Singing Expression Transfer from One Voice to Another for a Given Song

Korea Advanced Institute of Science and Technology

Sangeon Yong, Juhan Nam

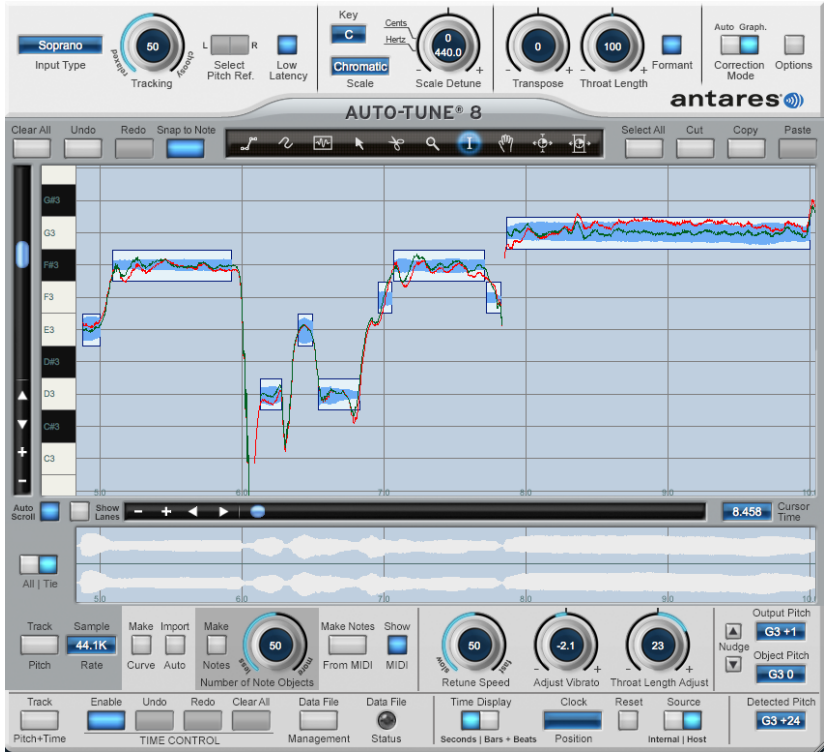
Introduction

Introduction

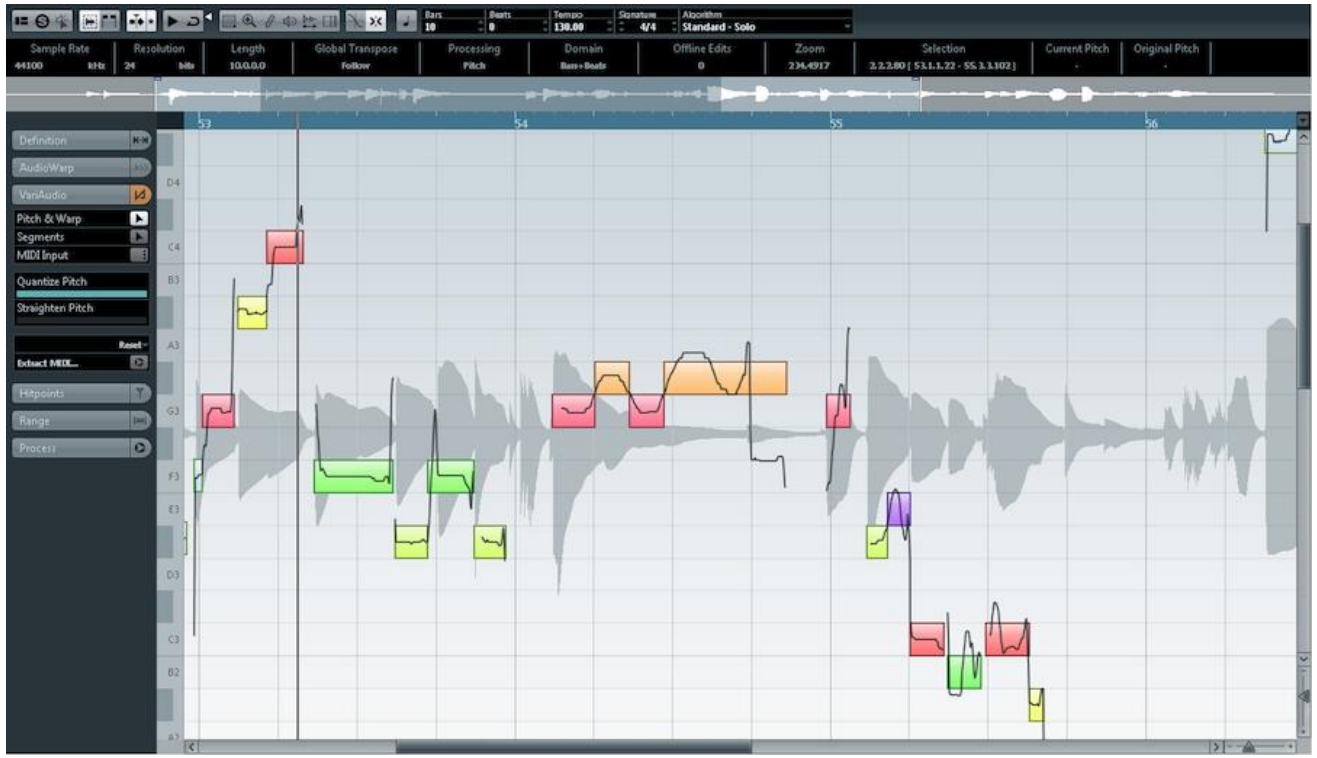
source

target

Related Works

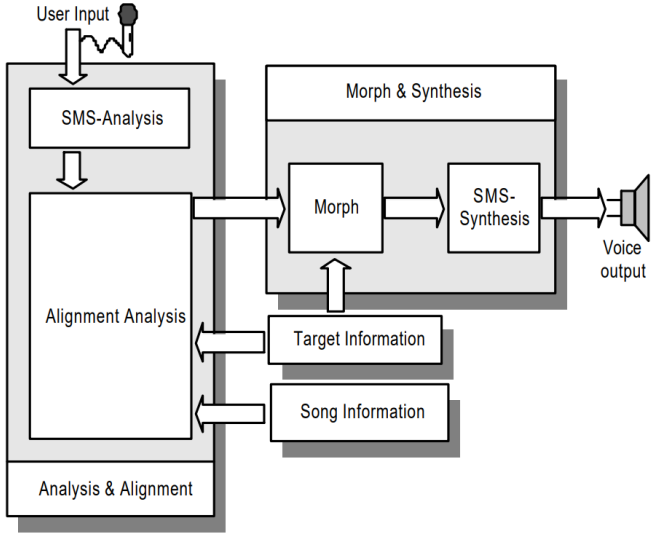


Antares Autotune 8 graphical mode



Steinberg Variaudio

Related Works



- Cano et al. (ICMC, 2000)
  - Voice morphing system with source and target voice
  - Score information is used for temporal alignment
- Nakano et al. (SMC, 2009)
  - Similar with above but using a singing synthesizer instead of the source voice (i.e. Vocaloid)
  - Tune synthesizer parameter with the lyric information of the song
- However, they require additional score information!

