Guiding LLMs by speech melody

David Porteš

Prompt:

Generate a catchy slogan for a car company.

Audio input:

The pre-crafted melody of the slogan

Motivation: Prosody in speech translation

- Prosody – Intonation, stress and rhythm of speech
- Two main approaches for ST:
  - Cascade – transcribe, translate, synthesize

    + Lots of parallel text data available

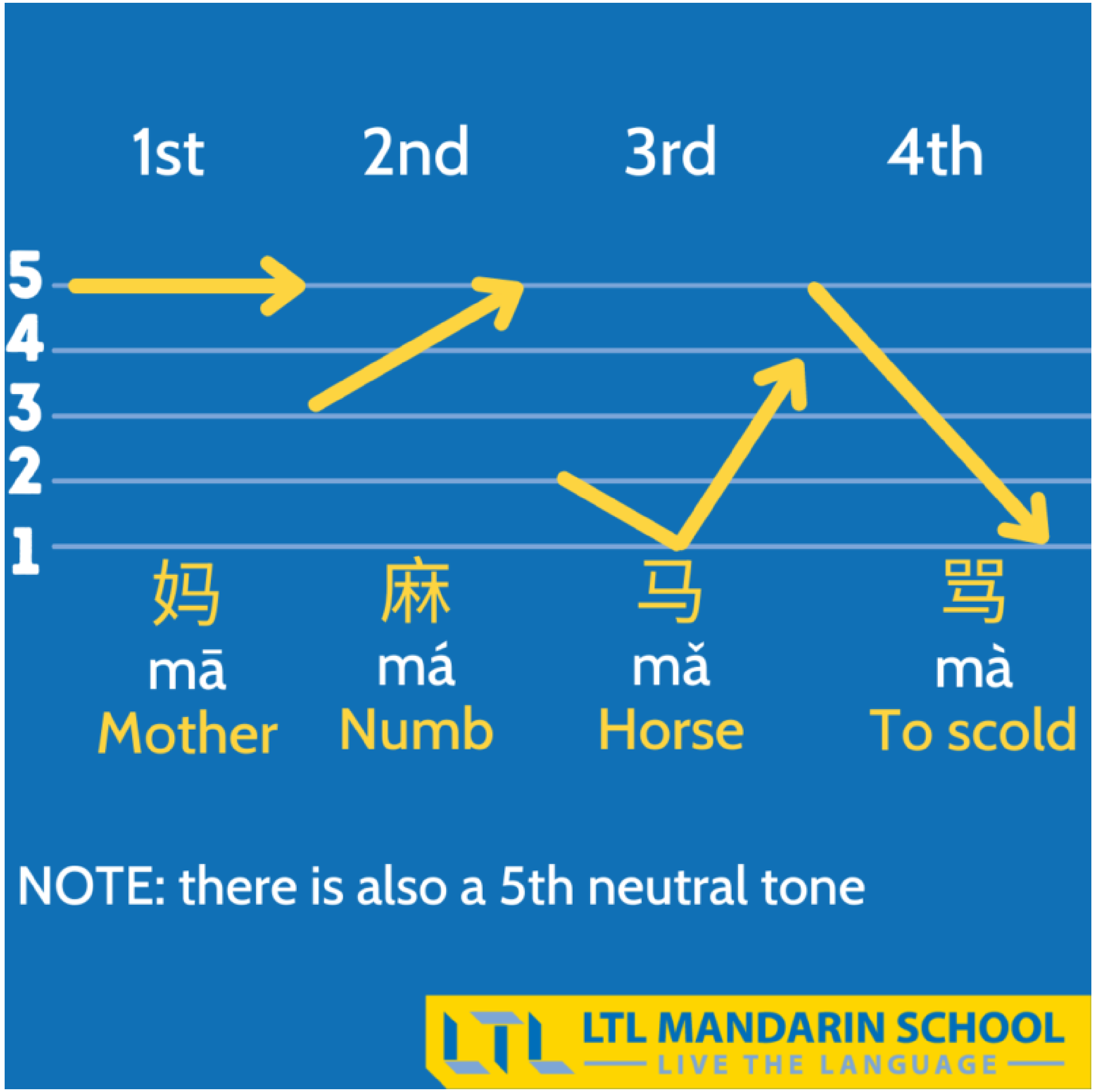
    - Discards all prosodic information from the audio
  - End-to-end – use parallel speech-to-speech data

