Speech to Singing Synthesis

Yufei Zhang, Yoon Mo Yang, Mingqing Yun

Department of Electrical and Computer Engineering, University of Rochester, NY, 14627

{yzh242, yyang106,myun5}@ur.rochester.edu

What to do?

Input

Our system

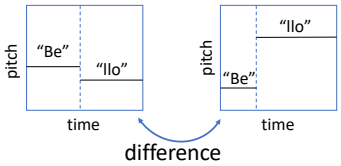
Output

How to do?

speaking

singing

1. Detect the pitch of each phoneme



2. Pitch shifting and tempo change

Pitch Detection

Yin algorithm

Step1: Autocorrelation

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau}$$

Step2: Difference Function

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2$$

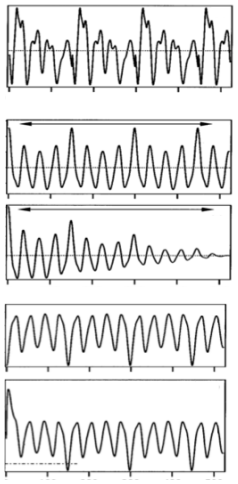
where τ is the lag time

Step3: Cumulative mean

$$d'_t(\tau) = \begin{cases} 1, & if\ \tau = 0 \\ \frac{d_t(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} d_t(j)}, & otherwise \end{cases}$$

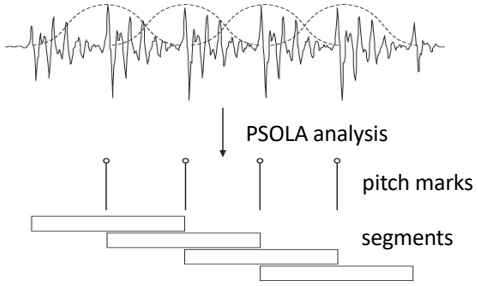
Step4: Absolute threshold

Step5: Parabolic interpolation



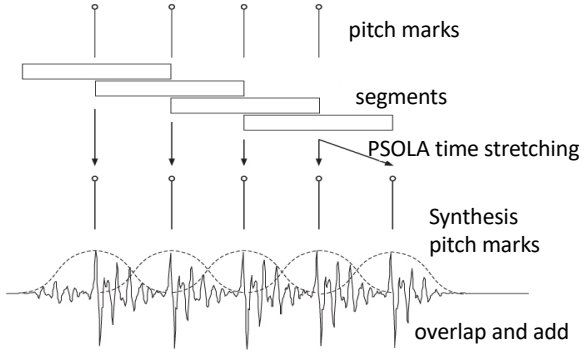
Pitch shifting and tempo change

PSOLA



(a) Determination of the pitch period P(t) of the input signal and of time instants (pitch marks) t_i

(b) Extraction of a segment centered at every pitch mark t_i by using a Hanning window with length $L_i = 2P(t_i)$



(a) Choice of the corresponding analysis segment i to minimize the time distance $|at_i - \tilde{t}_k|$

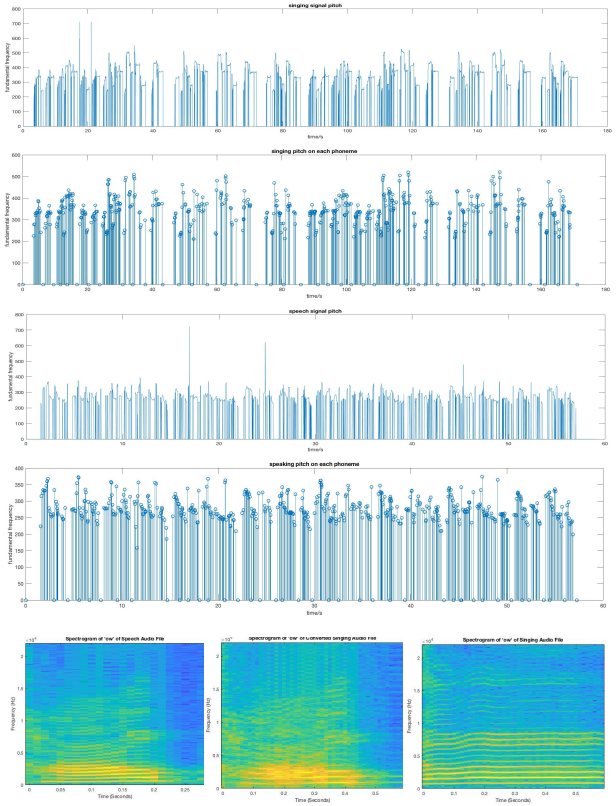
(b) Overlap and add the selected segment.

(c) Determination of the time instant where the next synthesis segment will be centered.

Data

NUS Sung and Spoken Lyrics Corpus.
Speaking lyrics , singing voice, and temp steps baseline of each phoneme.

Result



Listen to the demo!