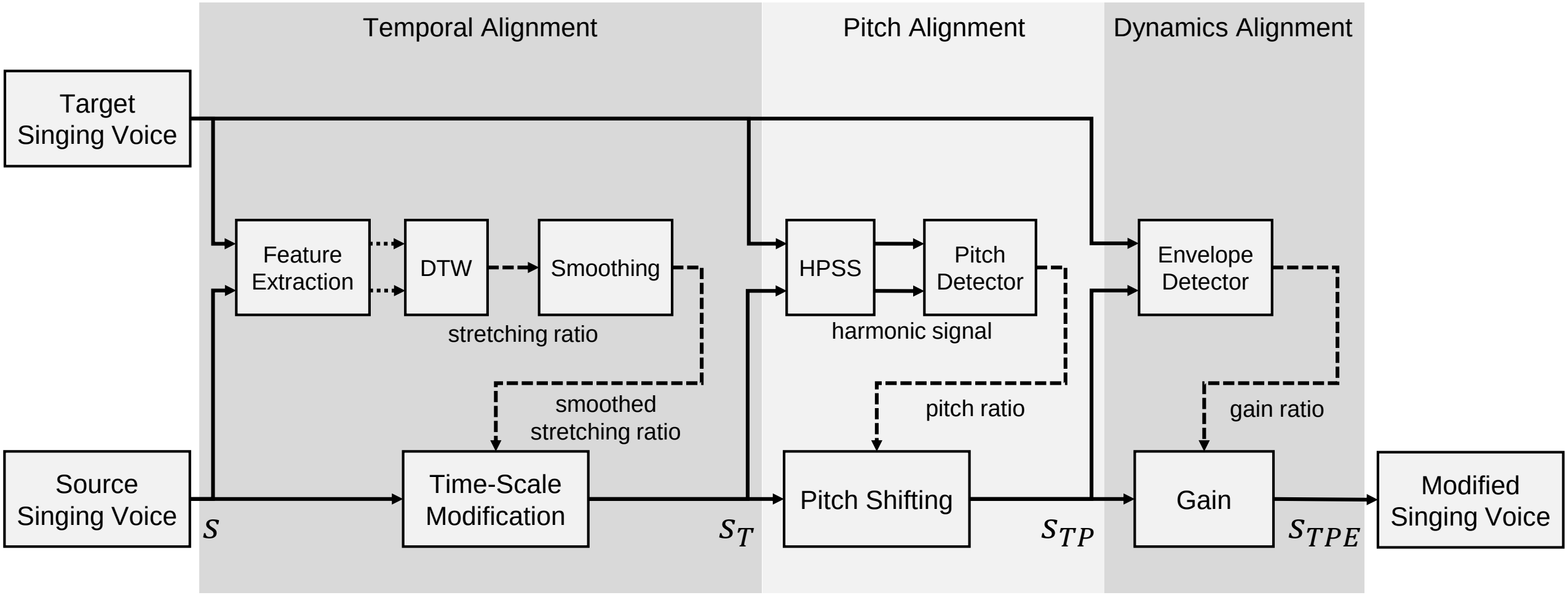
Research Goal

Voice color

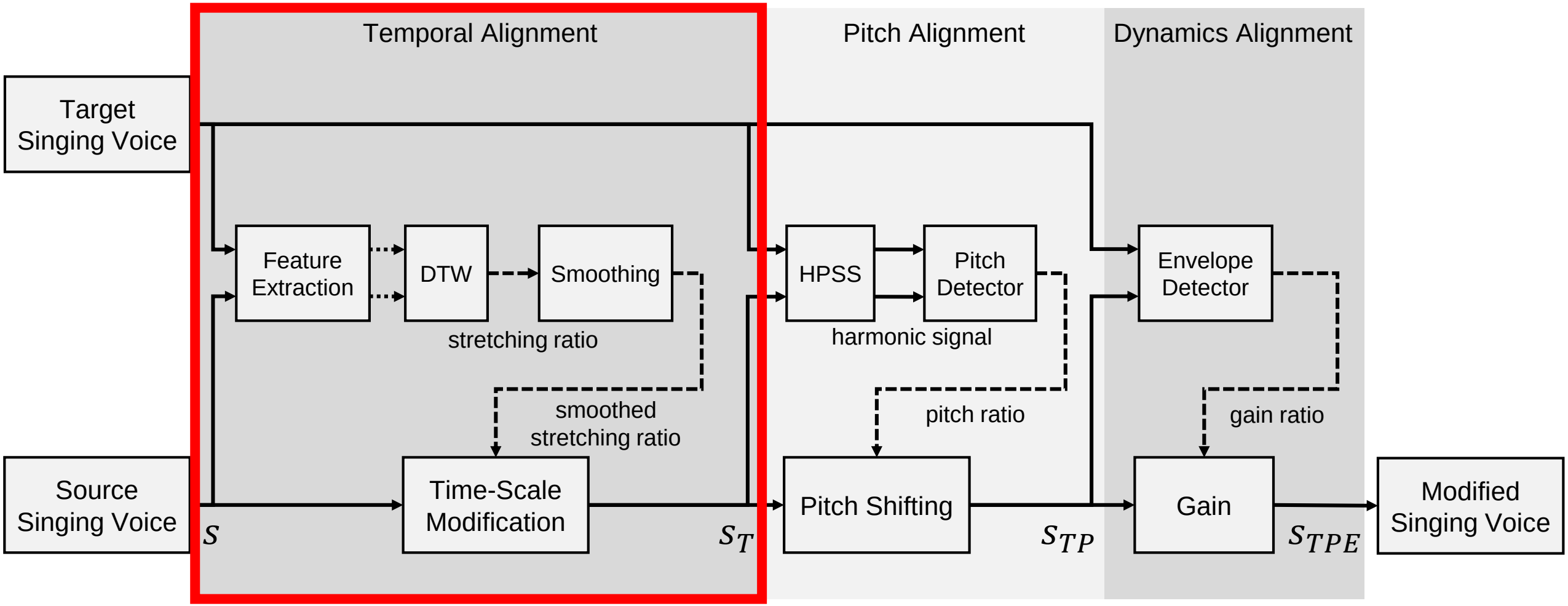
Rhythm, Pitch, Dynamics

Transfer musical expressions **without any additional information**

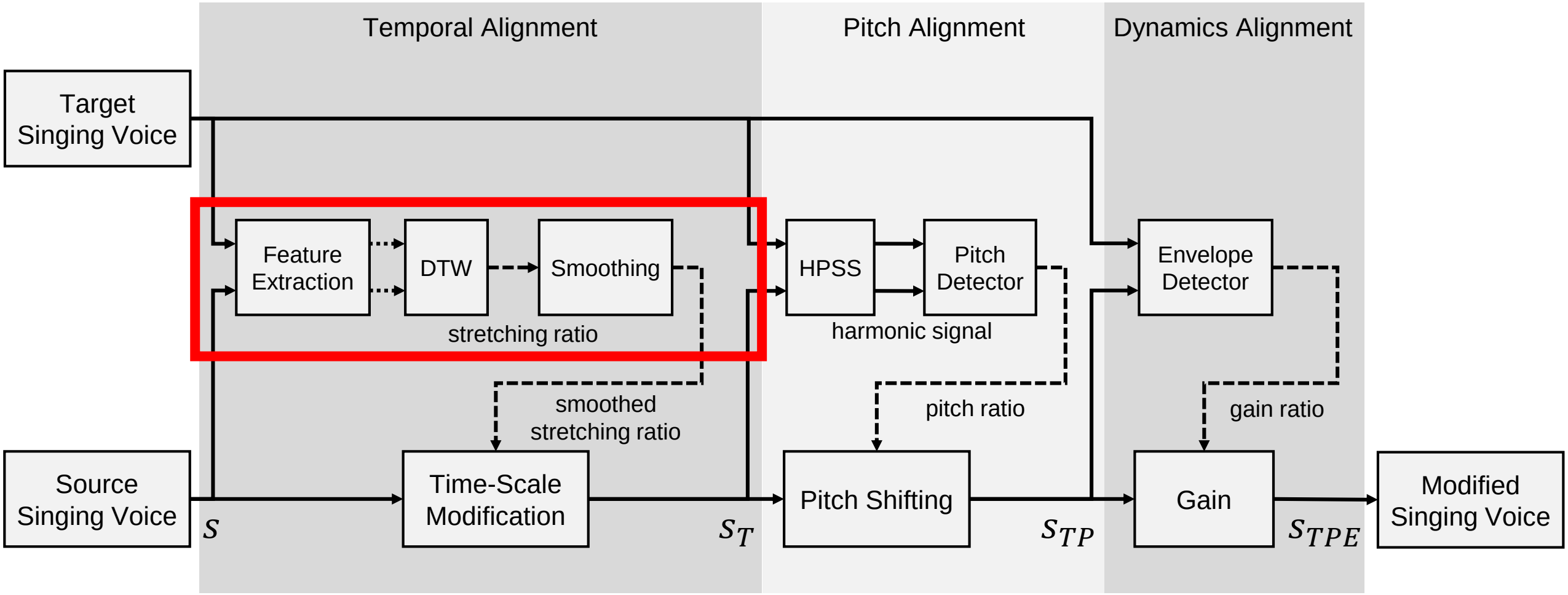
System Structure



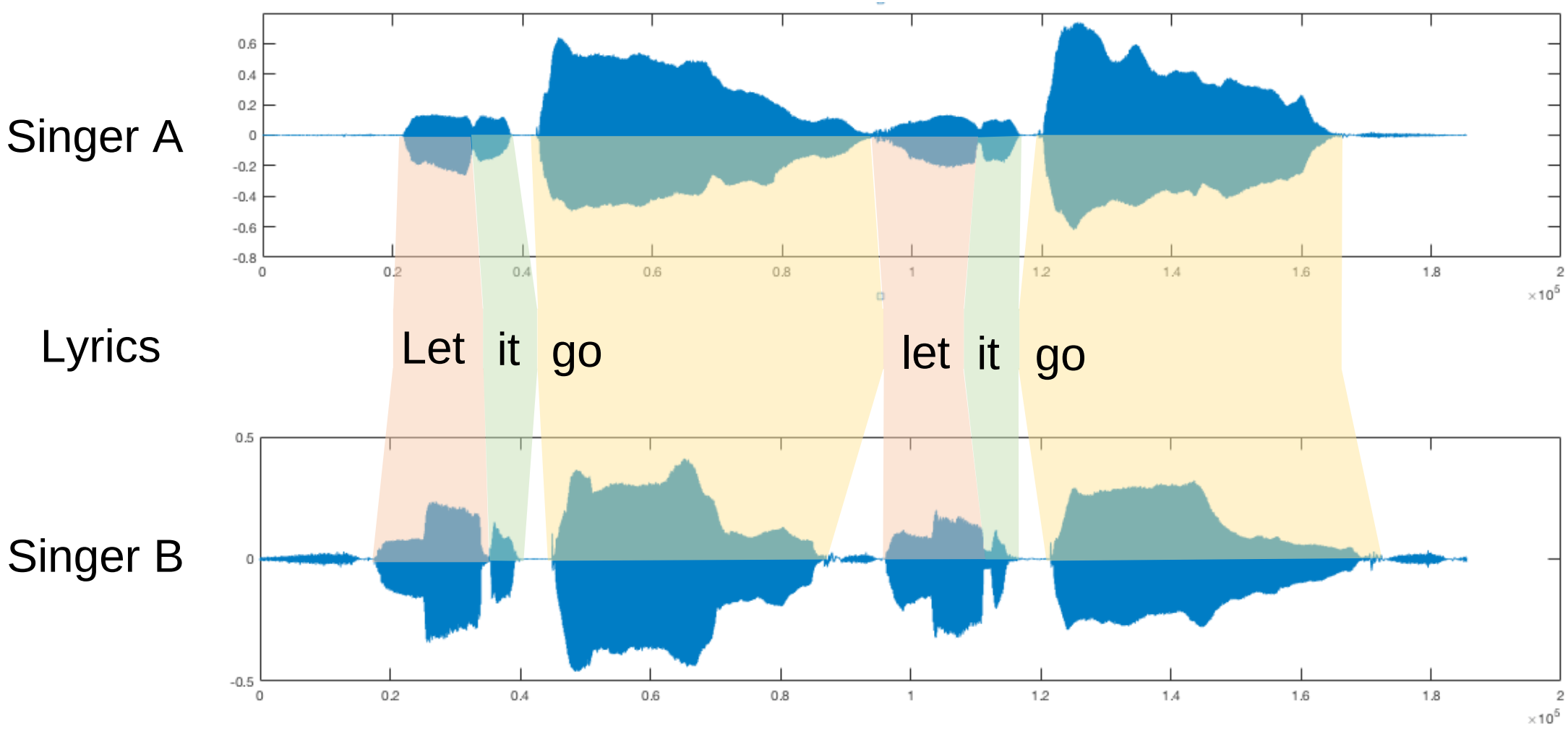
System Structure



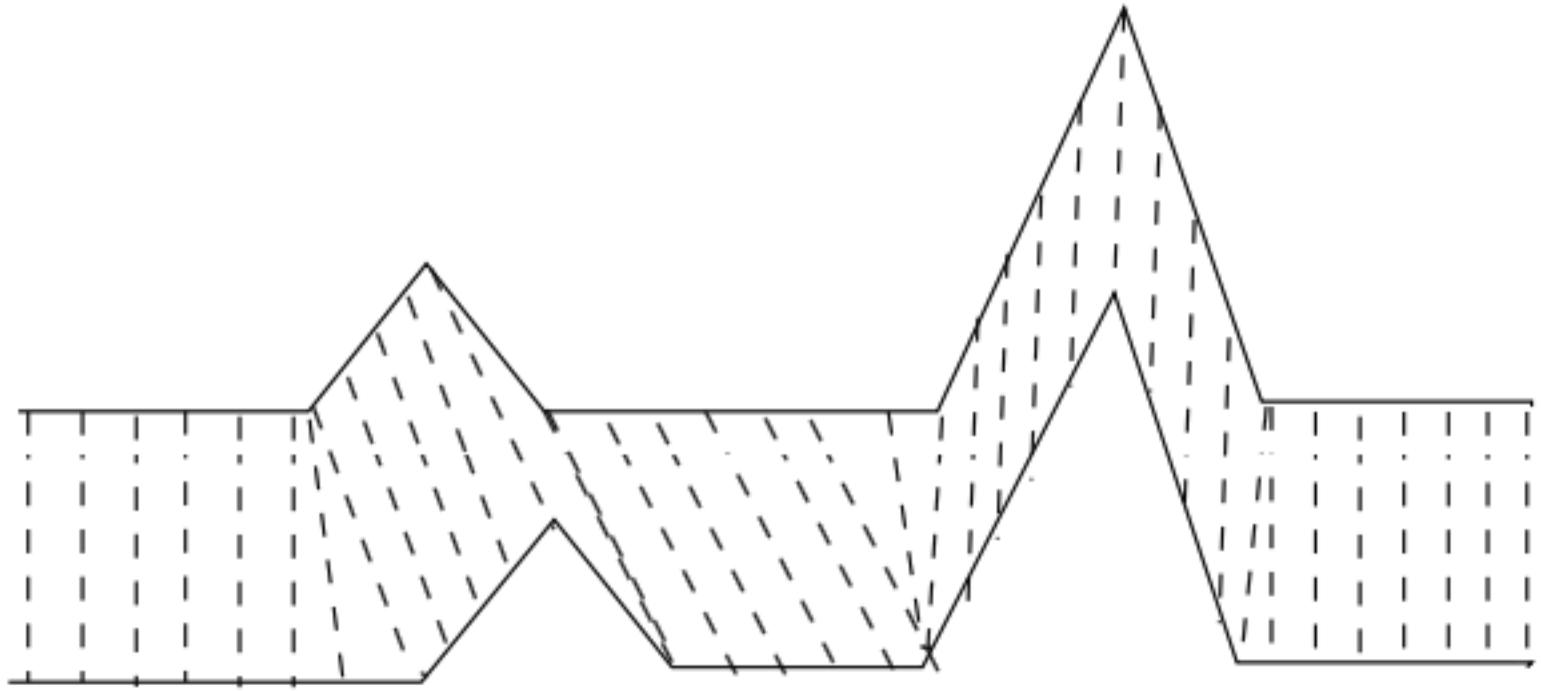
System Structure



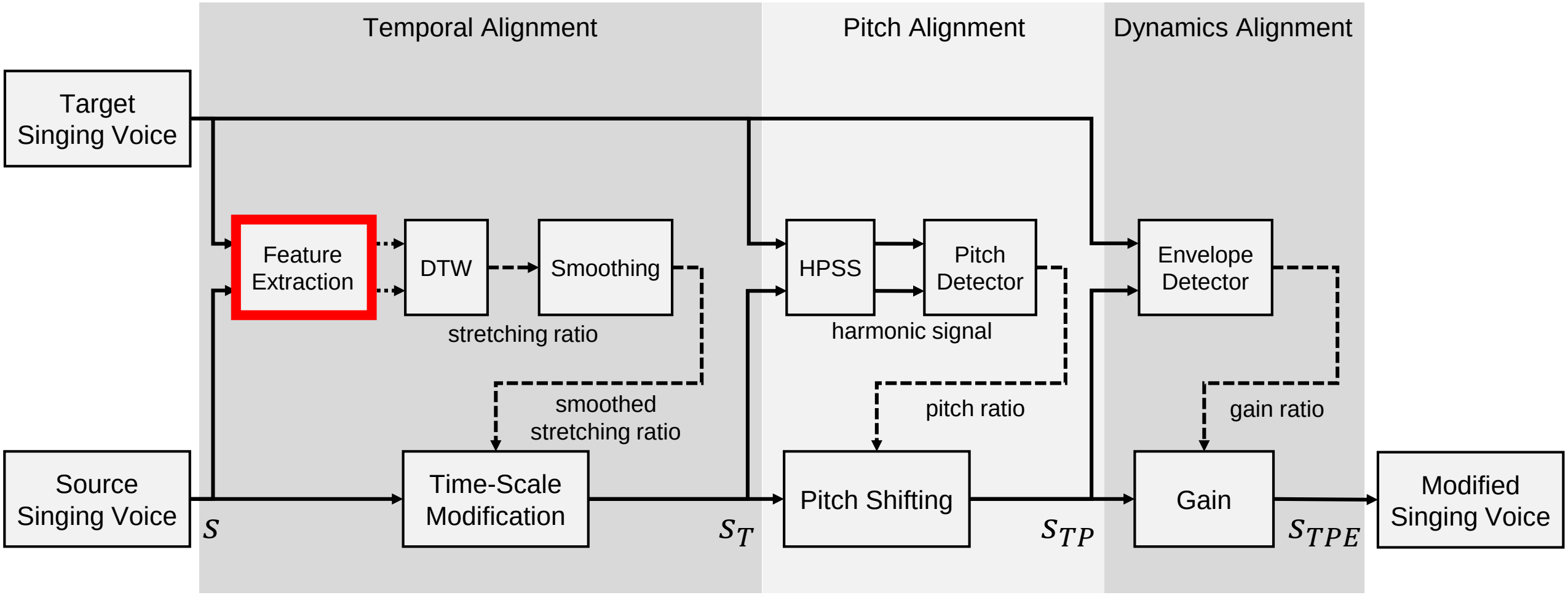
Temporal Alignment



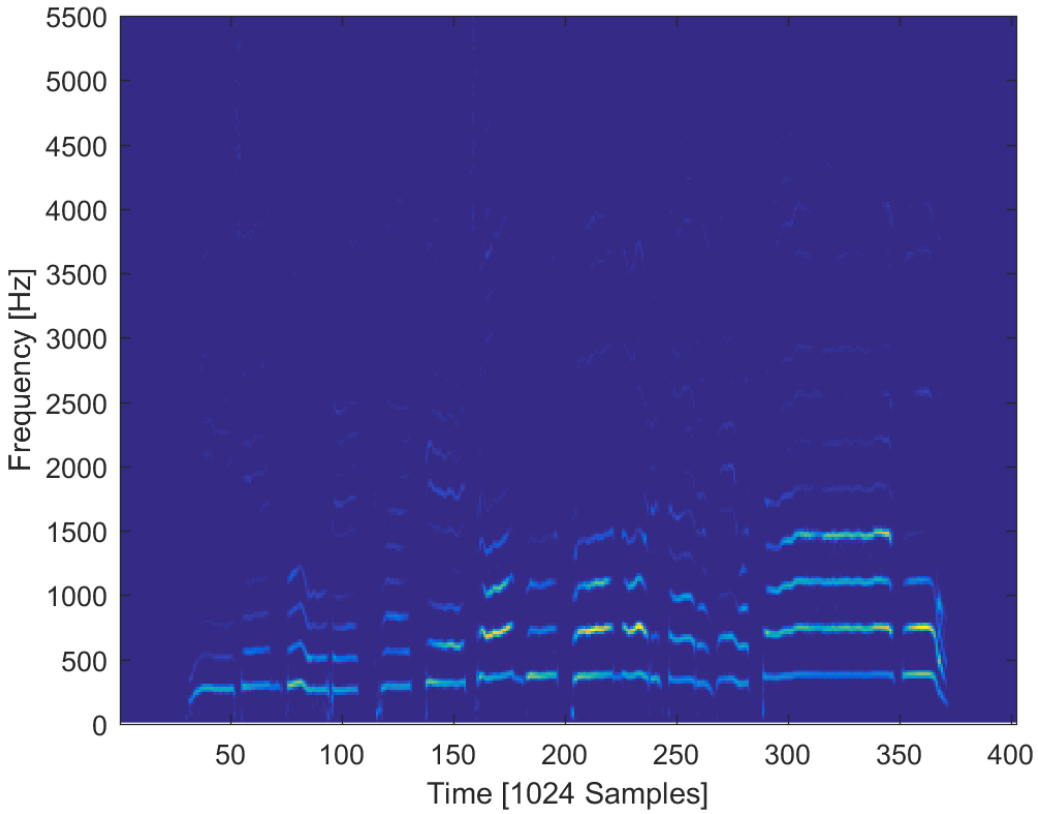
Temporal Alignment – Dynamic Time Warping