    + Translation can be influenced by prosodic features

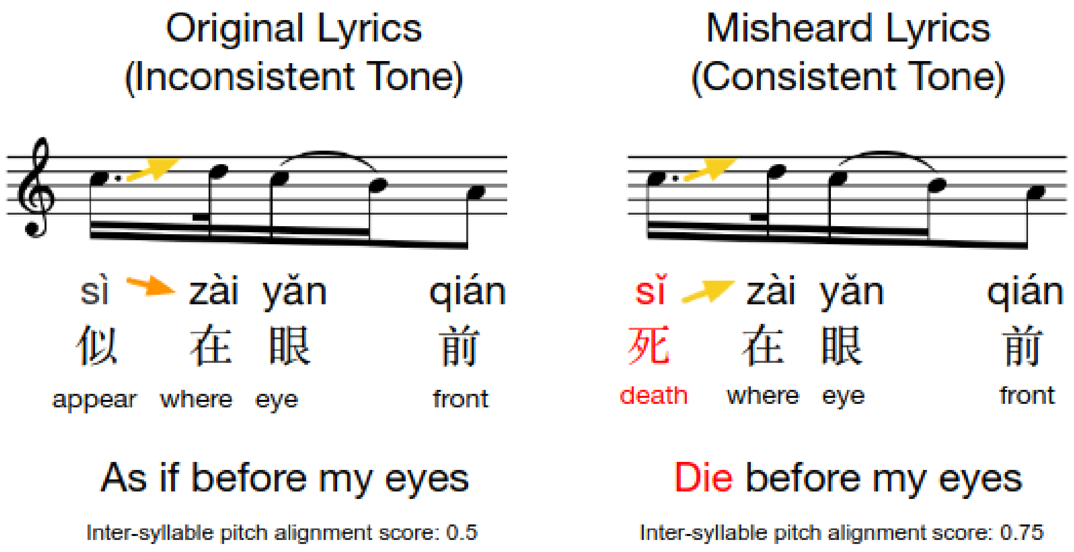
    - Data scarcity
- Can we get the best of both worlds?
- If only we could use prosodic features to guide **text-to-text** translation...

Inspiration: Song translation into tonal languages

- In tonal languages, intonation influences meaning
- If song lyrics do not match the melody, misunderstandings happen



