Educational Song Composition using Large Language Model (LLM) with Synthesized Vocal Singing

Nagyung KIM

Communication Systems, École polytechnique fédérale de Lausanne (EPFL)

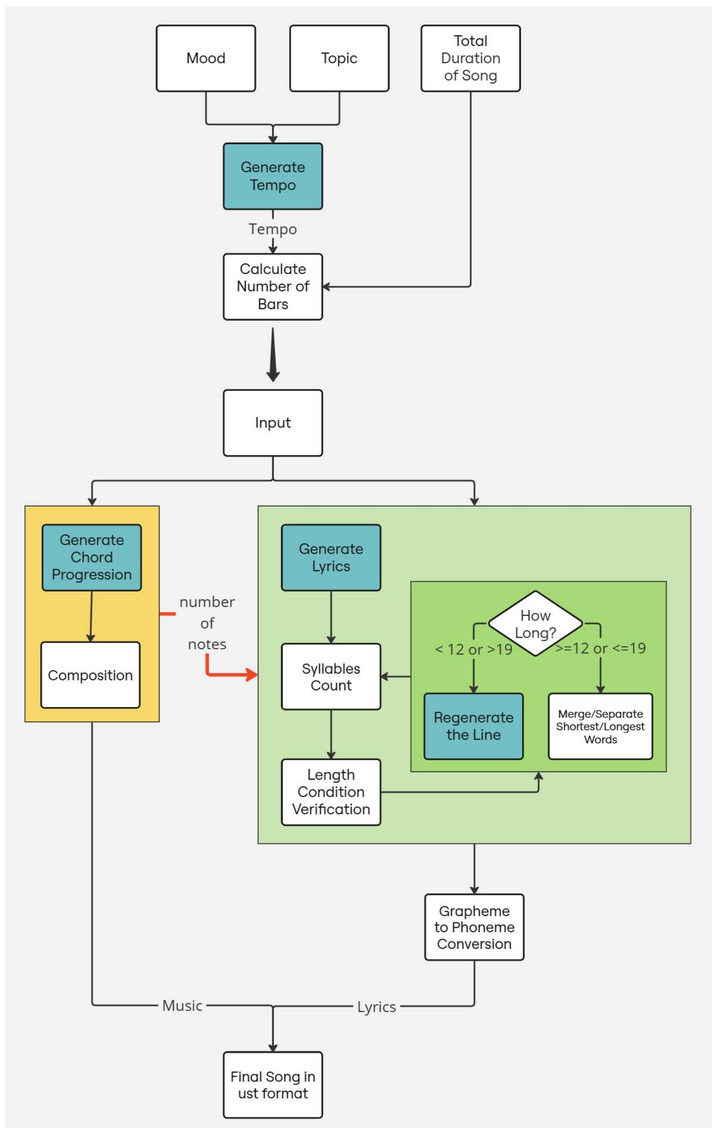
1. Introduction

This program uses a **large language model (LLM), Chat GPT 3.5 turbo** to generate chord progressions and lyrics based on user input, such as mood and topic. The generated song is then sung using **DiffSinger, a diffusion probabilistic model for singing voice synthesis (SVS).** This software empowers educators to instantly create songs that enhance student learning by teaching vocabulary, concepts, rhythm, and musical skills through engaging, real-time music experiences.

Key features

- **Automatic Song Composition**: Teachers with no musical background can create songs on any topic and mood instantly.
- **Singing Voice**: A fully synthesized sung version allows students to sing along, aiding in language acquisition and musical understanding.
- **Interactive Learning Tool**: Teachers can use this as an engaging, interactive tool to teach not only musical concepts and vocabulary but also enhance student participation and retention through music.

2. Model A: Static Song & Dynamic Lyrics



Strengths	Weaknesses
Consistent Musical Quality The music maintains a high standard as it has no restriction	**Reduced Lyric Quality** Repeated regeneration of lyrics may lower the quality
Performance Adaptability Lyrics can change without altering the music	**Lyric Overlap** Lyrics from one measure may spill over into the next measure
Scalable Composition Flexibly adjusts from simple tunes to complex arrangements without losing structural integrity	**Creativity Limits** Lyrics are constrained by the fixed musical structure

Potential

Significant **enhancements in musical quality**, such as the use of **Markov chains** for better composition, could allow this model to surpass others by providing both dynamic lyrics and elevated musical creativity.