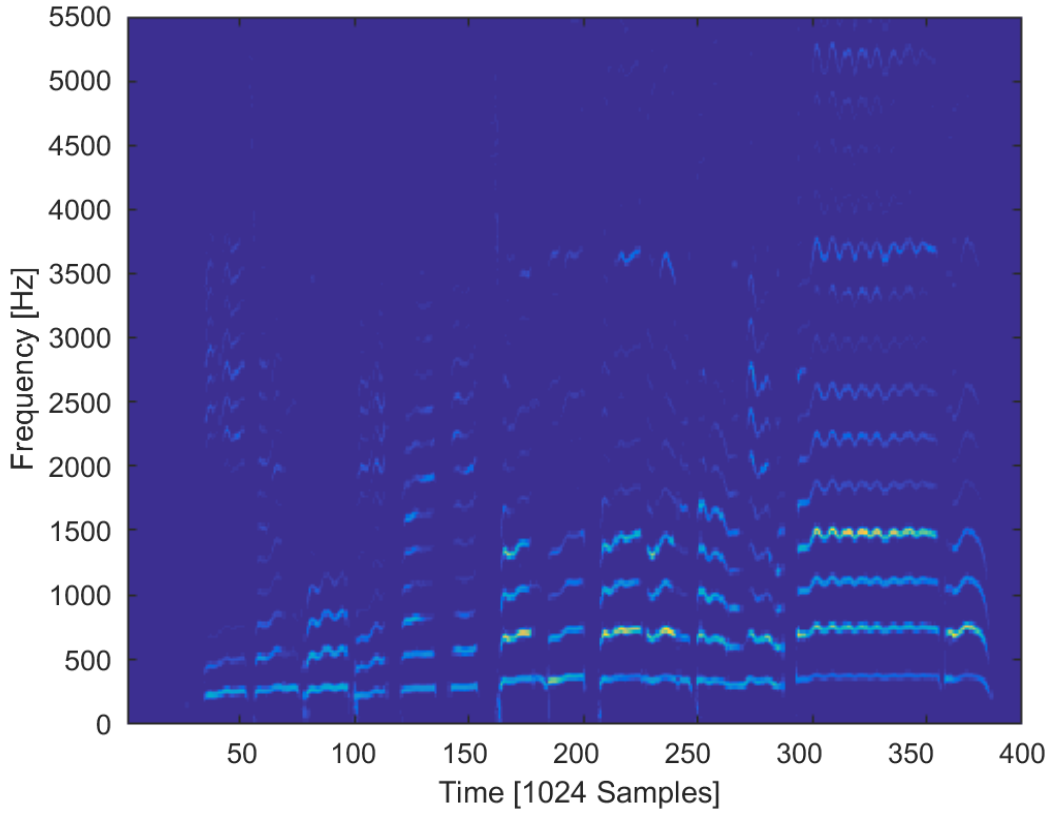
System Structure



Temporal Alignment – Feature Extraction

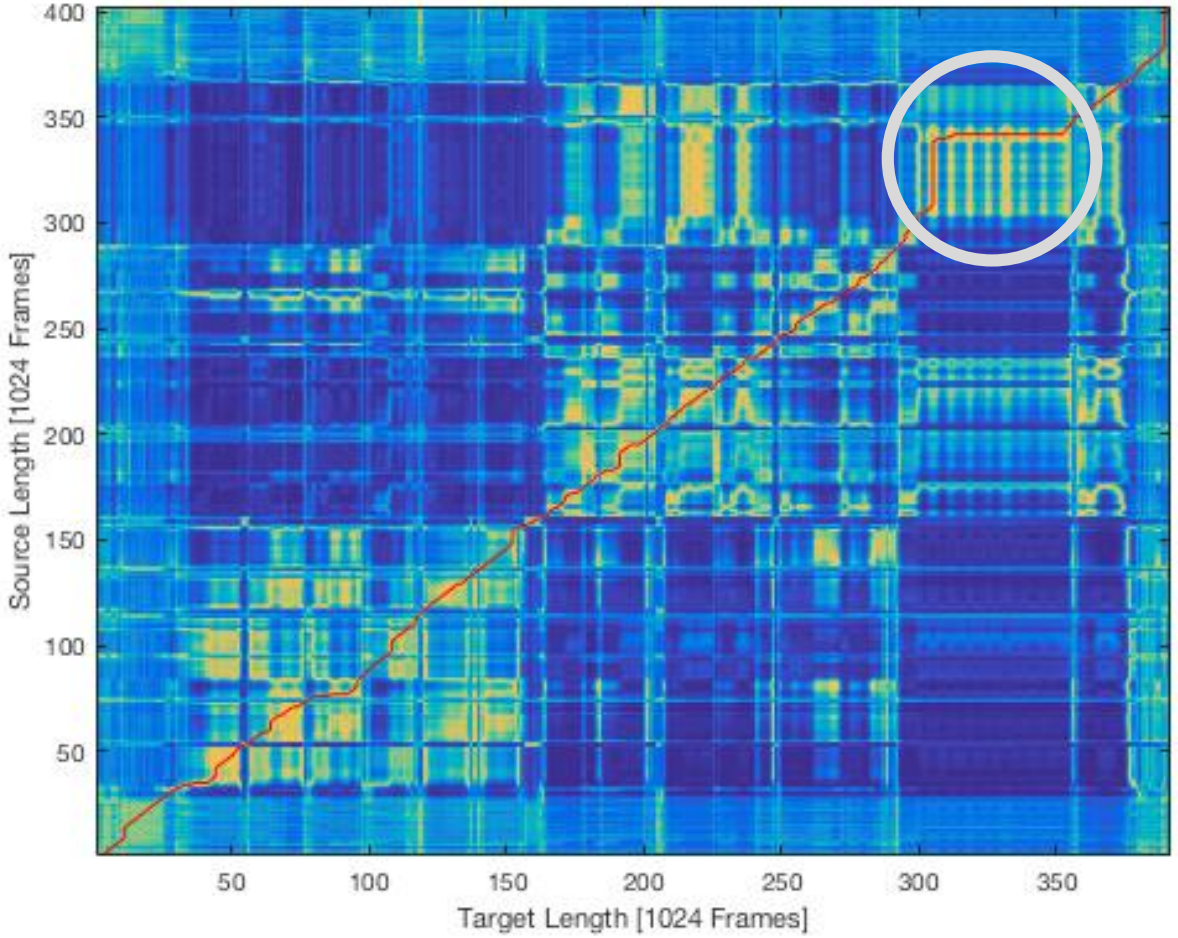


Spectrogram of Source



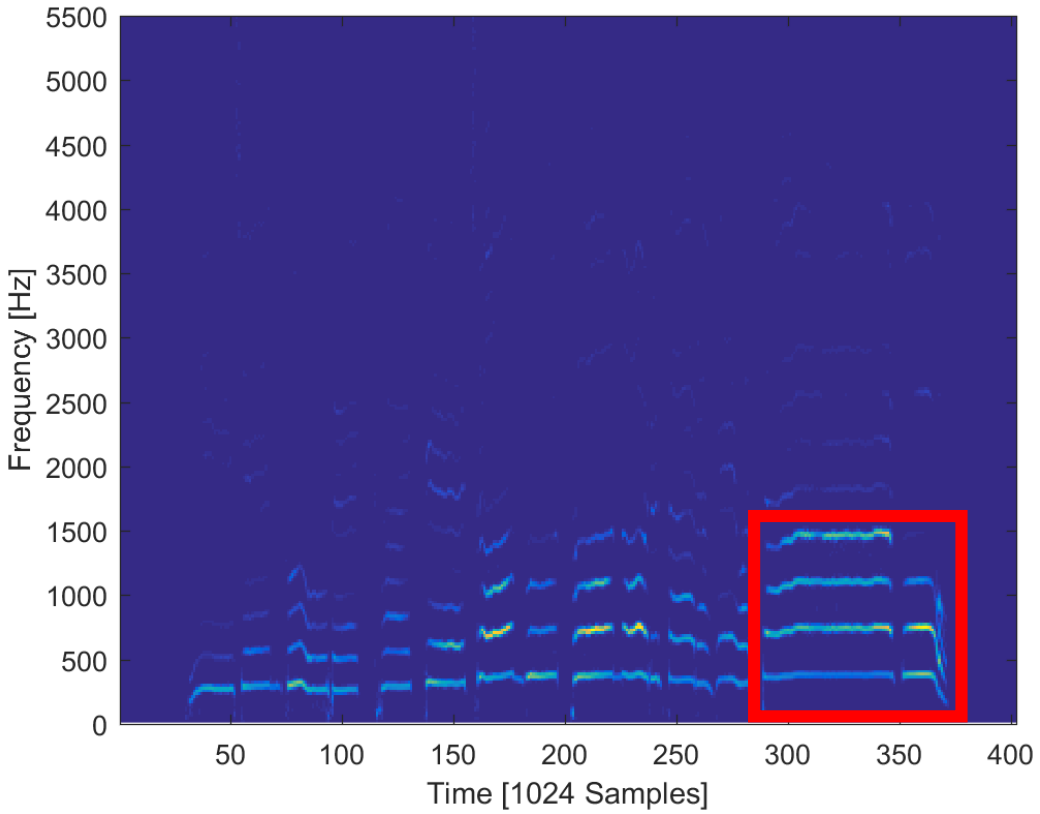
Spectrogram of Target

Temporal Alignment – Feature Extraction

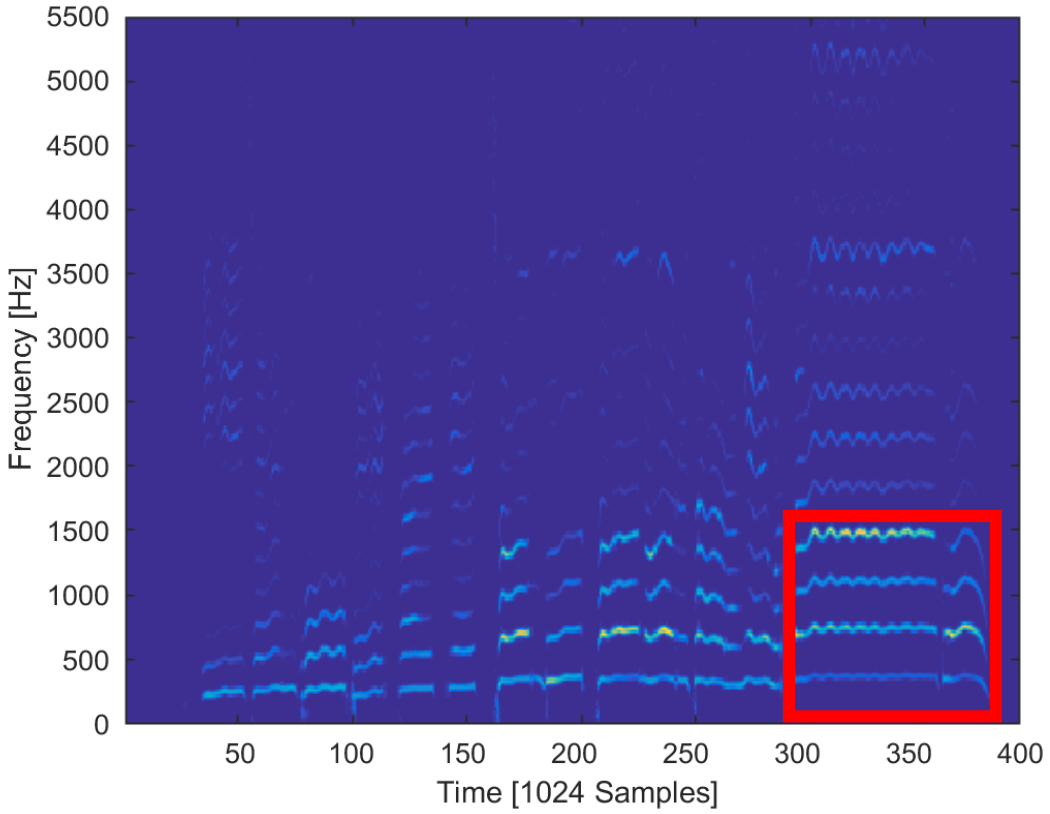


Similarity matrix with spectrogram

Temporal Alignment – Feature Extraction



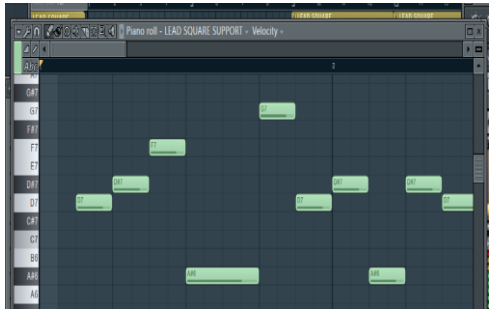
Spectrogram of Source

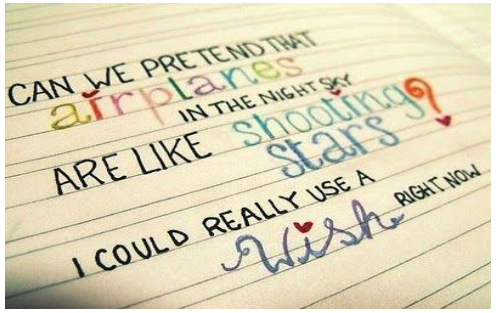


Spectrogram of Target

Feature Extraction Strategy

- Preserving common elements
  - Note-level melody
  - Lyrics

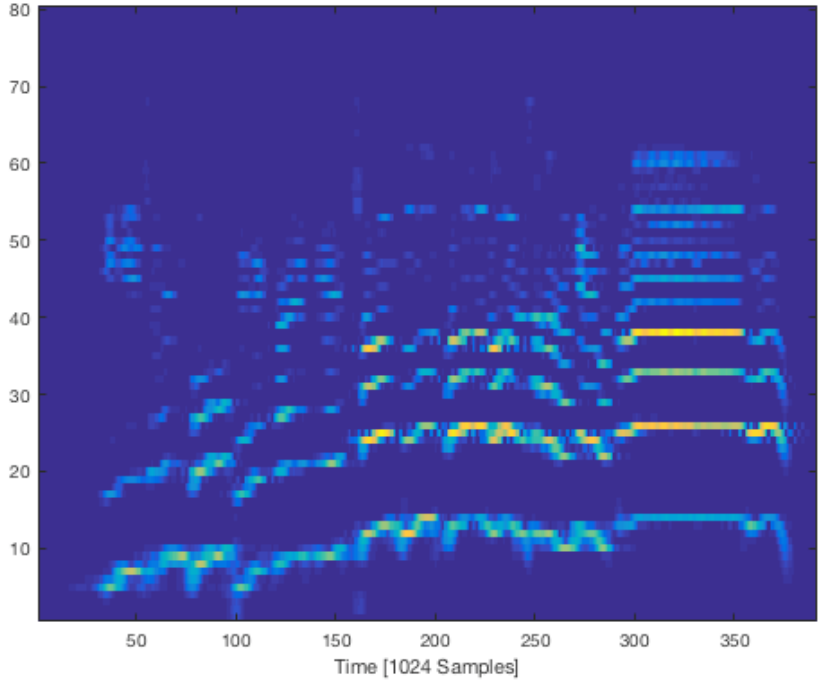




