# Speech-to-Singing Conversion in an Encoder-Decoder Framework

Jayneel Parekh[⋆†] Preeti Rao[†] Yi-Hsuan Yang[§]

[⋆] Télécom Paris [†] IIT Bombay [§] Academia Sinica

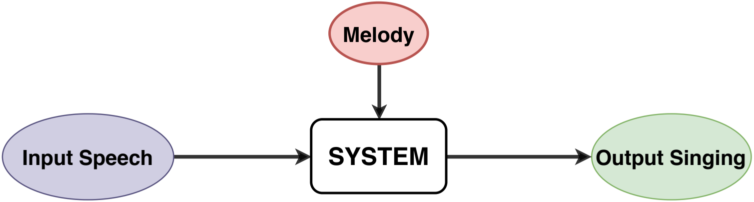
April, 2020

**ICASSP 2020**
**Audio and Acoustic Signal Processing**

# Outline

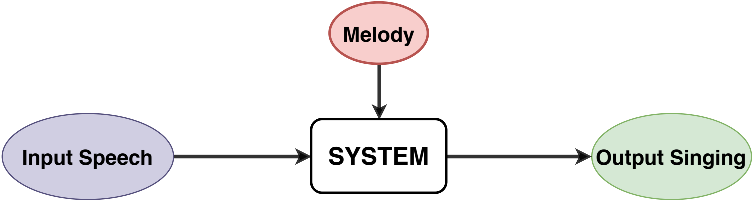
# Problem Formulation

**Aim**: To design a system that transforms speech into a song based on a given melody.



# Problem Formulation

**Aim**: To design a system that transforms speech into a song based on a given melody.




**Desired characteristics of output**

- Preserve speaker's timbre
- Preserve speech intelligibility with plausible phoneme durations.
- Follow the given melody

