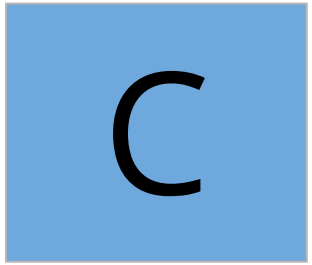
Speech-to-Singing

Group 17

The Team (Group 17)



Kangning Chen
Background: M.S. Computer Science

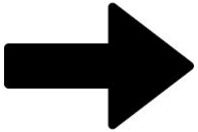
Ziyi Liu
Background: M.S. Computer Science

Yinghao Ma
Background: M.S. Music & Tech

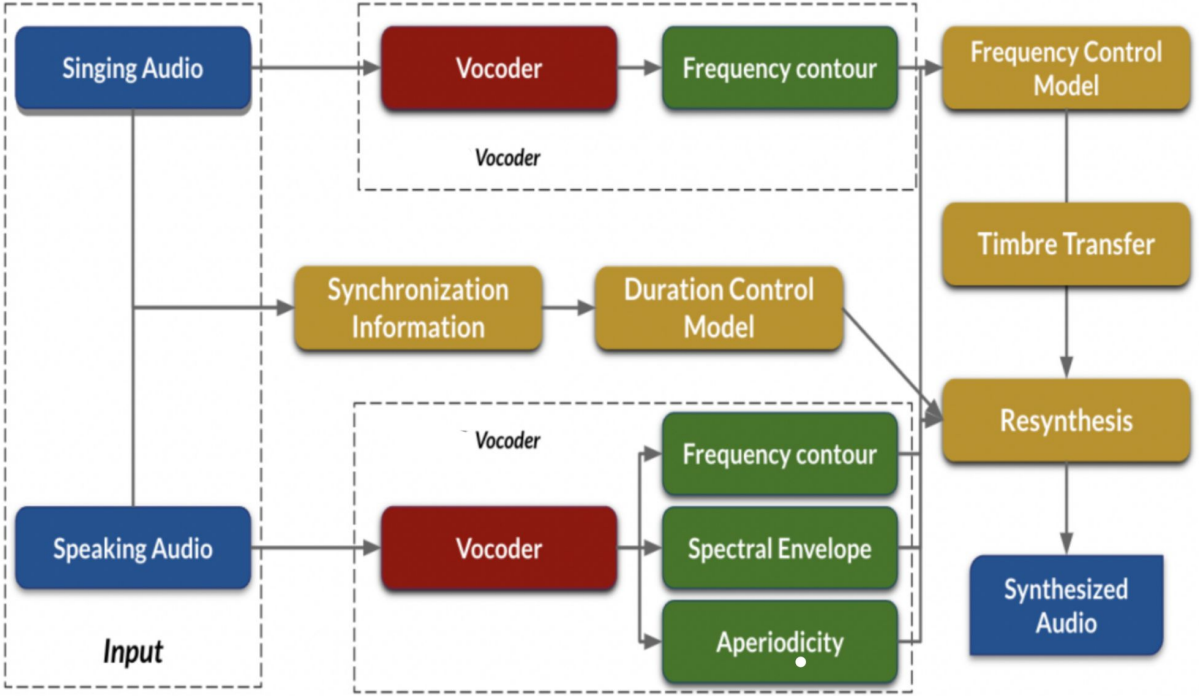
Ruibin Yuan
Background: M.S. Music & Tech

Problem Statement: speech-to-singing (STS)

- Personal specific: preserve content and partial style
- Difficulty: few-shot on testing data
- Preventing from out of tune