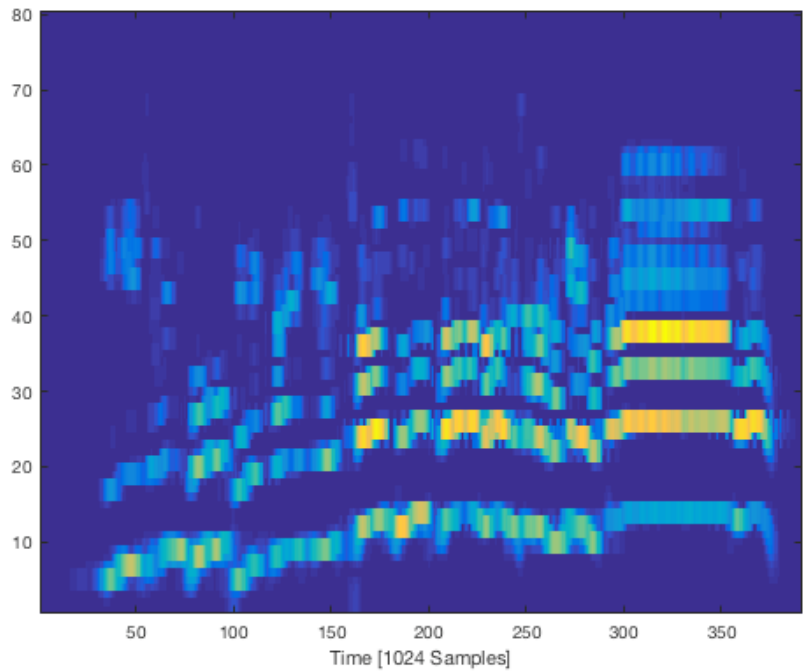
- Suppressing different characteristics
  - Vibrato or other pitch-related articulations
  - Singer timbre

Proposed Features

- Max-filtered Constant-Q transform
  - Semi-tone pitch resolution: vibrato with less than one semi-tone
  - Frequency-wise max-filtering: vibrato with more than one semi-tone



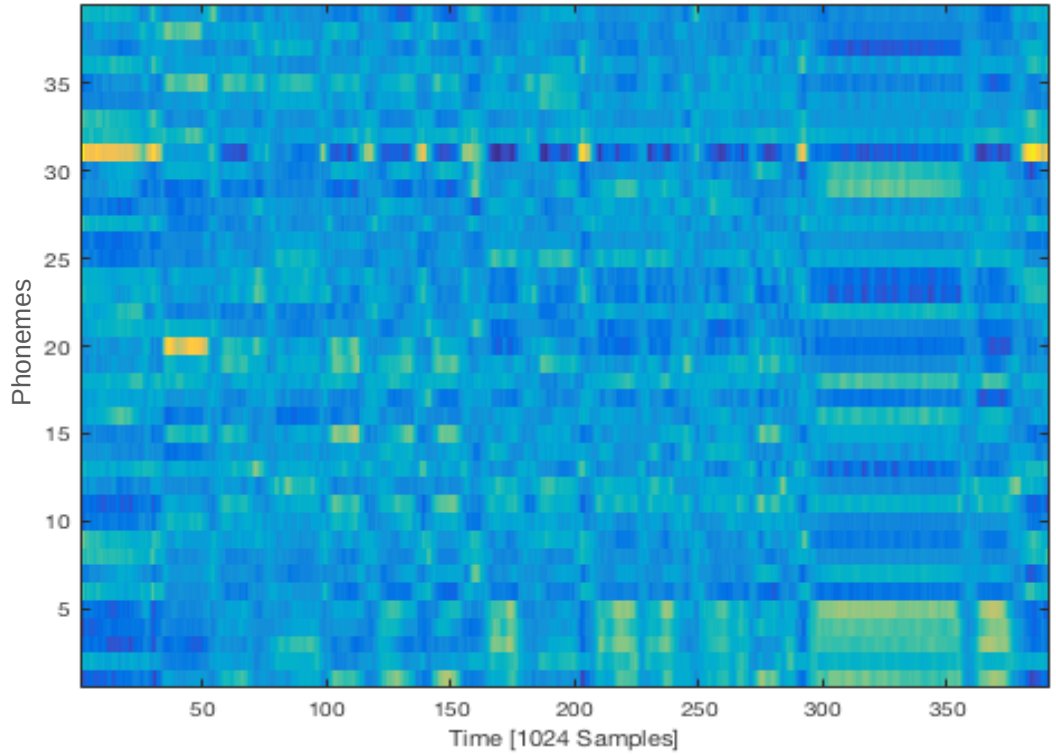
Constant-Q Transform



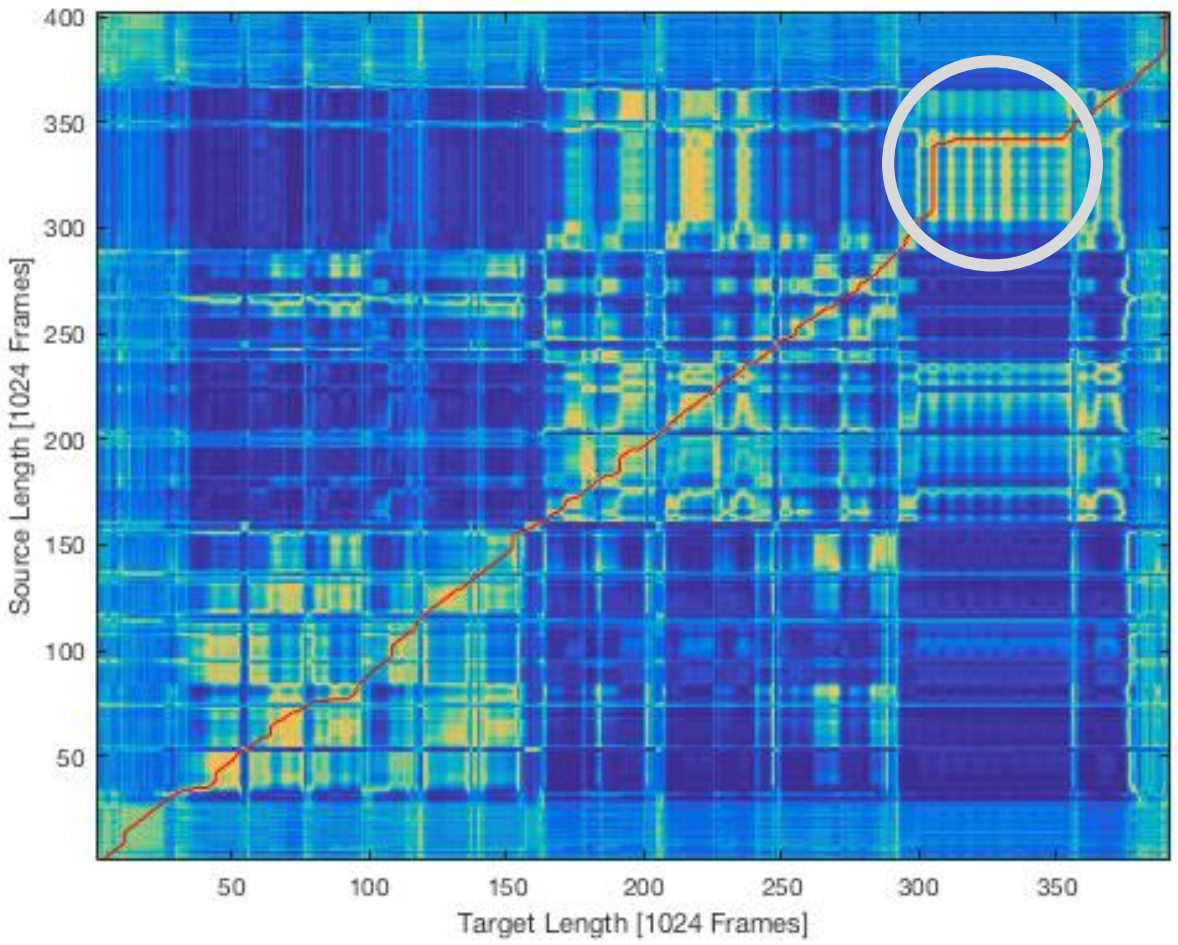
Const-Q Trans with Maximum Filtering

Proposed Features

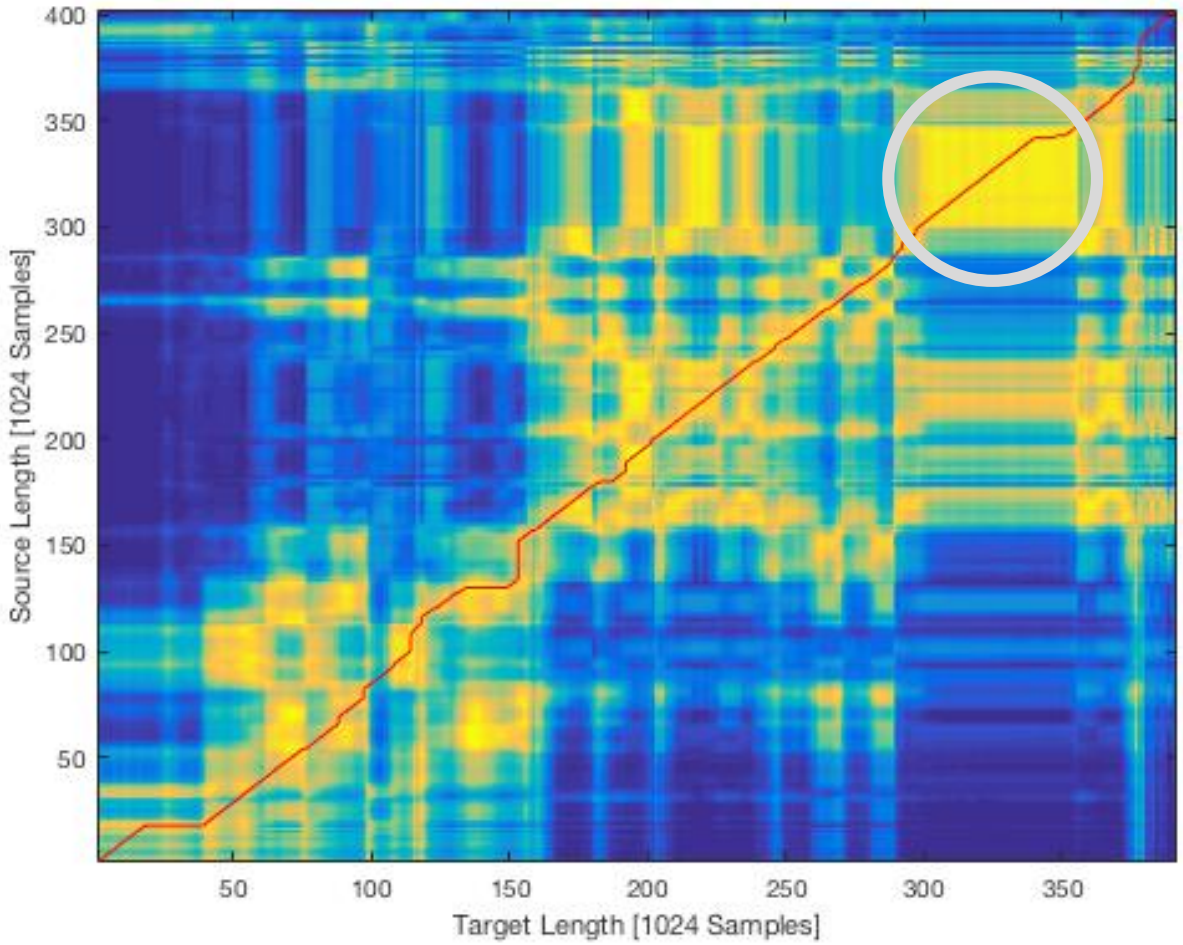
- Phoneme score (phoneme classifier posteriorgram)
  - Frame-level features for accurate temporal alignment
  - Singer invariant lyrical features