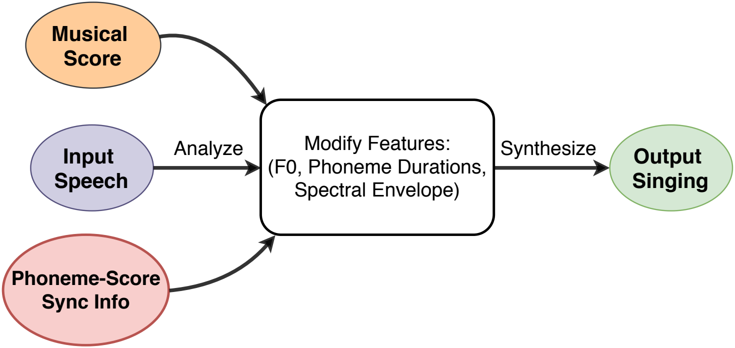
# Motivation & Applications


Songify

Music production

# Outline

# Model-based STS [Saitou et al., 2007]

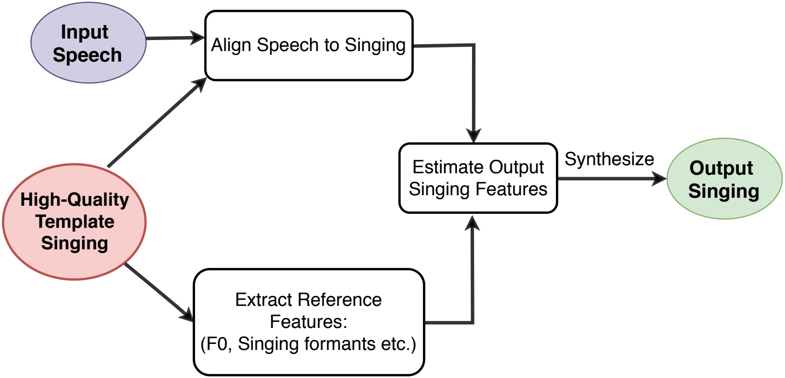



**Musical score**: Sequence of musical notes (pitch and duration).

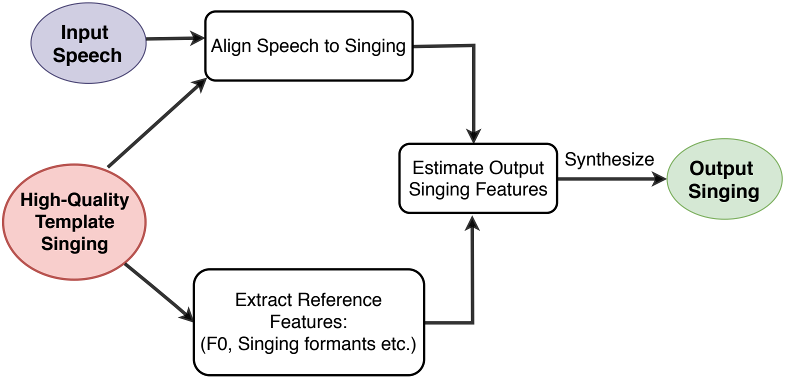
---

**Phoneme-Score Sync Info**: Association of each phoneme in speech with a musical note in the score.

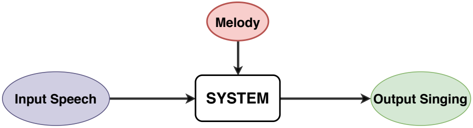
# Template-based STS [Cen et al., 2012], [Gao et al., 2019]



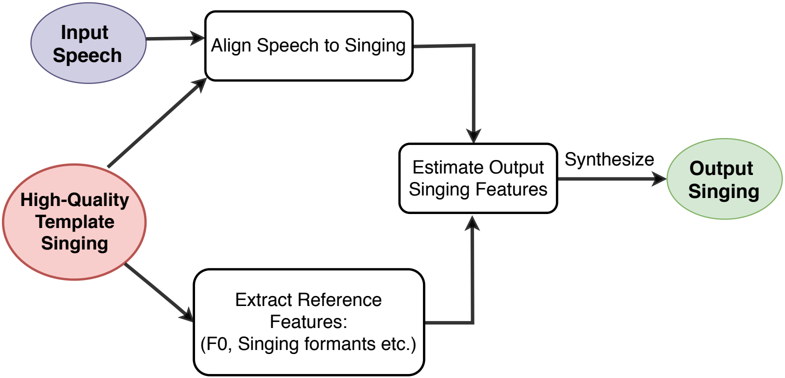
# Template-based STS [Cen et al., 2012], [Gao et al., 2019]




**Key difference in our formulation**: Minimal input information.



# Template-based STS [Cen et al., 2012], [Gao et al., 2019]

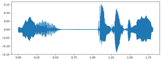



**Key difference in our formulation**: Minimal input information. Use Melody + Input Speech. We do not require singing templates or synchronization information. First to attempt such a transformation!

