Long-Form Text-to-Music Generation with Adaptive Prompts: A Case of Study in Tabletop Role-Playing Games Soundtracks

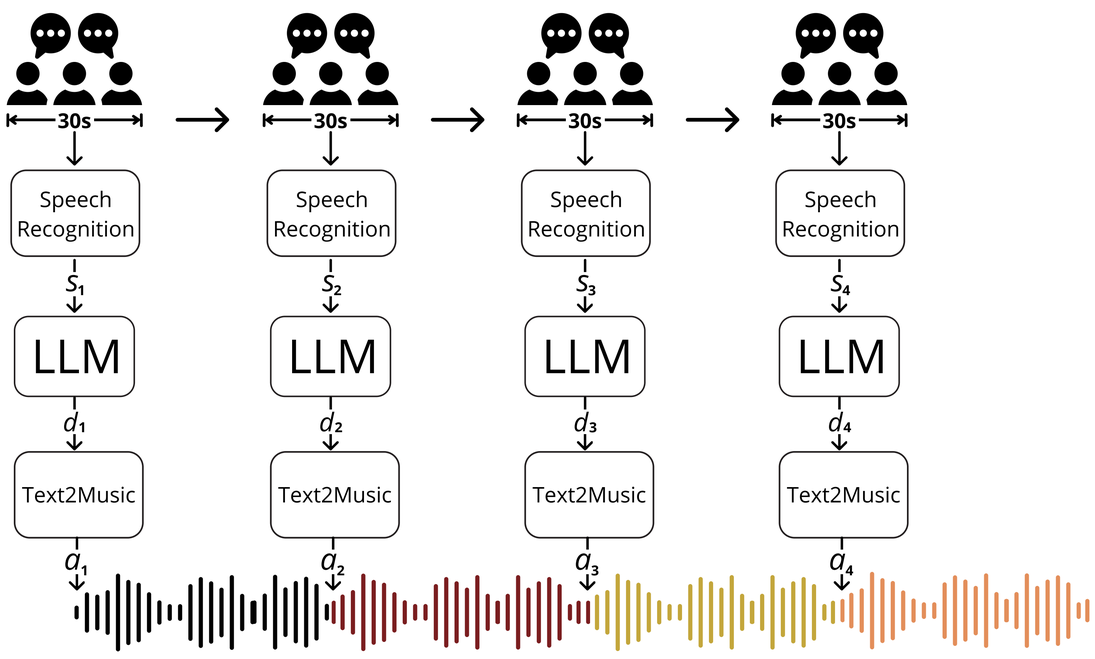
Felipe Marra, Lucas N. Ferreira
Universidade Federal de Viçosa, Minas Gerais, Brazil

Overview

We introduce a system called Babel Bardo, to investigate the capabilities of text-to-audio music generation models in producing long-form music with prompts that change over time.

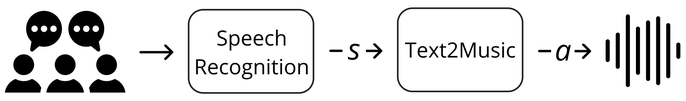
Babel Bardo

Babel Bardo listens to players speeches, leveraging an LLM to generate a music description that is fed into a text-to-music model.

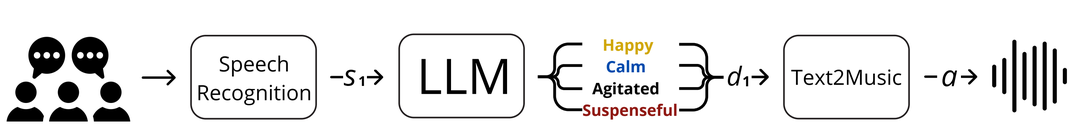


Methodology

Babel Bardo - Baseline (B)

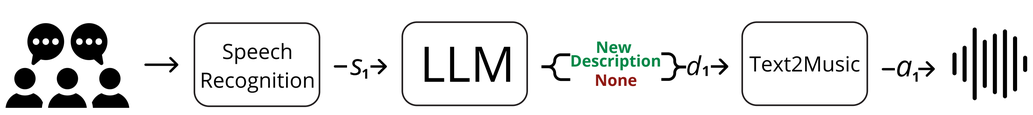


Babel Bardo - Emotion (E)



Babel Bardo - Description (D)

Babel Bardo - Description Continuation (DC)



Experiments

We evaluate the system in two Table-top RPGs:

- Call of the Wild (American English)
- O Segredo na Ilha (Brazilian Portuguese)

Audio Quality

TRPG	↓ FAD score				
	Babel Bardo				Human
	B	E	D	DC	
COTW	9.66	5.99	6.25	**5.82**	3.00
OSNI	9.55	6.11	5.63	**5.13**	4.18

Table 1. FAD for each Babel Bardo version in COTW and OSNI in contrast with Human music.

Story Aligment

TRPG	↓ Mean KL-Divergence			
	Babel Bardo			
	B	E	D	DC
COTW	4.84±2.98	**3.34±1.89**	4.26±2.65	4.23±2.51
OSNI	5.65±3.23	**4.16±2.12**	4.85±2.62	4.96±2.86
	↓ Mean Transition KLDs			
COTW	2.33±2.1	**1.33±1.19**	2.41±2.27	2.19±1.93
OSNI	2.11±1.65	**1.37±1.05**	1.88±1.47	2.09±1.71

Table 2. Mean/Standard Deviation of transition KLD for each Babel Bardo version in both COTW and OSNI.

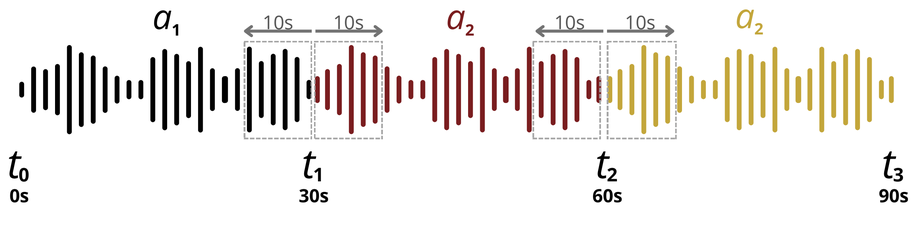


Figure 2. The transition KLD is computed between the 10 seconds before and after every transition moment t_i.

Results

While detailed music descriptions help improve audio quality, consistency across consecutive descriptions is important for smoother transitions. Moreover, emotion is a strong signal for generating soundtracks for RPGs.