Temporal Alignment – Feature Comparison

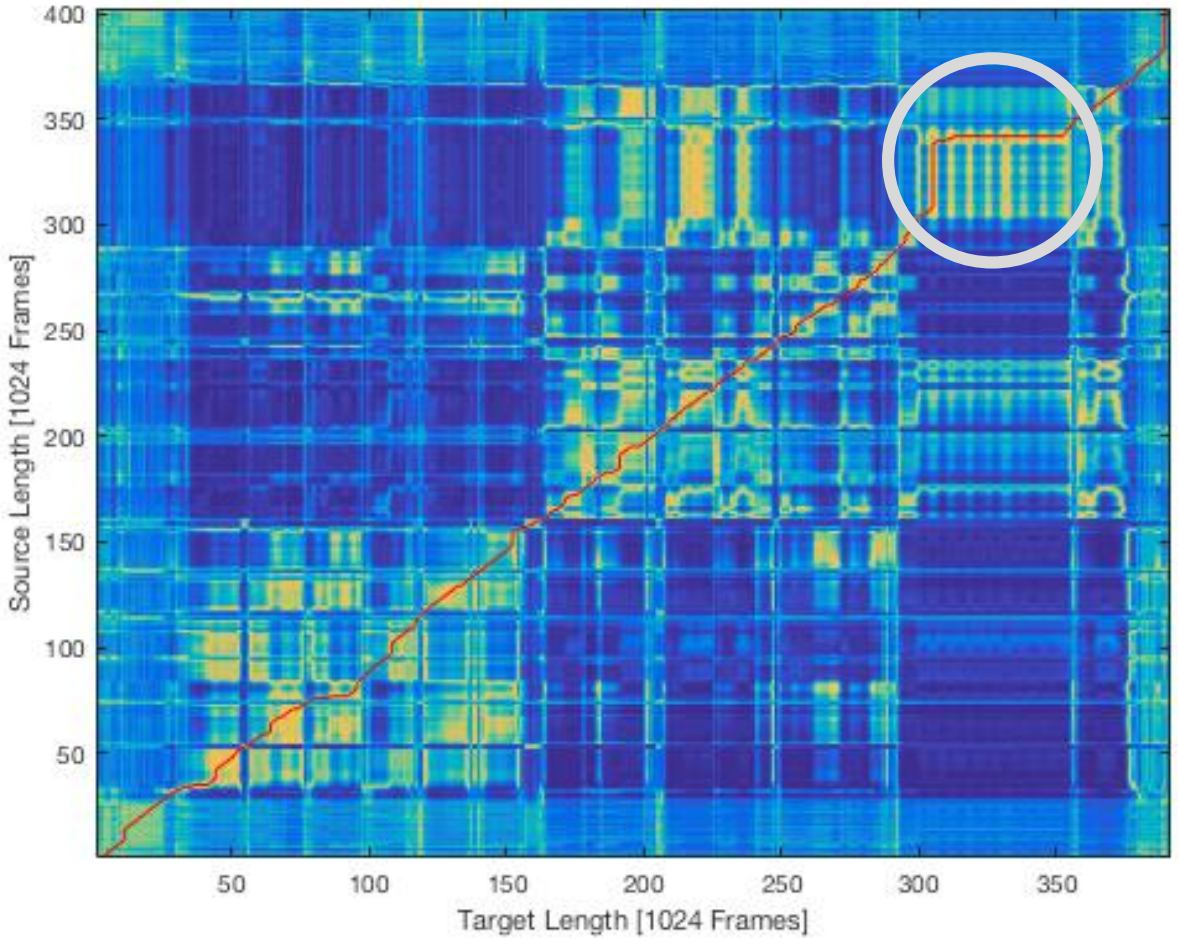


Spectrogram

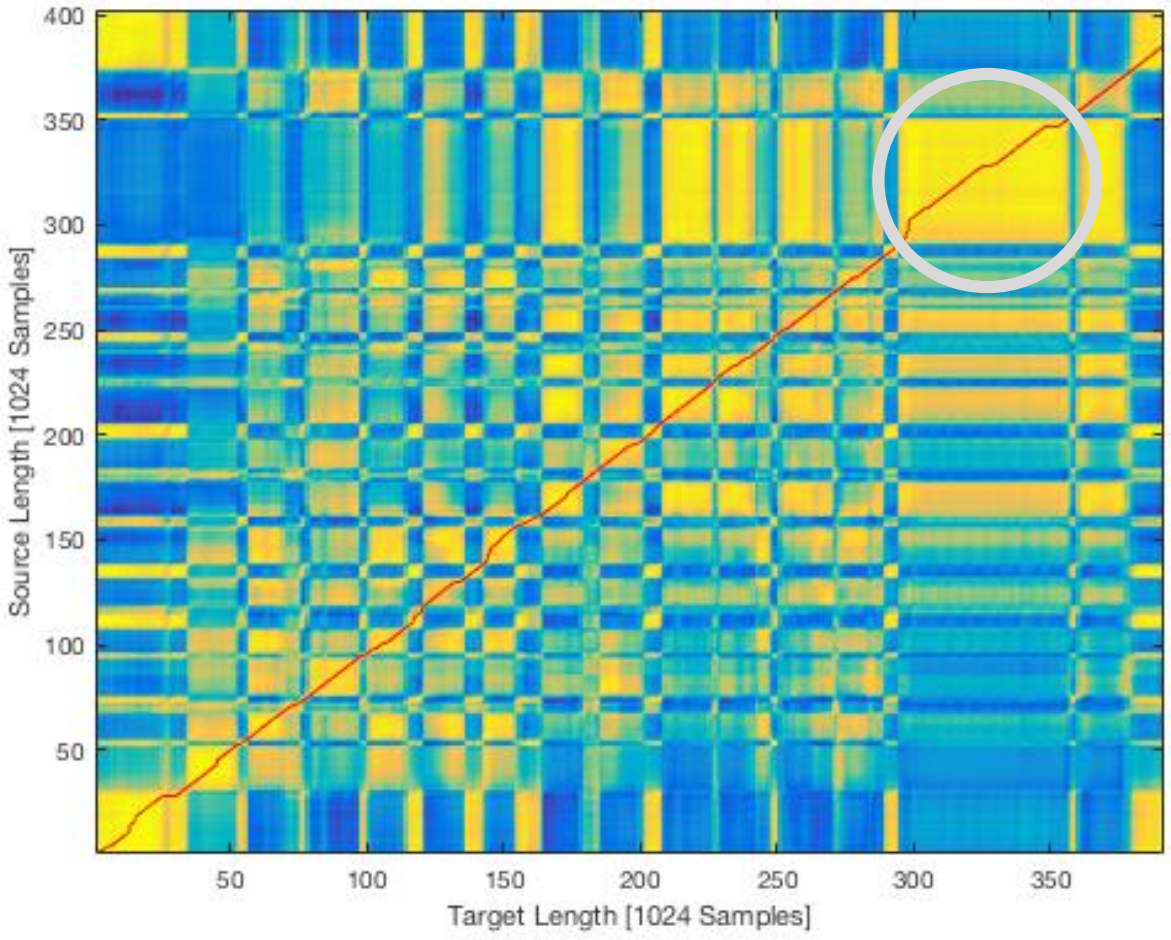


Max-filtered Constant-Q Transform

Temporal Alignment – Feature Comparison

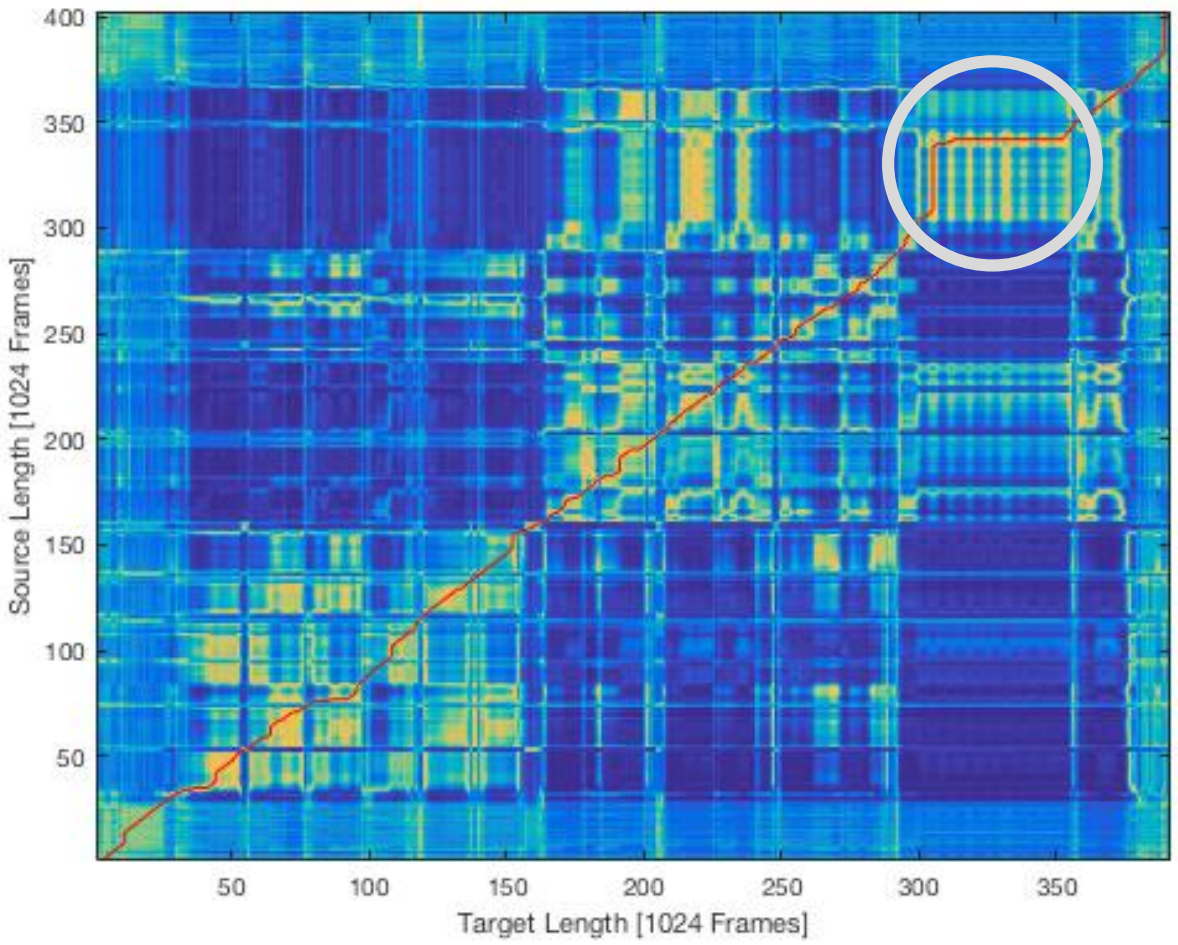


Spectrogram

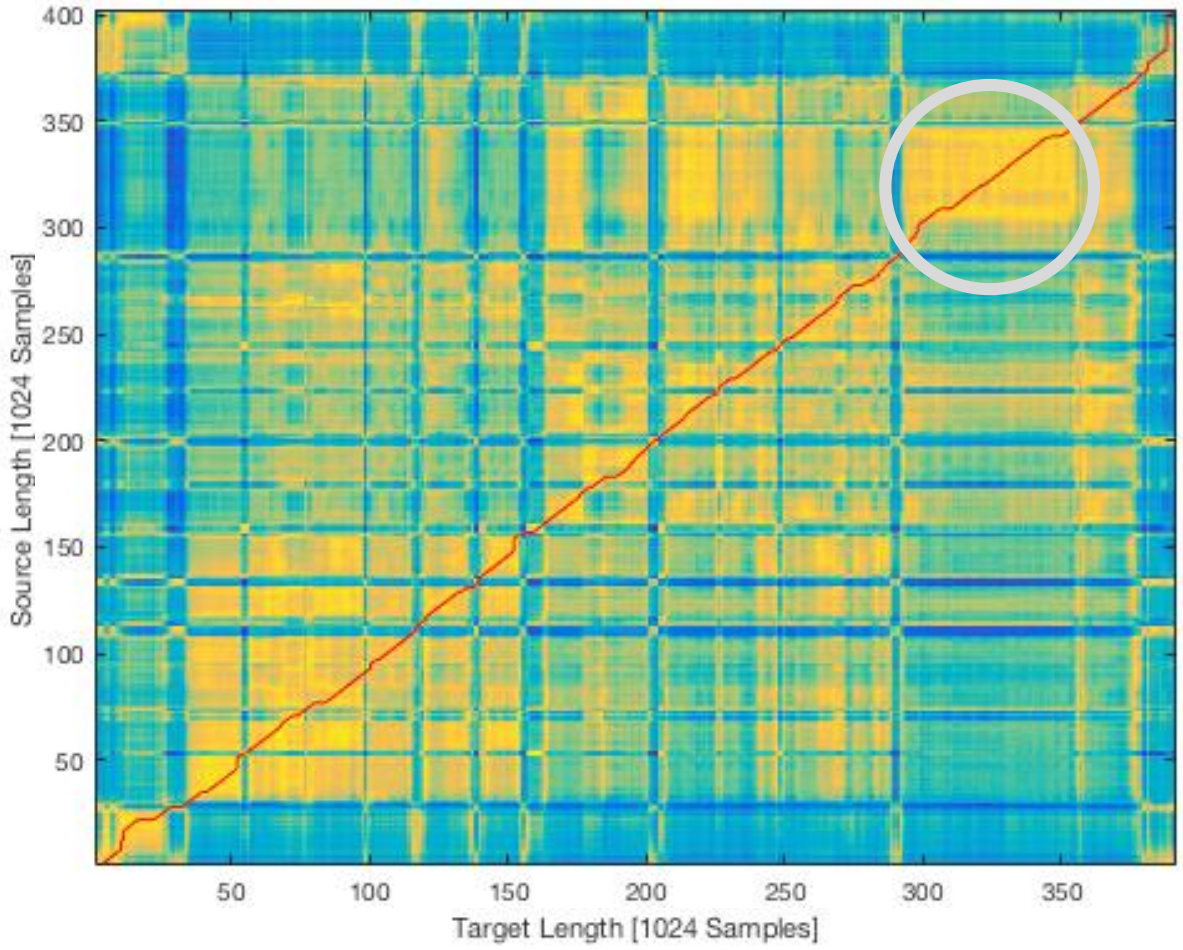


phoneme score

Temporal Alignment – Feature Comparison

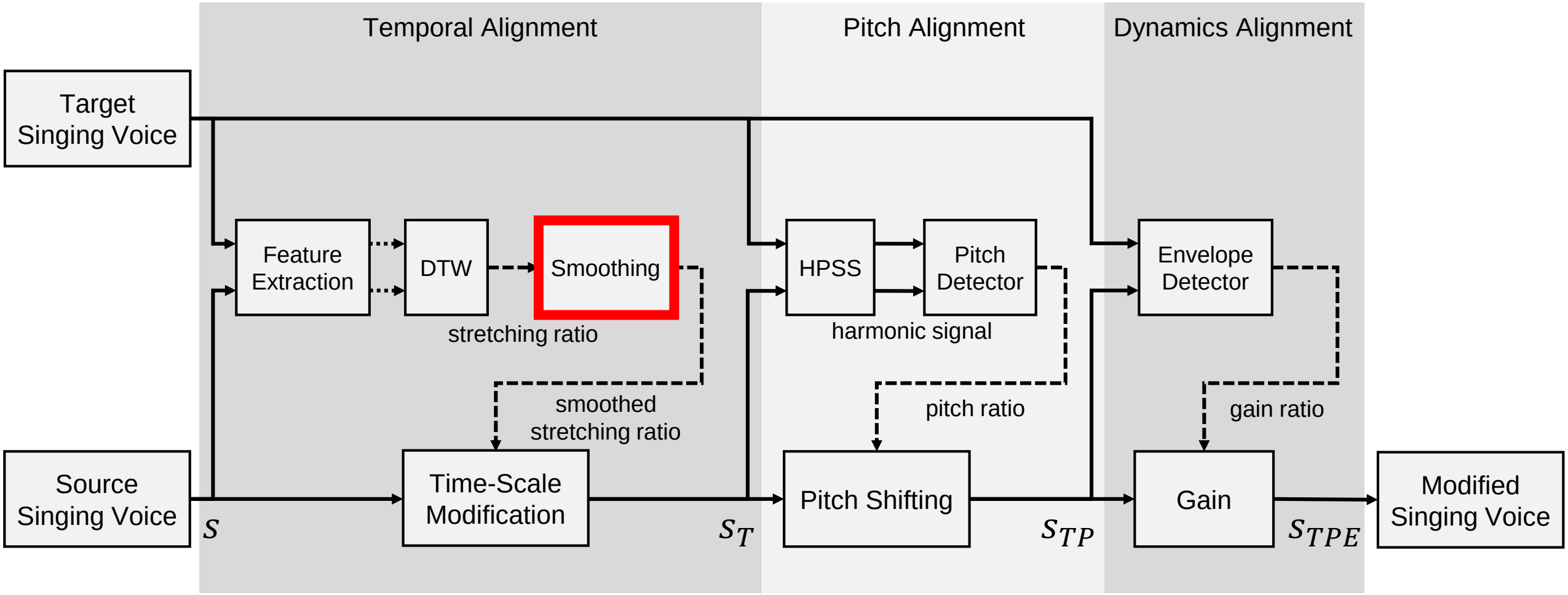


Spectrogram

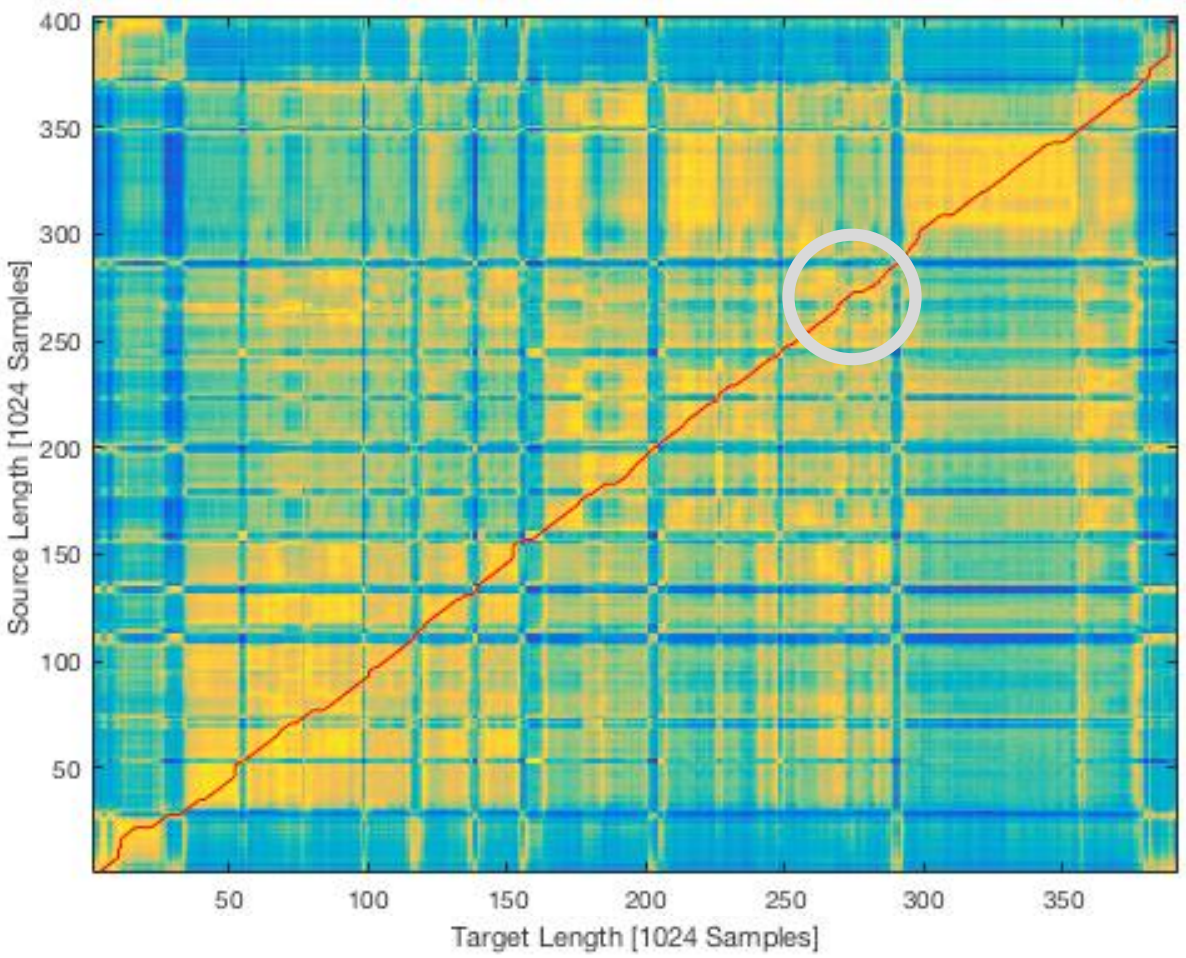


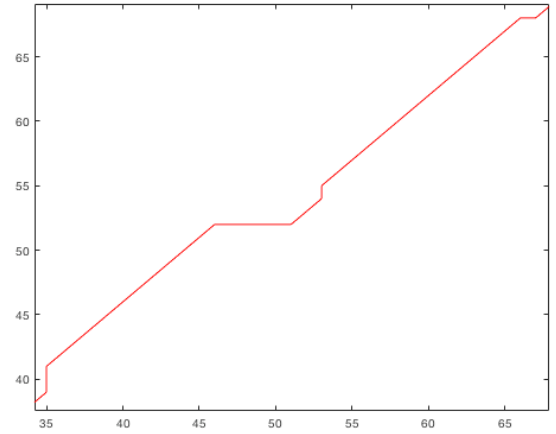
Phoneme Score +Const-Q Trans

