Our goal: a model that can characterize the dynamics of various singing behaviors in a sung melodic contour.

F0 contours of singing-voices include melody components and several types of fluctuations. Our model is designed to capture these complexities.

**F0 control model for singing voices**

- **Idea 1:** Modeling of target note sequence using Hidden Markov Model (HMM).
  
  \[ u_{T1} = \sum_{m=1}^{M} a_{Tm} y_{m,j-1} = \eta_{T1} \]
  
  Considering the error (fine fluctuations) between estimated target note sequence \( \tilde{u}_{T1} \) and \( u_{T1} \),
  
  \[ \tilde{u}_{T1} = u_{T1} - u_{T1} = \sum_{m=1}^{M} a_{Tm} y_{m,j-1} + \sum_{m=1}^{M} a_{Tm} y_{m,j-1} - u_{T1} \approx \mathcal{N}(\eta_{Tm}, \sigma^2) \]

- **Idea 2:** Modeling of fluctuation factors using all-pole model.

  Difference equation:
  
  \[ a_{T1} + \sum_{j=1}^{M} a_{Tm} y_{m,j-1} = u_{T1} \]

- **Idea 3:** Maximum likelihood estimation by iteration of following two steps:
  
  1. Step 1: Viterbi training.
  2. Step 2: LPC-like solver.

**Maximum likelihood parameter estimation**

- **Log likelihood**
  
  \[ \log P(Y^{(1)}, \ldots, Y^{(T)}) \text{ (\omega)} = \sum_{i=1}^{T} \log P(Y^{(i)}(\theta^{(i)}), \omega) \]
  
  where:
  
  \[ \omega = \{ \eta_{T1}, \ldots, \eta_{Tm} \} \]

- **Prior probability**
  
  \[ P(\theta) = P(\theta^{(1)}, \ldots, \theta^{(T)}) P(\theta^{(T)}) P(m_{S1}, \ldots, m_{Sj}) \]

- **Objective function**
  
  \[ J = \sum_{i=1}^{T} \log P(Y^{(i)}(\theta^{(i)}), \omega) + \log P(e_{1}, \ldots, e_{N}) \]

- **Schematic view for updates of parameters**

  1. Update of \( \theta^{(1)} \)
  2. Update of \( q^{(1)}, \ldots, q^{(N)} \)
  3. Update of \( m_{S1}, \ldots, m_{Sj} \)

**Experimental setup**

- **Singing voice database**
  
  - Singers: Male and female classical singers
  - Male and female pop songs
  - Amateur male and female singers
  - Songs: "Twinkle, twinkle, Little Star" and "Ode to Joy" sung with Japanese lyrics and hummed while listening to melody (guide tones) with headphones.
Estimation results

- Order $M$ of all-pole model: 3,
- Frame length $L$: 100ms

![Amplitude responses of all-pole models](image)

Estimation performance

- Correct rate based on RMS

\[
\text{RMS}_{u(t)} = \frac{\text{Frames where } \text{RMS}_{u(t)} < \xi \times 100 \%}{\text{Total frames}}
\]

\[
\text{RMS}_{\hat{y}(t)} = \frac{\text{Frames where } \text{RMS}_{\hat{y}(t)} < \xi \times 100 \%}{\text{Total frames}}
\]

- Evaluation of frame length $L$. Evaluation of order $M$ of impulse response ($M = 5$)

<table>
<thead>
<tr>
<th>$L$</th>
<th>$\text{Corr}_{u(t)}$</th>
<th>$\text{Corr}_{\hat{y}(t)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50ms</td>
<td>82.1</td>
<td>65.9</td>
</tr>
<tr>
<td>100ms</td>
<td>80.7</td>
<td>55.8</td>
</tr>
<tr>
<td>200ms</td>
<td>79.6</td>
<td>28.9</td>
</tr>
</tbody>
</table>

- Evaluation of order $M$ of impulse response ($L = 50\text{ms}$)

<table>
<thead>
<tr>
<th>$M$</th>
<th>$\text{Corr}_{u(t)}$</th>
<th>$\text{Corr}_{\hat{y}(t)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>82.2</td>
<td>66.5</td>
</tr>
<tr>
<td>5</td>
<td>82.1</td>
<td>65.9</td>
</tr>
<tr>
<td>10</td>
<td>81.2</td>
<td>52.5</td>
</tr>
</tbody>
</table>

Analysis of singer’s individuality

- Frequent distribution of resonant frequency in amplitude responses

![Relative frequency](image)

Demonstration

- Convolute singing style of professionals to singing voice of amateurs

![Proposed algorithm](image)

Conclusion

- Parameter estimation method of F0 control model for singing voices
  - Decomposed F0 contours of singing voices to note sequence and various fluctuations
  - Confirmed convergence of proposed algorithm and estimation performance based on RMS
- Future works
  - Optimization of frame length and order of impulse response
  - Evaluation of our model’s ability to automatically detect particular singing behaviors such as vibrato and overshoot