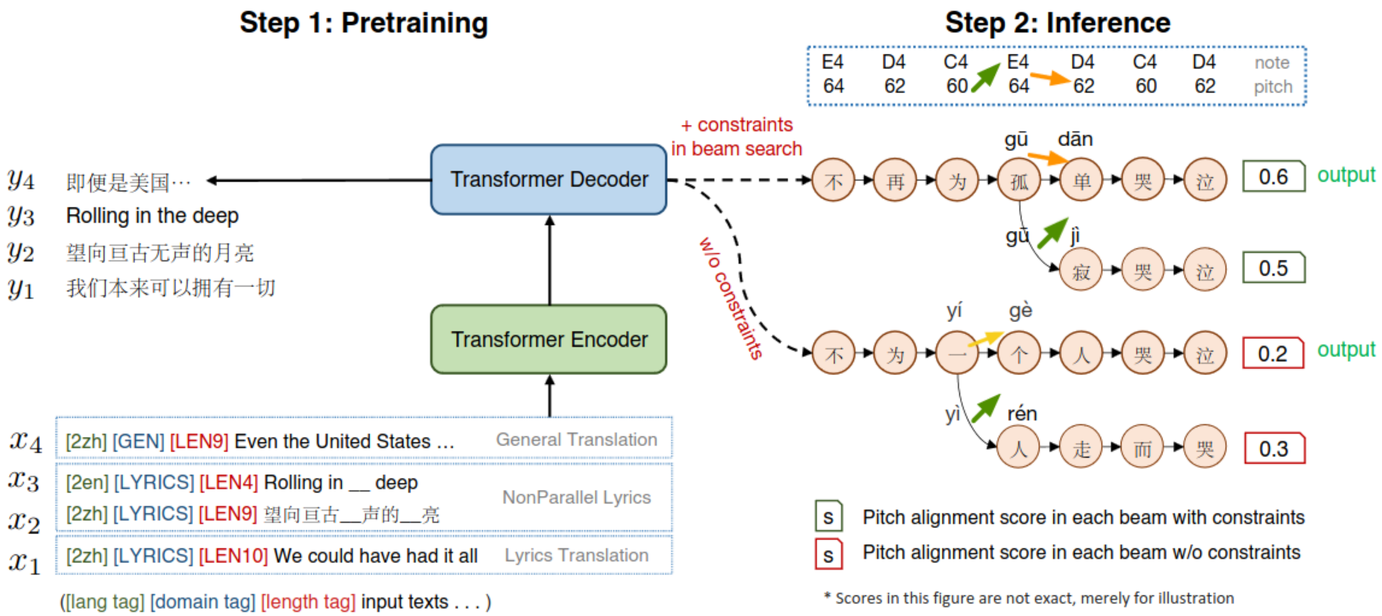
Lyrics-melody mismatch example



Automatic Song Translation for Tonal Languages (Guo et al. 2022)

Inspiration: GagaST song translator

- GagaST song translator
  - Automatic Song Translation for Tonal Languages (Guo et al. 2022)
- Rescores each token during the decoding phase
- Aligns translated lyrics to song melody


Step 1: Pretraining
Step 2: Inference
E4 64
D4 62
C4 60
E4 64
D4 62
C4 60
D4 62
note pitch
y_4 即便是美国…
y_3 Rolling in the deep
y_2 望向亘古无声的月亮
y_1 我们本来可以拥有一切
Transformer Decoder
Transformer Encoder
x_4 [2zh] [GEN] [LEN9] Even the United States … General Translation
x_3 [2en] [LYRICS] [LEN4] Rolling in __ deep
x_2 [2zh] [LYRICS] [LEN9] 望向亘古__声的__亮
NonParallel Lyrics
x_1 [2zh] [LYRICS] [LEN10] We could have had it all Lyrics Translation
([lang tag] [domain tag] [length tag] input texts . . .)
+ constraints in beam search
w/o constraints
gū dān
不 再 为 孤 单 哭 泣
0.6 output
gū jì
寂 哭 泣
0.5
yí gè
不 为 一 个 人 哭 泣
0.2 output
yì rén
人 走 而 哭
0.3
s Pitch alignment score in each beam with constraints
s Pitch alignment score in each beam w/o constraints
* Scores in this figure are not exact, merely for illustration

Automatic Song Translation for Tonal Languages (Guo et al, 2022)

Adapting GagaST

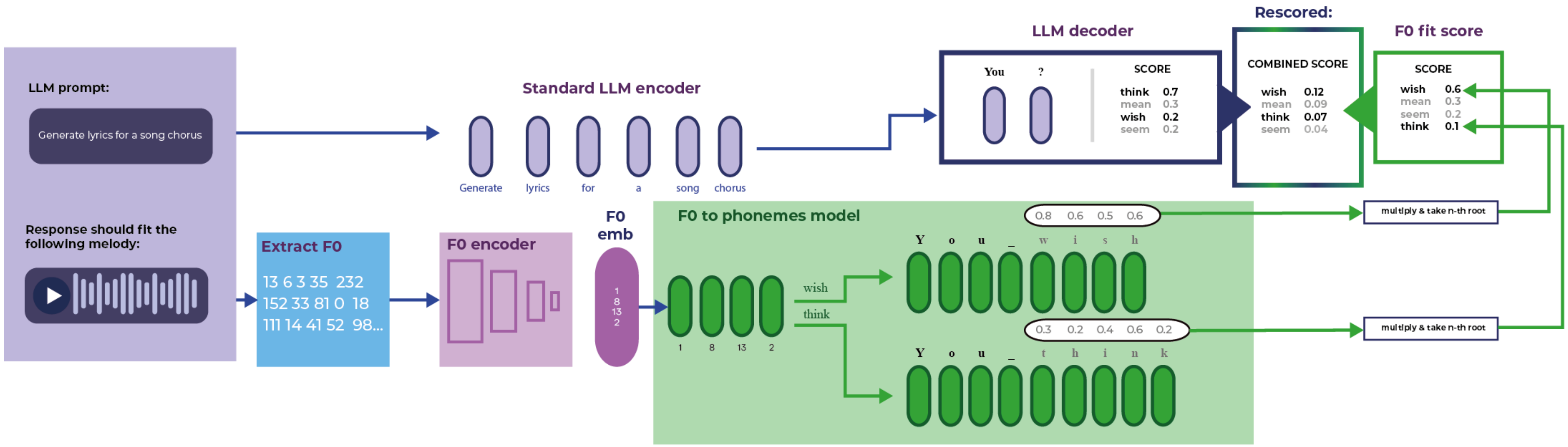
- GagaST uses prosodic features (intonation) to guide text-to-text translation of song lyrics