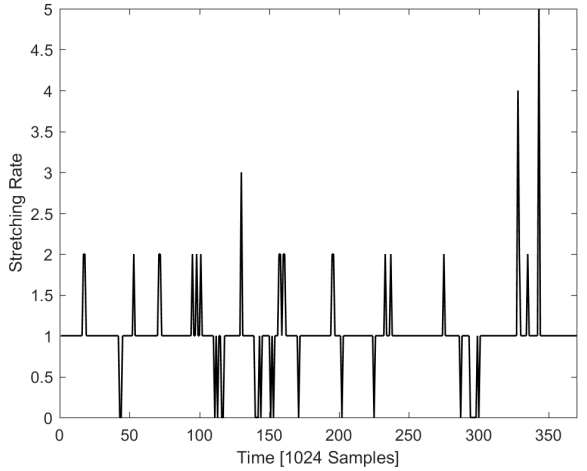
System Structure



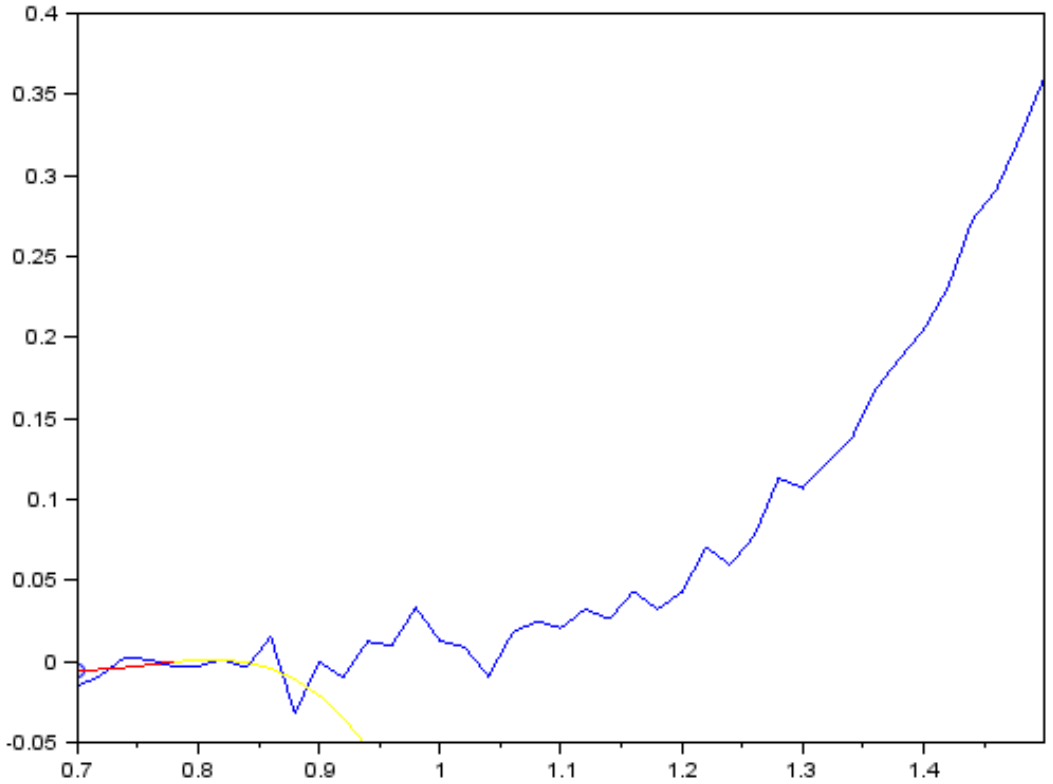
Temporal Alignment – Path Smoothing





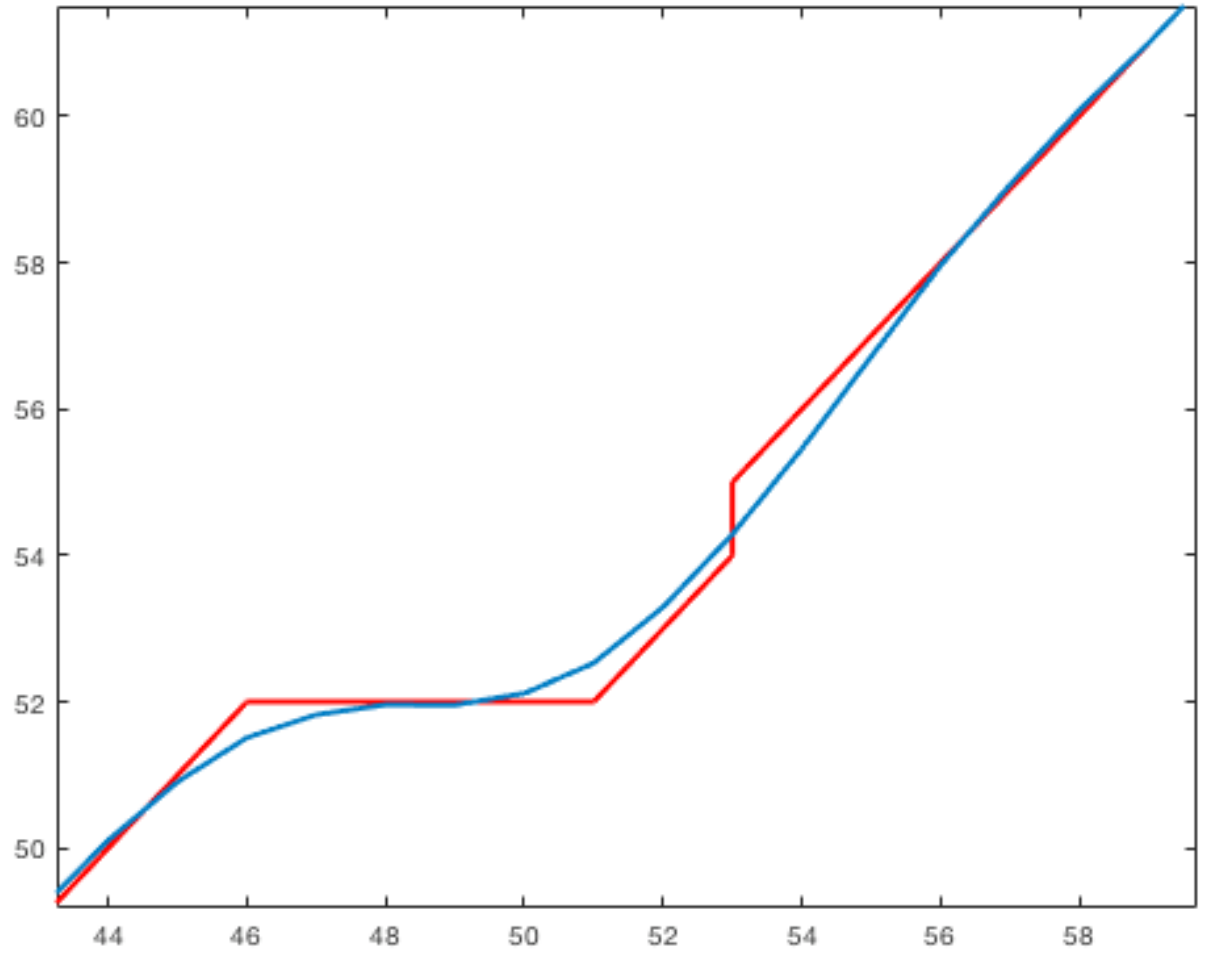


Temporal Alignment – Path Smoothing

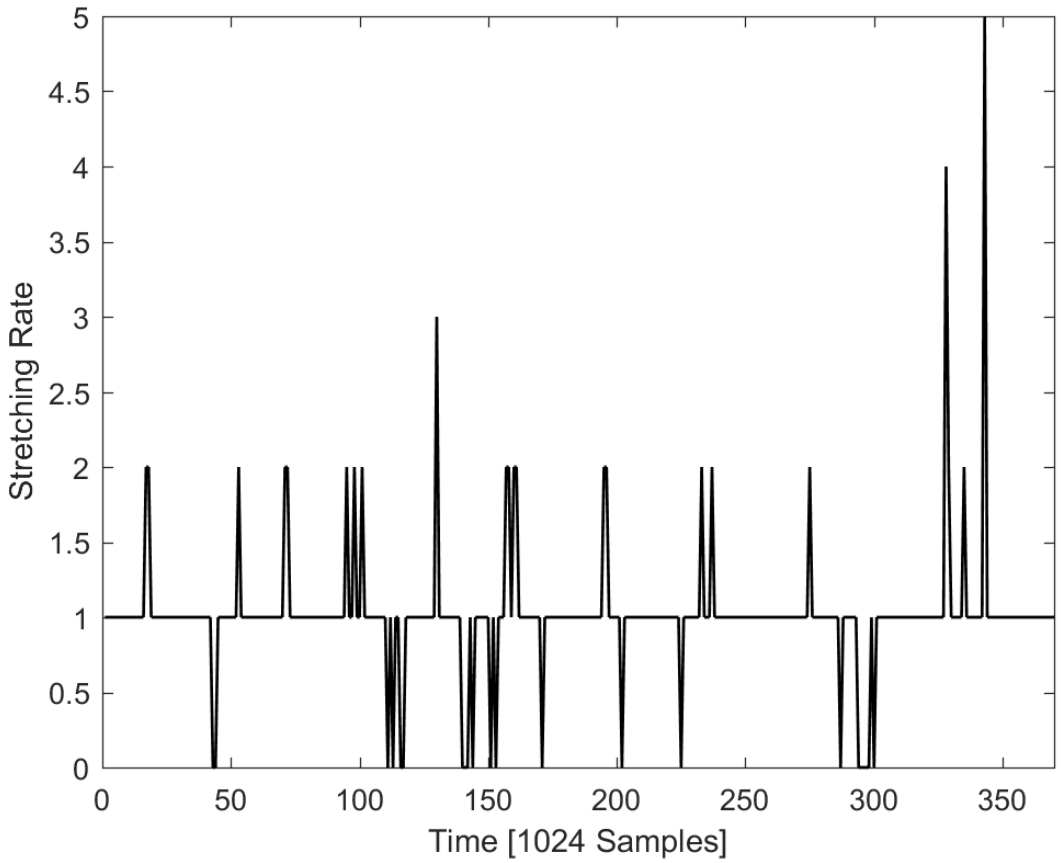


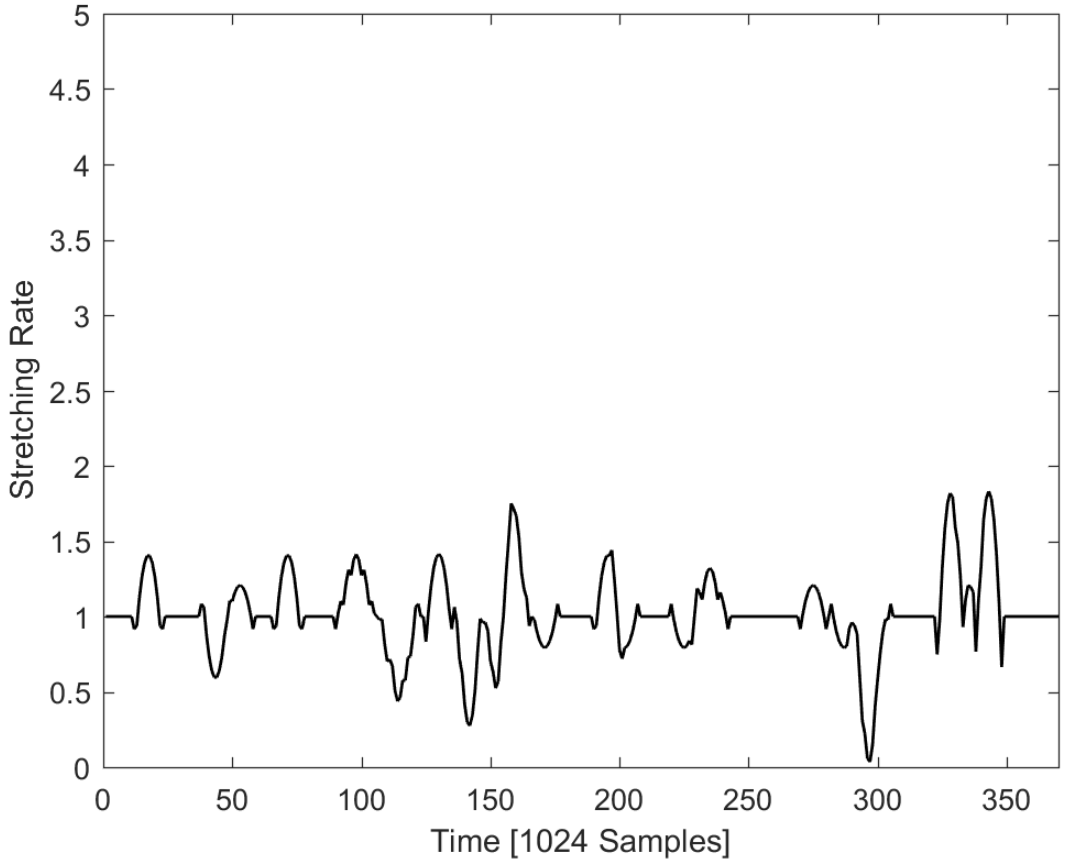
Savitzky, Abraham, and Marcel JE Golay. "Smoothing and differentiation of data by simplified least squares procedures." *Analytical chemistry* 36.8 (1964): 1627-1639.

Temporal Alignment – Path Smoothing