Traditional Methods

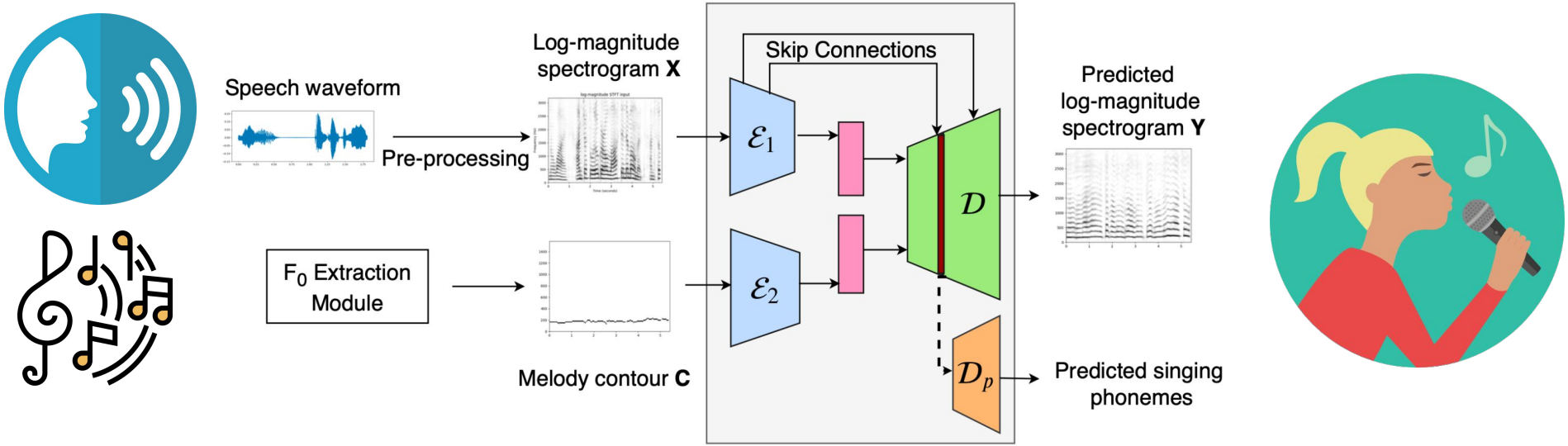


End-to-end Methods

- Deep learning architecture on (audio) style transfer
  - U-Nets, cycle-GAN etc.
- Strategy on the usage of singing input
  - Combine with acoustics features (F0, durations)
  - Using ground truth singing recording as a template (totally end2end)

Baseline

Baseline uses an encoder-decoder model:



Dataset

NUS-48E Sung and Spoken Corpus, which is a 169-min collection of audio recordings: sung and spoken lyrics of 12 subjects, 48 (20 unique) English songs, fully annotated at the phone-level (25,474 phone instances.)

Lyrics:
I'm sitting here in a boring room
It's just another rainy Sunday afternoon
I'm wasting my time I got nothing to do

Singing

Reading

Transcript

https://smcnus.comp.nus.edu.sg/nus-48e-sung-and-spoken-lyrics-corpus/

Results

	Original Speech	Predicted Singing	True Singing
Sample 1			
Sample 2			
Sample 3			
Sample 4			

System	LSD (dB) ↓	RCA ↑
Baseline 1 (B1)	14.19	0.221
Baseline 2 (B2)	11.71	0.769
AllNorm	11.72	0.771
Proposed MSE (P–MSE)	11.22	0.829
Proposed MTL (P–MTL)	**10.97**	**0.857**
Proposed MSE + PhSync	10.91	0.833
Singing Autoencoder	5.51	0.991

LSD (log-spectral distance)
- measures phone intelligibility
- Lower the better

RCA (F0 raw chroma accuracy)
- Evaluate melody transfer
- Higher the better

Ideas for Improvement (Project 4)

- Existing end2end systems failed to converge on this low resource task
- To deal with domain specific low resource task:
  - Expert knowledge + non-e2e systems
  - E2e pre-training with general knowledge + finetune on downstream tasks
- My thought: Data > Model
- We propose to adapt ”pre-train & finetune” paradigm to this task:
  - Self supervised pre-training with:
    - Large speech data
    - Medium size singing data
  - Finetune with:
    - Speech2singing data

Potential problems that we have identified in the baseline:

- Too many parameters(U-Net) with too small data
  a. -> Pretrain & Finetune
- Bad disentanglement between linguistic content & F0, easily overfit
  a. -> Disentangled representation learning, e.g. bottleneck, MMI
- MSE is not good enough for reconstruction
  a. -> Perceptual loss
- Model needs GT F0 as input, but what about inference?
  a. -> MIDI to F0 Prediction module
- The encoding of F0 is too sparse
  a. -> Dense representation / one-hot with bins
- Explicit duration control?
- Better objective metrics?