# Outline

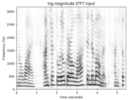
# System Overview

Speech waveform

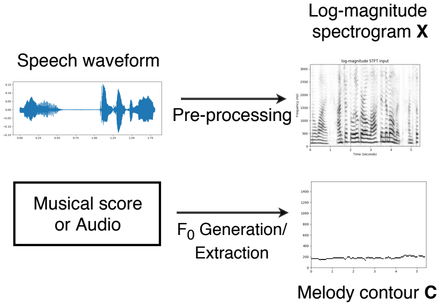


→ Pre-processing

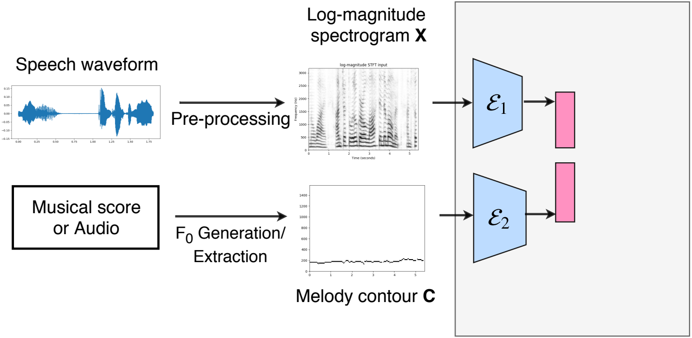
Log-magnitude spectrogram **X**



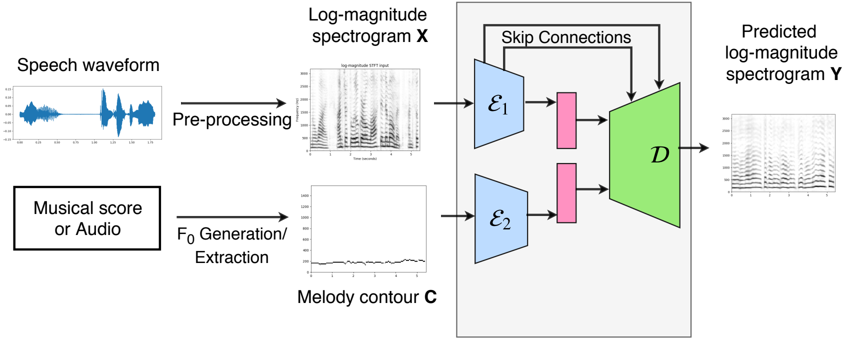
# System Overview



# System Overview

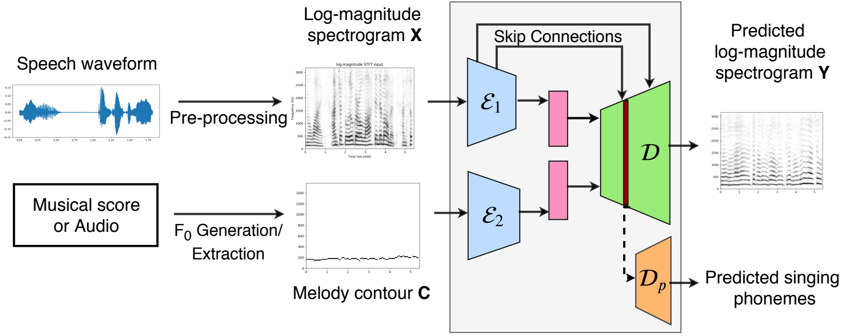


# System Overview




$$\mathbf{Y} = \mathcal{D}(\mathcal{E}_1(\mathbf{X}), \mathcal{E}_2(\mathbf{C}))$$

# System Overview

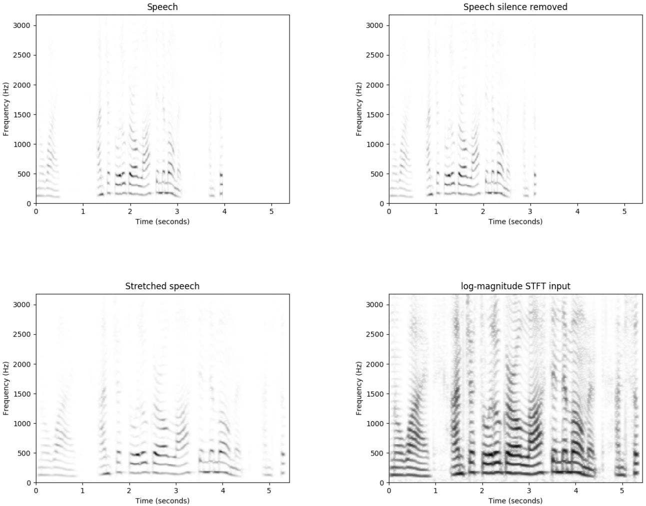



$$\mathbf{Y} = \mathcal{D}(\mathcal{E}_1(\mathbf{X}), \mathcal{E}_2(\mathbf{C}))$$

$\mathcal{D}_\mathbf{p}$ for Multi Task Learning (MTL) based objective

# Input Pre-processing

- Silent-frame removal
- Time stretching to singing length (Phase Vocoder)
- $\log(1 + x)$ transformation on magnitude spectrogram



# Network Architecture

- Adaptation of encoder-decoder network based on U-net [Ronneberger et al., 2015]
- Fully convolutional architecture with 1D convolutions.
- Skip connections between encoder $\mathcal{E}_1$ and decoder $\mathcal{D}$. Use of Instance Normalization (IN) layers before recurrent layers
- Encourage viewers to look at detailed architecture for each sub-component on our companion website.
  https://jayneelparekh.github.io/icassp20/