- Let's adapt it to speech translation!
- The original audio could be the 'song', and the translation could be the 'lyrics'

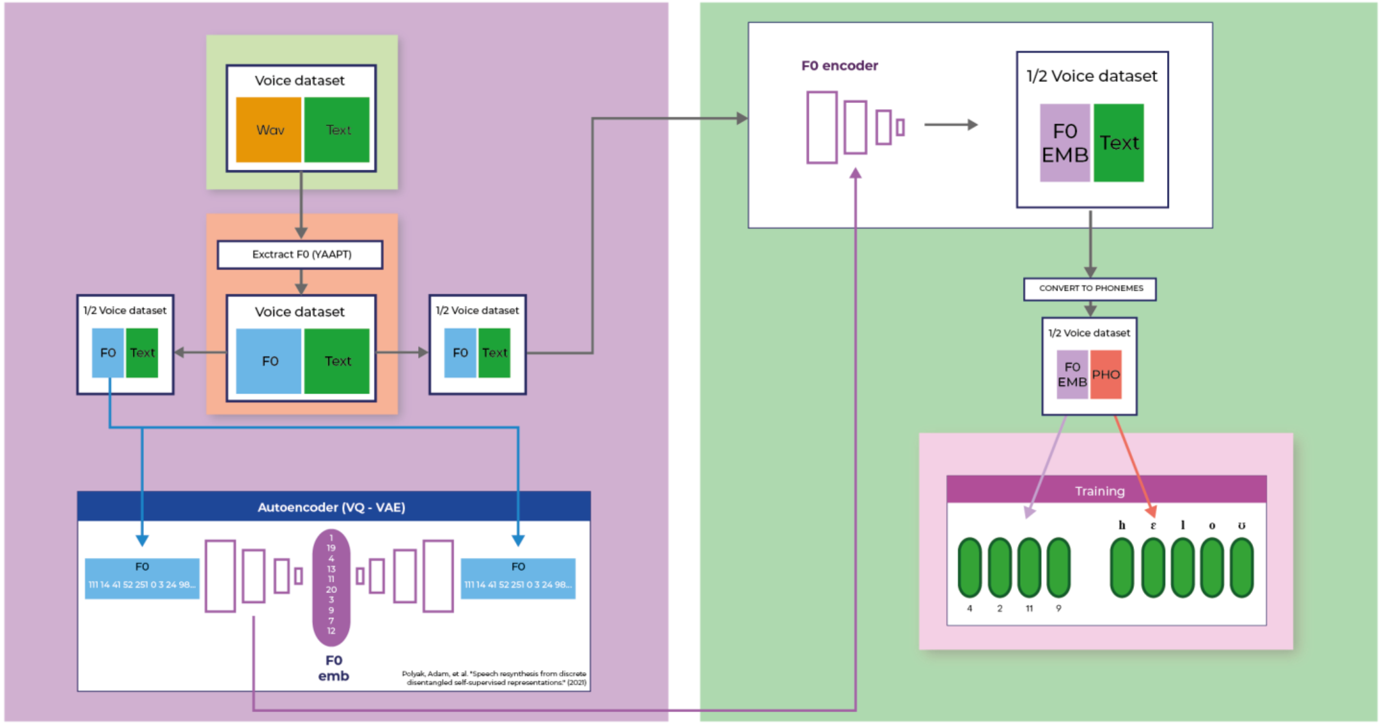
- But it uses the musical score of the song...
- And it uses Mandarin, where tone-word relationship falls into 4 pre-defined categories...

