% of cases compared to 31.0% for a baseline model.

Index Terms: speech synthesis, style modelling, prosody

1. Introduction

Contextual Word Embeddings (CWEs) encode semantic and syntactic word relations within a context (input text) and have been used for text-to-speech (TTS) prosody prediction [1, e.g.]. Large Language Models (LLMs) [2, e.g.] utilize CWEs for word representations and are, when sufficiently large, capable few-shot learners in many language-related tasks [2]. Such LLMs are highly flexible since the target task can be fully defined using only natural language. The task description and the task itself (the prompt) are then used to query the LLM for a solution without any fine-tuning or model parameter updates.

In the current work, we explore whether a LLM can be queried to directly predict prosodic modifications, which can be implemented in a text-to-speech model (here, a modified FastSpeech-2[3]). We prompt the InstructGPT LLM, which has been fine-tuned to follow natural language instructions [2]), to suggest context-appropriate modifications to the acoustic features (F₀, energy and duration) that are employed in our proposed model architecture. The flexibility of our design allows for changing the task of the LLM arbitrarily. We explore its capabilities for two types of conditional information: 1) a target speaking style described using natural language and 2) a previous line in an expressive dialogue. As far as we can tell, the proposed method is novel and the first one to prompt an LLM, using only natural language, to solve a specific task in TTS.

2. Method

The TTS model separately predicts F₀, energy and duration. Following [4], we replace the FastSpeech-2 variance adapters and duration model with the low-level prosody predictor and Gaussian upsampling modules. This allows us to make independent modifications to F₀, energy and duration predictions at the phone-level. We propose to prompt an LLM to suggest more appropriate settings of those values based on the target text and optional contextual information. Instead of setting absolute F₀, energy and duration values, we instead make changes relative to the values initially predicted by the TTS model. Those initial values reflect the statistics of the training corpus, something the LLM knows nothing about.

Given the sequences of predicted per-phone duration, F₀ and energy for an input sequence of length T \( \{d_i, p_i, e_i\}_{i=1:T} \) we aim to suggest more appropriate values (\( \{d'_i, p'_i, e'_i\} \)). We propose a 2-level modification procedure to control both utterance (‘global’) and per-word (‘local’) prosodic effects. Global modifications are applied to all phones in the utterance using three parameters. All durations and energies are scaled by \( G_d, G_e \), respectively and shift all per-phone \( F_0 \) values by \( G_p \) to make the most use of the speaker’s \( F_0 \) range while preserving the quality of the voice. \( G_d \) and \( G_e \) are each limited to \([0.5, 2]\) while \( G_p \) is limited to a range such that all \( p'_i \) remain within the natural \( F_0 \) range of the target speaker.

Local word modifications are realised using three parameter sequences, \( \delta_{1:W}, \pi_{1:W} \), and \( \epsilon_{1:W} \) for an input text consisting of \( W \) words. The values resulting from both global and local modifications for phone \( i \) appearing in word \( j \) is given by:

\[
\begin{align*}
    d'_i &= d_i \cdot (G_d \cdot \delta_j), \quad G_d \in [0.5, 2], \quad \delta_j \in [1.0, 2.0] \\
    e'_i &= e_i \cdot (G_e \cdot \epsilon_j), \quad G_e \in [0.5, 2], \quad \epsilon_j \in [1.0, 2.0] \\
    p'_i &= p_i + (G_p + \pi_j), \quad G_p + \pi_j \in [p_{\text{min}}, p_{\text{max}}]
\end{align*}
\]

where \( p_{\text{min}} \) and \( p_{\text{max}} \) are the minimum and maximum allowed changes in \( F_0 \) determined by corpus statistics. After choosing...
Figure 2: The LLM suggests acoustic modifications given the target text and, optionally, a speaking style or dialogue context. The rules, supplied to the LLM in the prompt, make the LLM produce reasoning for the modifications.

3. Subjective valuation

We compare the proposed method to: 1) an unmodified FastSpeech-2 model (baseline) and 2) random, where modification parameters are pseudo-randomly drawn from an appropriate distribution for each parameter. We evaluate random to account for any listener bias resulting from arbitrary acoustic variation. We evaluate appropriateness with a preference A/B/C design. We recruit native English-speaking participants and each screen is evaluated by 8 different listeners. All models are trained on LJSpeech [7].

3.1. Target style task

In this task, participants are shown the target text and speaking style. We create a small text corpus for this where we first select 7 speaking styles and, for each one, then generate 10 target texts that we deem would be appropriate for that style. Here we define a speaking style as a distinct manner of speaking appropriate for the argument. The flexibility of the approach means that we can define highly specific speaking styles to evaluate. Our list includes styles such as “frightened” and “in a hurry”. Participants are asked to base appropriateness on how well they thought the overall quality of the voice, emotion and attitude fit the target text. The results for this task are shown in the first row in Table 1. The proposed method is preferred at a

Table 1: A/B/C appropriateness preference results for both the target style and dialogue tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Proposed</th>
<th>Baseline</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target style</td>
<td>51.4%</td>
<td>30.9%</td>
<td>17.7%</td>
</tr>
<tr>
<td>Dialogue</td>
<td>48.4%</td>
<td>31.0%</td>
<td>20.6%</td>
</tr>
<tr>
<td>Mean</td>
<td>49.9%</td>
<td>31.0%</td>
<td>19.1%</td>
</tr>
</tbody>
</table>

>70% higher rate than baseline. Raters did not indicate a high preference for random.

3.2. Dialogue task

Here, the LLM is instructed to make modifications based only on a previous line in a dialogue and the target text. We generate 60 two-line dialogues where the target text would naturally fit a hidden speaking style. We evaluate the proposed method again against baseline and random. Participants indicate a preference for the proposed method as shown in Table 1.

4. Acknowledgements

This work was supported in part by Huawei and the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology and Language Sciences.

5. References