# Training

1. **Loss**
   - MSE on predicted and true log-magnitude spectrograms
   - Cross entropy loss for phoneme decoder (for frame $t$: $c_t$ – true phoneme, $\hat{y}_t^p$ – predicted phoneme probability distribution)

$$\mathcal{L}_{\text{MTL}} = \mathcal{L}_{\text{MSE}}(\mathbf{Y}, \hat{\mathbf{Y}}) + \frac{\lambda}{T} \sum_{t=1}^{T} \mathcal{L}_{\text{CE}}(\hat{y}_t^p, c_t)\,,$$

$$\mathcal{L}_{\text{MSE}}(\mathbf{Y}, \hat{\mathbf{Y}}) = ||\mathbf{Y} - \mathcal{D}(\mathcal{E}_1(\mathbf{X}), \mathcal{E}_2(\mathbf{C}))||^2\,,$$

$$\mathcal{L}_{\text{CE}}(\hat{y}_t^p, c_t) = -\hat{y}_t^p(c_t) + \log\Big(\sum_{m \in P} \exp(\hat{y}_t^p(m))\Big)\,.$$

# Training

## 1. Loss

- MSE on predicted and true log-magnitude spectrograms
- Cross entropy loss for phoneme decoder (for frame $t$: $c_t$ – true phoneme, $\hat{y}_t^p$ – predicted phoneme probability distribution)

$$\mathcal{L}_{\text{MTL}} = \mathcal{L}_{\text{MSE}}(\mathbf{Y}, \hat{\mathbf{Y}}) + \frac{\lambda}{T} \sum_{t=1}^{T} \mathcal{L}_{\text{CE}}(\hat{y}_t^p, c_t)\,,$$

$$\mathcal{L}_{\text{MSE}}(\mathbf{Y}, \hat{\mathbf{Y}}) = ||\mathbf{Y} - \mathcal{D}(\mathcal{E}_1(\mathbf{X}), \mathcal{E}_2(\mathbf{C}))||^2\,,$$

$$\mathcal{L}_{\text{CE}}(\hat{y}_t^p, c_t) = -\hat{y}_t^p(c_t) + \log\Big(\sum_{m \in P} \exp(\hat{y}_t^p(m))\Big)\,.$$

## 2. Data augmentation

- Augmenting the training data to 2 times.
- Pitch-shifting input speech – target singing unchanged.
- Amount of pitch-shift sampled uniformly at random from $[-1, 1]$ semi-tones.

# Prediction Strategy

Network output (Log-magnitude spectrogram) $\rightarrow$ Time-domain signal

- Get magnitude spectrogram via element-wise transformation $f(x) = e^x - 1$
- Phase estimation using *Griffin-Lim*
- Modification [Wang et al., 2017]: Raise magnitude spectrogram to power 1.2 before *Griffin-Lim*

# Outline

# Data Generation

**NUS Sung and Spoken Lyrics Corpus** [Duan et al., 2013]

- 48 recordings for 20 unique songs, 12 subjects, each subject sings & reads 4 songs.

# Data Generation

**NUS Sung and Spoken Lyrics Corpus** [Duan et al., 2013]

- 48 recordings for 20 unique songs, 12 subjects, each subject sings & reads 4 songs.
- Each unique song covered by 2 or 4 singers. 19 songs for training, 1 for testing (2 recordings).

# Data Generation

**NUS Sung and Spoken Lyrics Corpus** [Duan et al., 2013]

- 48 recordings for 20 unique songs, 12 subjects, each subject sings & reads 4 songs.
- Each unique song covered by 2 or 4 singers. 19 songs for training, 1 for testing (2 recordings).
- Phone-level annotation file: Start and end time of each phone.

# Data Generation

**NUS Sung and Spoken Lyrics Corpus** [Duan et al., 2013]

- 48 recordings for 20 unique songs, 12 subjects, each subject sings & reads 4 songs.
- Each unique song covered by 2 or 4 singers. 19 songs for training, 1 for testing (2 recordings).
- Phone-level annotation file: Start and end time of each phone.

**Input-Target Sample Generation**

- Extract segments to remove silences from singing
- Generate multiple combinations of consecutive words (3 – 20 words).
- Refer to paper for precise details

# Evaluated Systems

- Baseline 1 (B1): Proposed network, MSE loss, and no melody information.
- Baseline 2 (B2): No IN layers, skip connections, MSE loss.