Introducing Meligner (WIP)

- We propose MElody aLIGNER (MELIGNER)
  - Instead of a musical score
    - Uses the YAAPT algorithm to extract melody (F0 curve) from source audio
    - Converts it into vector embeddings by VQ-VAE
  - Instead of the 4 tones of Mandarin
    - Uses a seq2seq model trained on parallel melody-text data
    - Learns the tone-word correspondence in the target language
  - Rescores each token during the LLM's decoding phase



Models' training

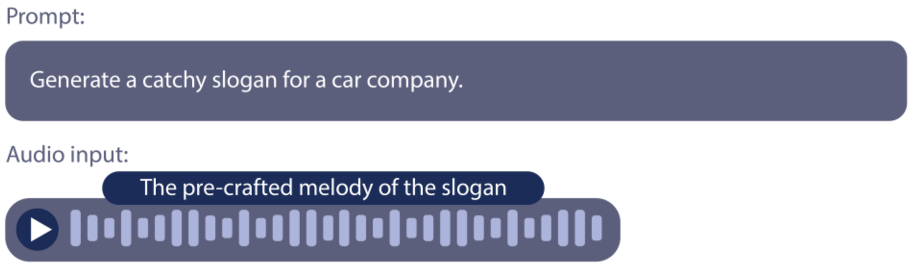


Polyak, Adam, et al. "Speech resynthesis from discrete disentangled self-supervised representations." (2021)

- We use two trainable models
  - VQ-VAE autoencoder
  - seq2seq model
    - Standard transformer architecture in our experiments

Why stop at speech translation?

- The ability to prescribe a target melody can be useful in other tasks
- We can use the LLM prompt for all kinds of tasks
  - Generate slogans
  - Do Automatic Speech Recognition (ASR)

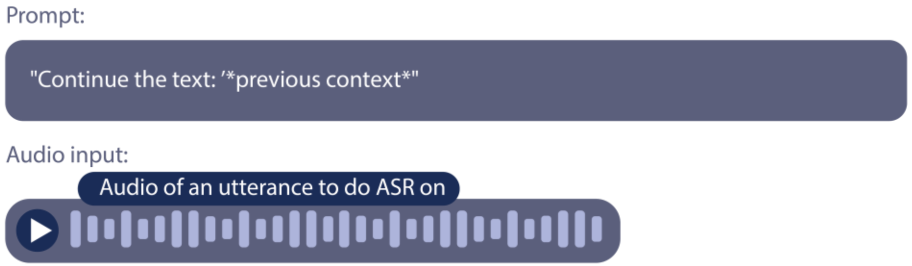


Prompt:

Generate a catchy slogan for a car company.

Audio input:

The pre-crafted melody of the slogan



Prompt:

"Continue the text: ’*previous context*"

Audio input:

Audio of an utterance to do ASR on

A new NLP task: Melodic Alignment

- Goal: Make LLM generated text fit a custom speech melody
- Evaluation metric: ASR- based 'Rank-shift score'
  - Fix an LLM
  - Take a speech-transcription pair
  - Let the LLM generate text without any prompt
  - Use the speech recording to rescore LLM outputs at each decoding stage
    - At each stage, observe the rank of the 'correct' token before and after rescoring
    - (the correct token is the token in the transcription at the current decoding position)
  - Report the average rank-shift – the difference in rank before and after rescoring, averaged over all stages

Conclusion

- A novel task: Melody Alignment
- Originally motivated for prosody-driven speech translation
    - But can be useful for other tasks, depending on the LLM prompt
- Defined as a wrapper over LLMs
    - Won't be made obsolete by LLM improvement
- A first-shot architecture at solving the task: Meligner
    - Reads in the melody in form of vector embeddings
    - Uses a melody-text seq2seq model to rescore LLM outputs

Future work

- Implement and evaluate Meligner using the rank-shift metric
- Evaluate on speech-translation benchmarks
- Also evaluate on other tasks, such as low-quality sound ASR or song lyrics generation
- Use the same approach for other prosodic features (stress, rhythm)

Thank you