- Process done by LLM
- Music Composition
- Lyrics Generation
- Optimizing Lyrics Length (method 2)

Optimizing Lyrics Length: Challenges and Techniques

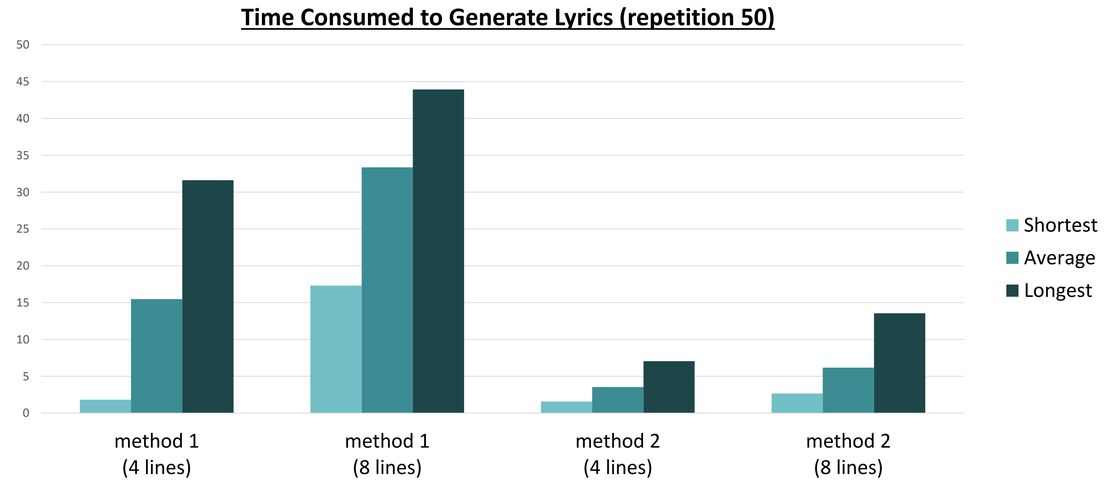
Method 1: Targeted Section Regeneration

- Only sections failing to meet the syllable conditions are regenerated.
- *Challenges*:
  - More time-consuming as it may require multiple iterations to correct specific parts.
  - Occasionally degrade the overall lyrical context, making sections seem out of place or disjointed.
  - In general, this methods gives the *best quality* music for model A.

Method 2: Targeted Section Regeneration with Hardcoded Word Manipulation

- Involves artificially merging the shortest words and splitting the longest words to adjust the syllable count.
- *Challenges*:
  - Merging/Separating distorts pronunciation and disrupt natural language flow
  - Generally, *the most efficient* among the three in terms of managing time constraints.

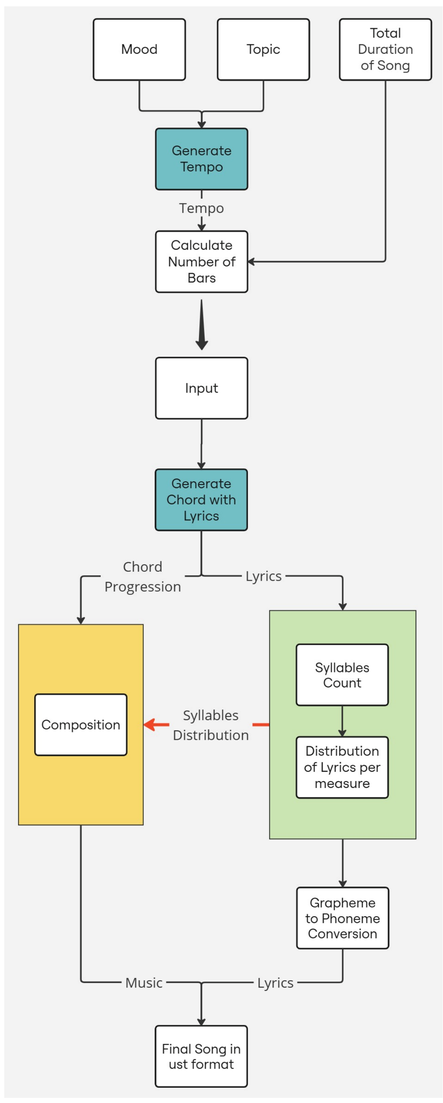
Time Consumed to Generate Lyrics (repetition 50)



Observation & Select of method

- **Method 1** increases time consumption proportionally as the lyric length grows.
- **Method 2** is much less affected by lyric length and remains more time-efficient.
  For short songs, Method 1 is preferable for quality, while Method 2 is better for longer songs due to its greater time efficiency.

3. Model B: Static Lyrics & Dynamic Song



Strengths	Weaknesses
Lyrics Quality Less restrictive, allowing for better lyrical flow	**Patterned Music** Music may become too predictable, reducing listener interest
One Word Per One Note Easier for children to learn and sing the song	**Extended Rest Periods** Long rests caused by shortage of lyrics can disrupt musical flow.
Harmonical Song Structure Each new word starts at the beginning of a bar, enhancing musical harmony	**Complexity in Synchronization** Aligning static lyrics with dynamic melodies is challenging, especially when creating multiple verses, which can compromise the musical quality.
Controlled Lyric Adaptability Offers precise control over lyrics, advantageous for educational content where specific themes or vocabulary need to be emphasized for learning purposes.	