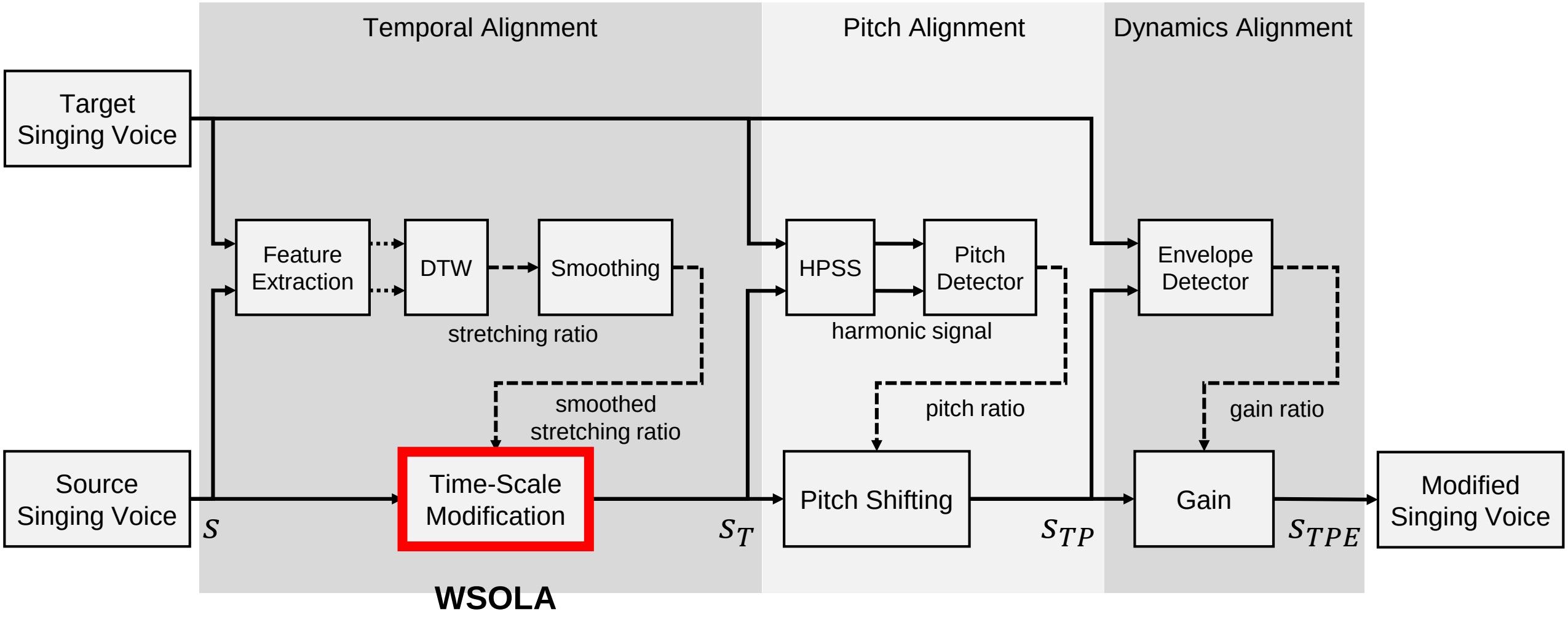


Temporal Alignment – Path Smoothing

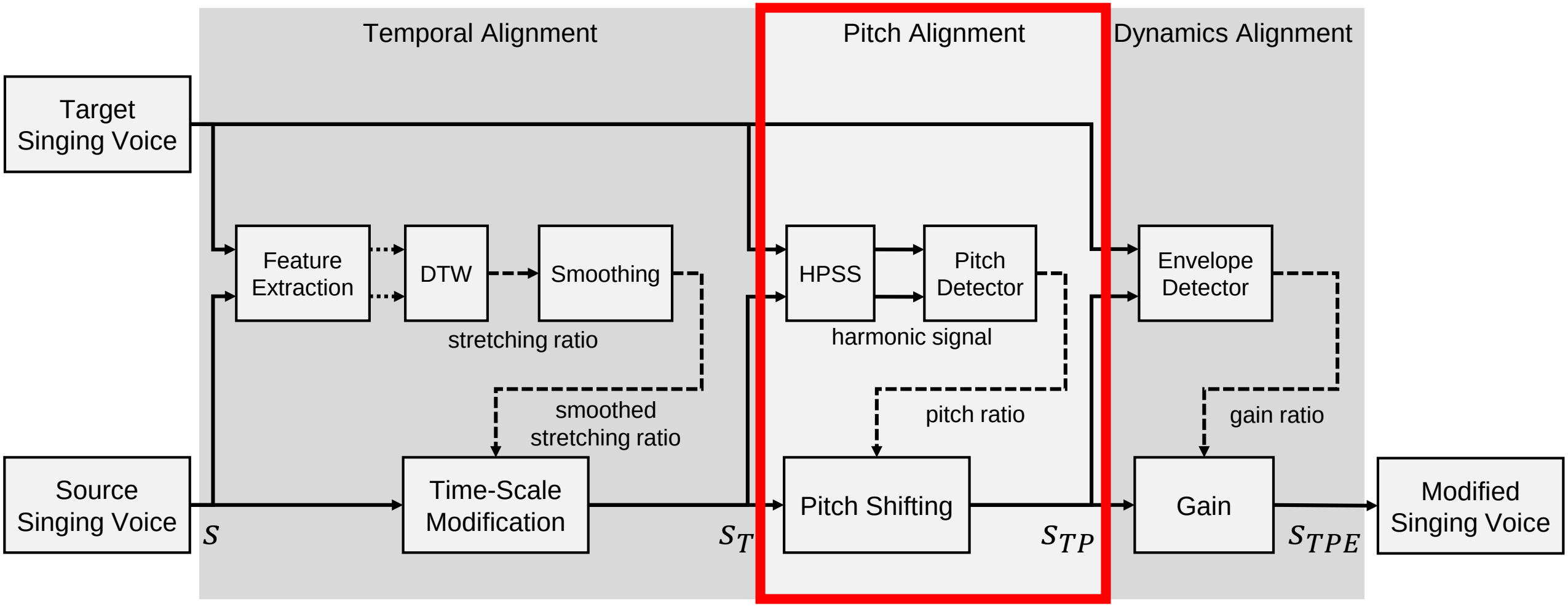




System Structure



System Structure



Pitch Alignment

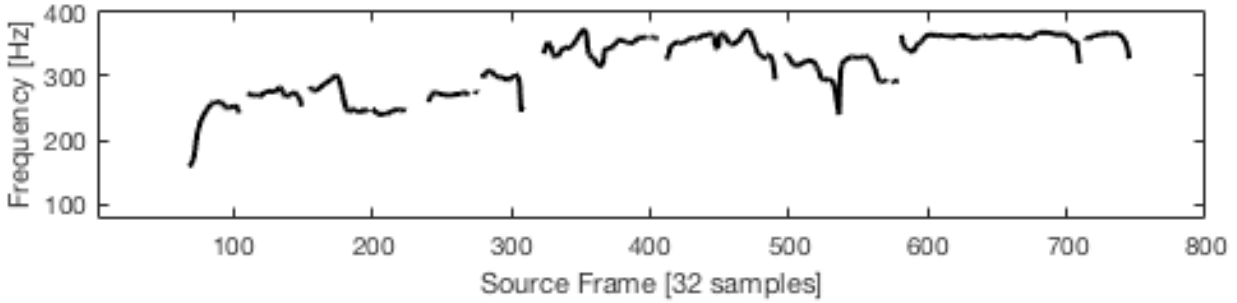
- Harmonic-Percussion Source Separation (HPSS)
  - Pre-processing of pitch detection to increase detection accuracy
  - Median filter (IEEE Signal Processing Letters 2014)

- Pitch Detector
  - YIN

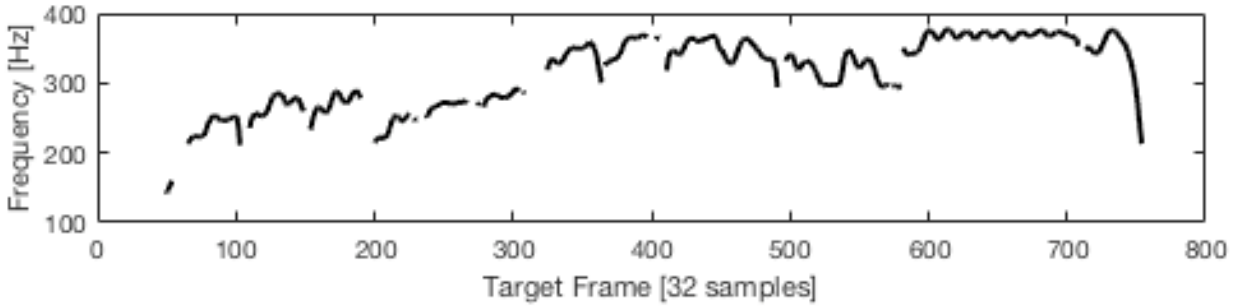
- Pitch shifting
  - Pitch-Synchronous Overlap-Add (PSOLA)
  - Formant preservation