# Evaluated Systems

- Baseline 1 (B1): Proposed network, MSE loss, and no melody information.
- Baseline 2 (B2): No IN layers, skip connections, MSE loss.
- Proposed MSE (P-MSE): Proposed network, MSE loss

# Evaluated Systems

- Baseline 1 (B1): Proposed network, MSE loss, and no melody information.
- Baseline 2 (B2): No IN layers, skip connections, MSE loss.
- Proposed MSE (P-MSE): Proposed network, MSE loss
- Proposed MTL (P-MTL): Proposed network, MTL loss

# Evaluated Systems

- Baseline 1 (B1): Proposed network, MSE loss, and no melody information.
- Baseline 2 (B2): No IN layers, skip connections, MSE loss.
- Proposed MSE (P-MSE): Proposed network, MSE loss
- Proposed MTL (P-MTL): Proposed network, MTL loss
- Singing Autoencoder

# Objective Evaluation

- Log-Spectral Distance (LSD): Average euclidean distance between true and predicted log-spectrogram frames over time, for frequencies between 100 Hz to 3.5 kHz.
- $F_0$ evaluation – Raw Chroma Accuracy (RCA): Determine how good is the model at preserving melody. Used RCA between predicted pitch contours of target and predicted singing.

# Objective Evaluation

- Log-Spectral Distance (LSD): Average euclidean distance between true and predicted log-spectrogram frames over time, for frequencies between 100 Hz to 3.5 kHz.

- $F_0$ evaluation – Raw Chroma Accuracy (RCA): Determine how good is the model at preserving melody. Used RCA between predicted pitch contours of target and predicted singing.

| System | LSD (dB) ↓ | RCA ↑ |
|---|---|---|
| Baseline 1 (B1) | 14.19 | 0.221 |
| Baseline 2 (B2) | 11.71 | 0.769 |
| Proposed MSE (P–MSE) | 11.22 | 0.829 |
| Proposed MTL (P–MTL) | **10.97** | **0.857** |
| Singing Autoencoder | 5.51 | 0.991 |

# Subjective Evaluation

- Preference test for subset of systems: B2, P–MSE, P–MTL.

# Subjective Evaluation

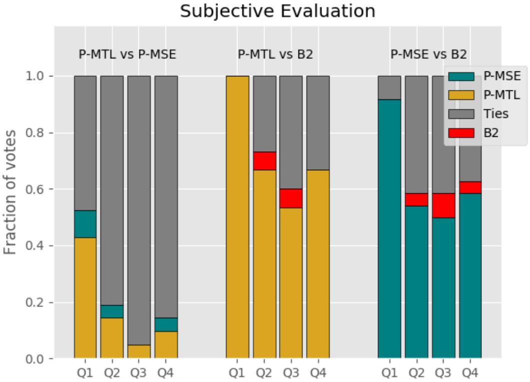
- Preference test for subset of systems: B2, P–MSE, P–MTL. Each participant (11 total) compared outputs of two random systems for 5 test samples

# Subjective Evaluation

- Preference test for subset of systems: B2, P–MSE, P–MTL. Each participant (11 total) compared outputs of two random systems for 5 test samples
- Lyrics/Phoneme intelligibility (Q1), Naturalness (Q2), Melodic similarity to target (Q3) and Speaker identifiability (Q4).

# Subjective Evaluation

- Preference test for subset of systems: B2, P–MSE, P–MTL. Each participant (11 total) compared outputs of two random systems for 5 test samples
- Lyrics/Phoneme intelligibility (Q1), Naturalness (Q2), Melodic similarity to target (Q3) and Speaker identifiability (Q4).



# Qualitative Observations

**Positives (for P-MTL)**

- Good Melody transfer
- Fair Naturalness
- Reasonable Phoneme duration modelling and intelligibility

**Limitations**

- Speaker identifiability
- Relatively small dataset, low generalizability

# Outline

# Conclusion

Key takeaways:

- Use only speech and melody for output singing. First to attempt such a transformation using a ML based method that does not use singing templates or synchronization information.
- Process the time-frequency representation via a deep neural network
- Multi Task Learning based objective to improve phoneme intelligibility
- Shows capability of transformation with significant room for improvement.

Code available on GitHub!
https://github.com/jayneelparekh/sp2si-code

Companion website
https://jayneelparekh.github.io/icassp20/

Code available on GitHub!
https://github.com/jayneelparekh/sp2si-code

Companion website
https://jayneelparekh.github.io/icassp20/

Thank you!