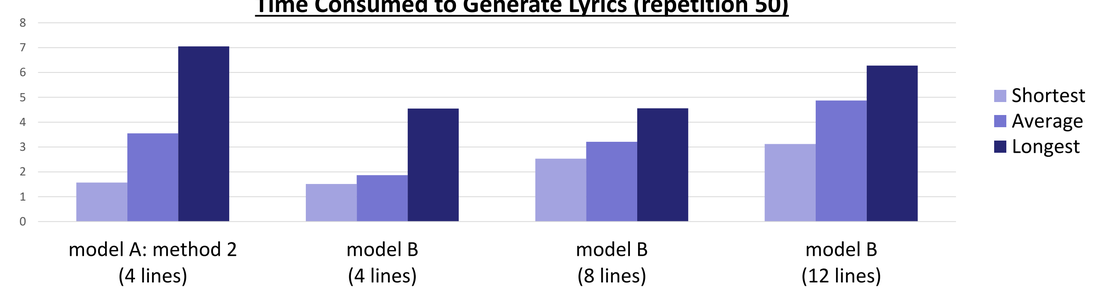
Potential

This model offers enhanced **flexibility in lyric composition**, enabling diverse thematic and stylistic exploration for educational purposes. This adaptability supports the creation of **contextually relevant educational content**, particularly suitable for **lyric-focused learning modules.**

- Process done by LLM
- Music Composition
- Lyrics Generation

YouTube Channel

Time Consumed to Generate Lyrics (repetition 50)



Observation & Select of method

Model B is the most time-efficient model and is minimally affected by the length of lyrics generated.

4. Enhanced Pronunciation Clarity through Phoneme Conversion

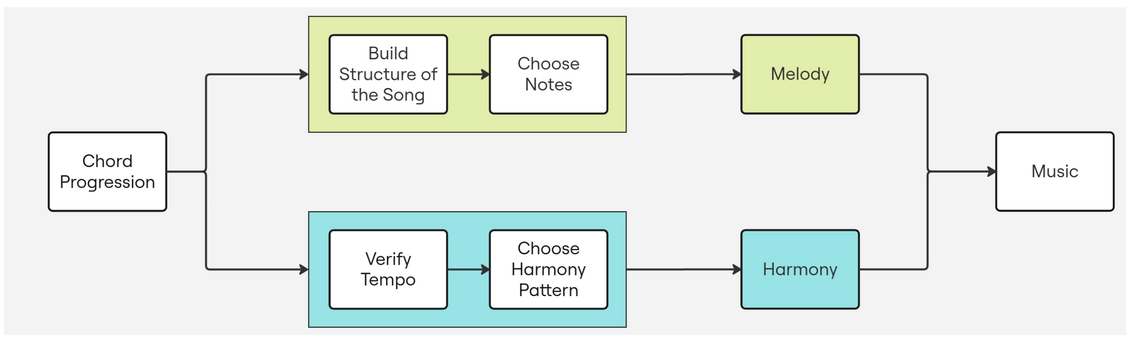
Observation

- **Issue**: Occasionally unclear pronunciation, which can hinder comprehension
- **Impact**: Particularly detrimental in educational contexts where precise pronunciation is crucial

Solution

- **Phoneme Conversion**: The software automatically converts lyrics into phonemes to ensure clarity and accuracy in pronunciation.

5. Melody and Harmony Music Composition Logic



	Melody	Harmony
Constraints	• Limits consecutive note gap to half an octave for ease of singing • Ensures no three consecutive notes are the same	• Weight distribution for harmony selection is adjusted based on tempo. • Slow tempo prioritizes lyrical harmonies and fast tempo prefers rhythmic backup

6. Further work

- ✓ **Develop a User-Friendly Interface**: Integrate a UST-SVS player into a front-end with a karaoke-like feature to enhance singing along for children.
- ✓ **Integration with Educational Robots**: Connect the software with Alpha-Mini robots for interactive use in classrooms, enhancing educational experiences.
- ✓ **Develop a Custom DiffSinger Voicebank**: Create DiffSinger voicebank to improve pronunciation and overall sound quality.
- ✓ **Enhance Music Quality with Markov Chains:** Investigate the use of Markov chains to potentially enhance music quality.

Acknowledgment

This study was supported by the Laidlaw Scholars Leadership and Research Program internship and conducted at the Computer-Human Interaction in Learning and Instruction (CHILI) lab, EPFL. I gratefully acknowledge their support.