Pitch Alignment

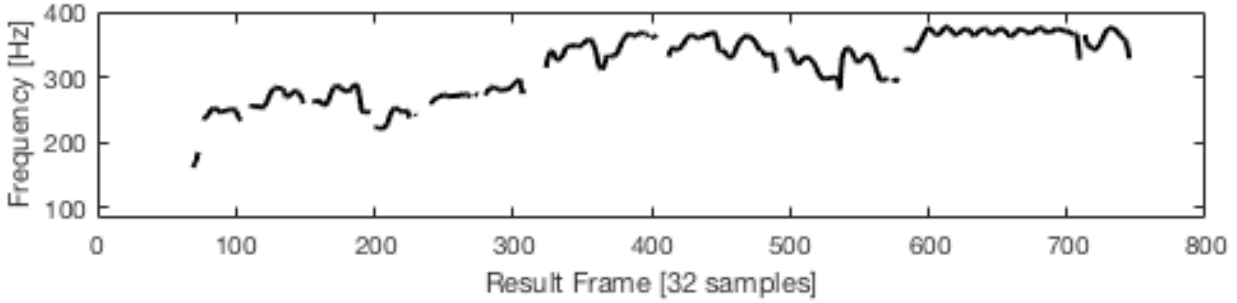
source



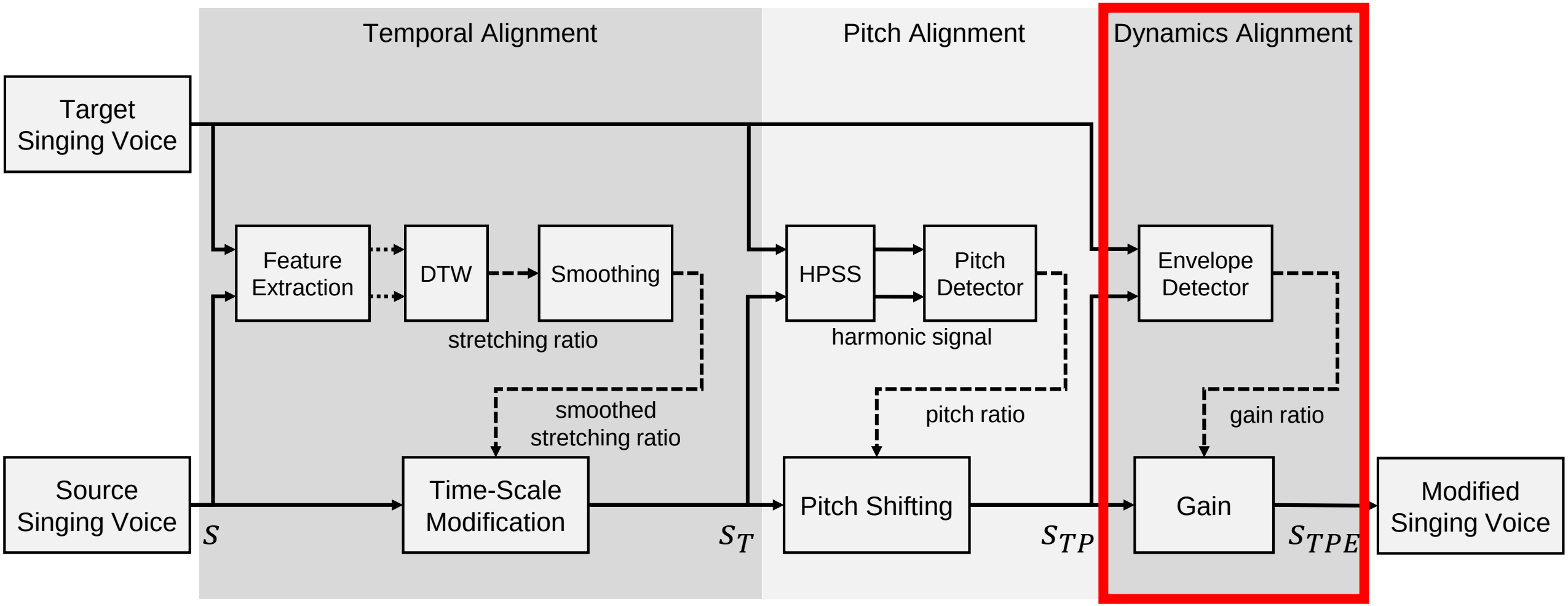
target



result



System Structure

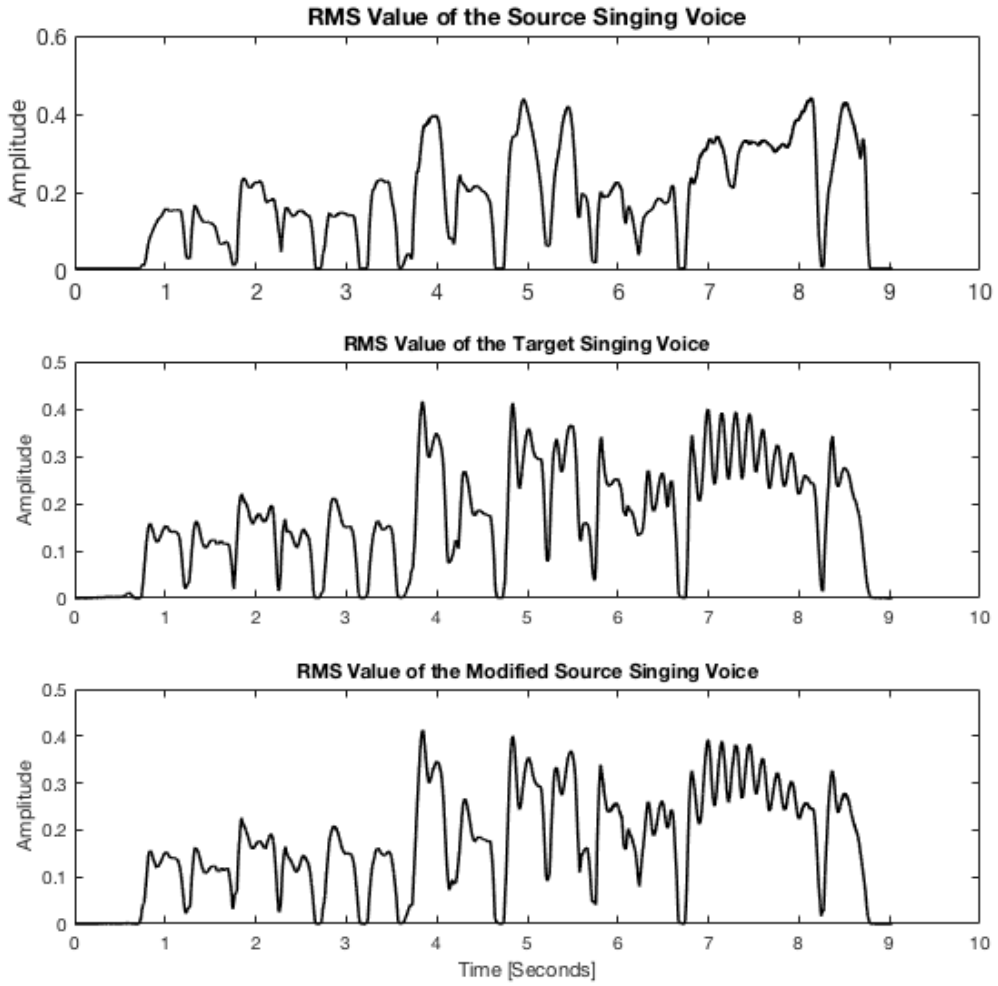


Dynamics Alignment

source

target

result



Evaluation

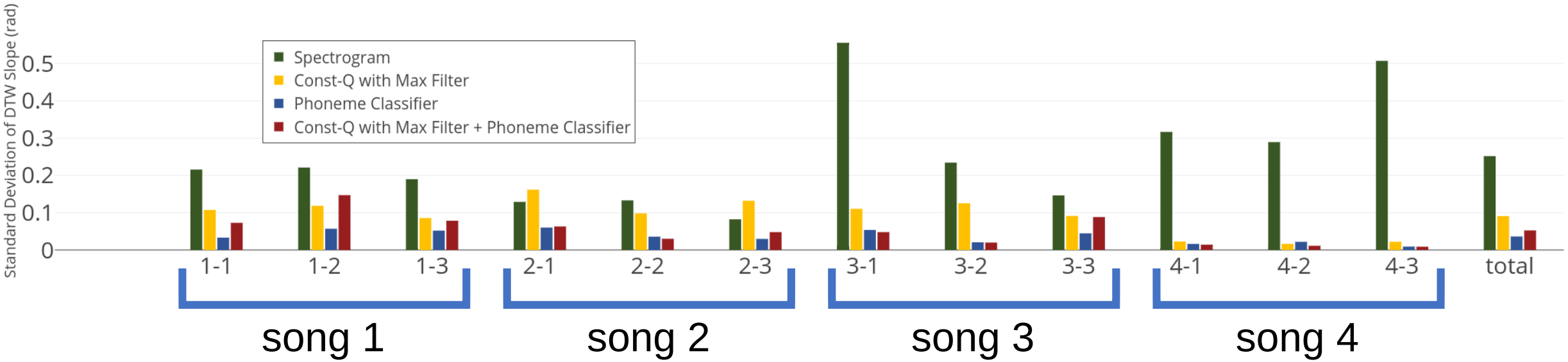
- Datasets
  - 4 recordings for each of 4 songs (total 16 recordings)
  - One of 4 recordings is a target singing voice (professional or skilled)
  - Totally 12 pairs of source-target singing voice

	Song 1	Song 2	Song 3	Song 4
Gender	female	male	male	male
No. of source	3	3	3	3
Remarks	high pitch English	low pitch English	swing rhythm Korean	swing rhythm Korean

Evaluation

	Song 1	Song 2	Song 3	Song 4
Gender	female	male	male	male
No. of source	3	3	3	3
Remarks	high pitch English	low pitch English	swing rhythm Korean	swing rhythm Korean

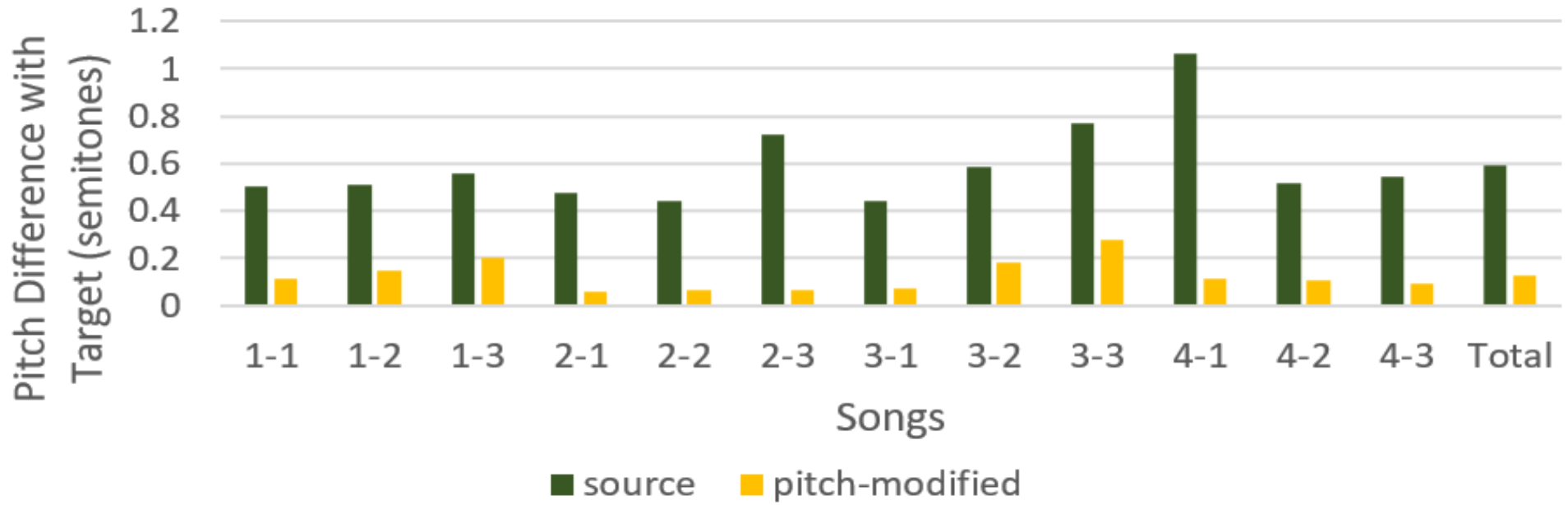
- Temporal alignment
  - Better alignment has less fluctuation of the DTW slope
  - Standard deviation of slope angle $\theta = \arctan(slope)$



Evaluation

- Pitch alignment

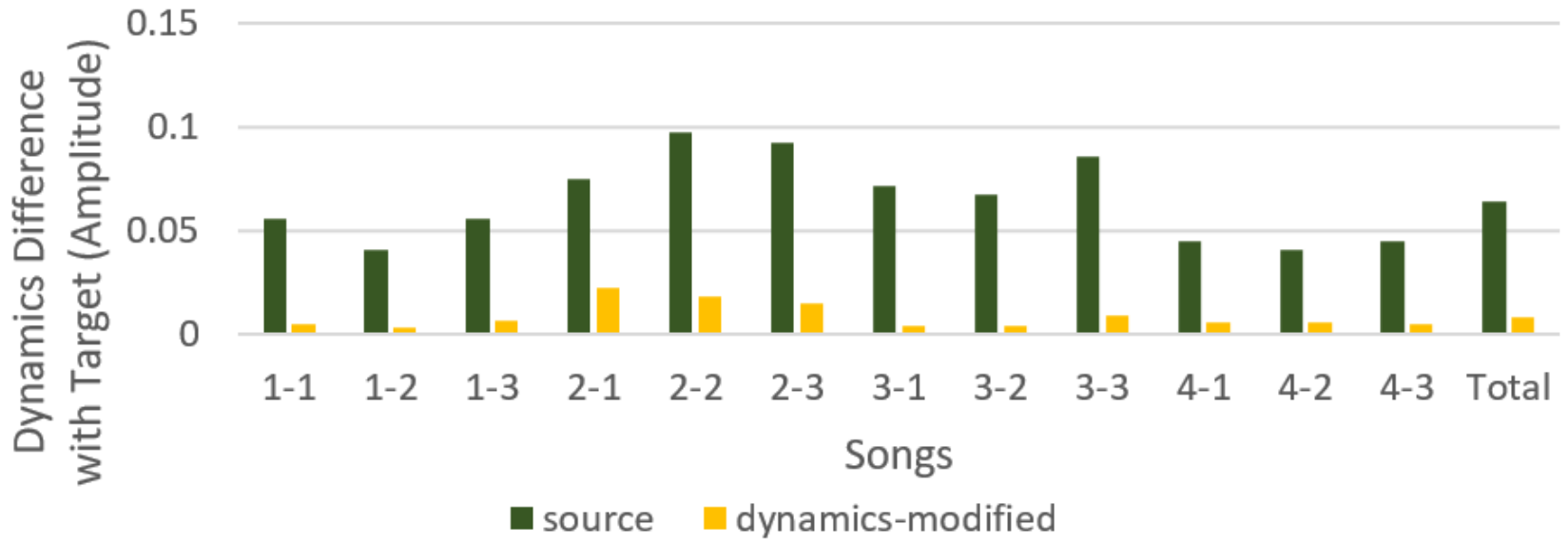
	Song 1	Song 2	Song 3	Song 4
Gender	female	male	male	male
No. of source	3	3	3	3
Remarks	high pitch English	low pitch English	swing rhythm Korean	swing rhythm Korean



Evaluation

- Dynamics alignment

	Song 1	Song 2	Song 3	Song 4
Gender	female	male	male	male
No. of source	3	3	3	3
Remarks	high pitch English	low pitch English	swing rhythm Korean	swing rhythm Korean



Audio Examples

	source	target	result
let it go			
cherry blossom ending			

More examples are available on https://seyong92.github.io/ICASSP2018

Summary

- Proposed a method to transfer vocal expressions from one voice to another in terms of tempo, pitch and dynamics without any additional information

- Showed the proposed method effectively transformed the source voices so that they mimic singing skills from the target voice

Future Plan

- The limitation of this work is that the target voice must be available

- A possible solution is to model a target singer model (e.g. singing synthesizer with natural expressions) and generate a target example using melody and lyrics information extracted from the source voice

- Improve the audio quality using other time-scale/pitch modification algorithms

Thank you