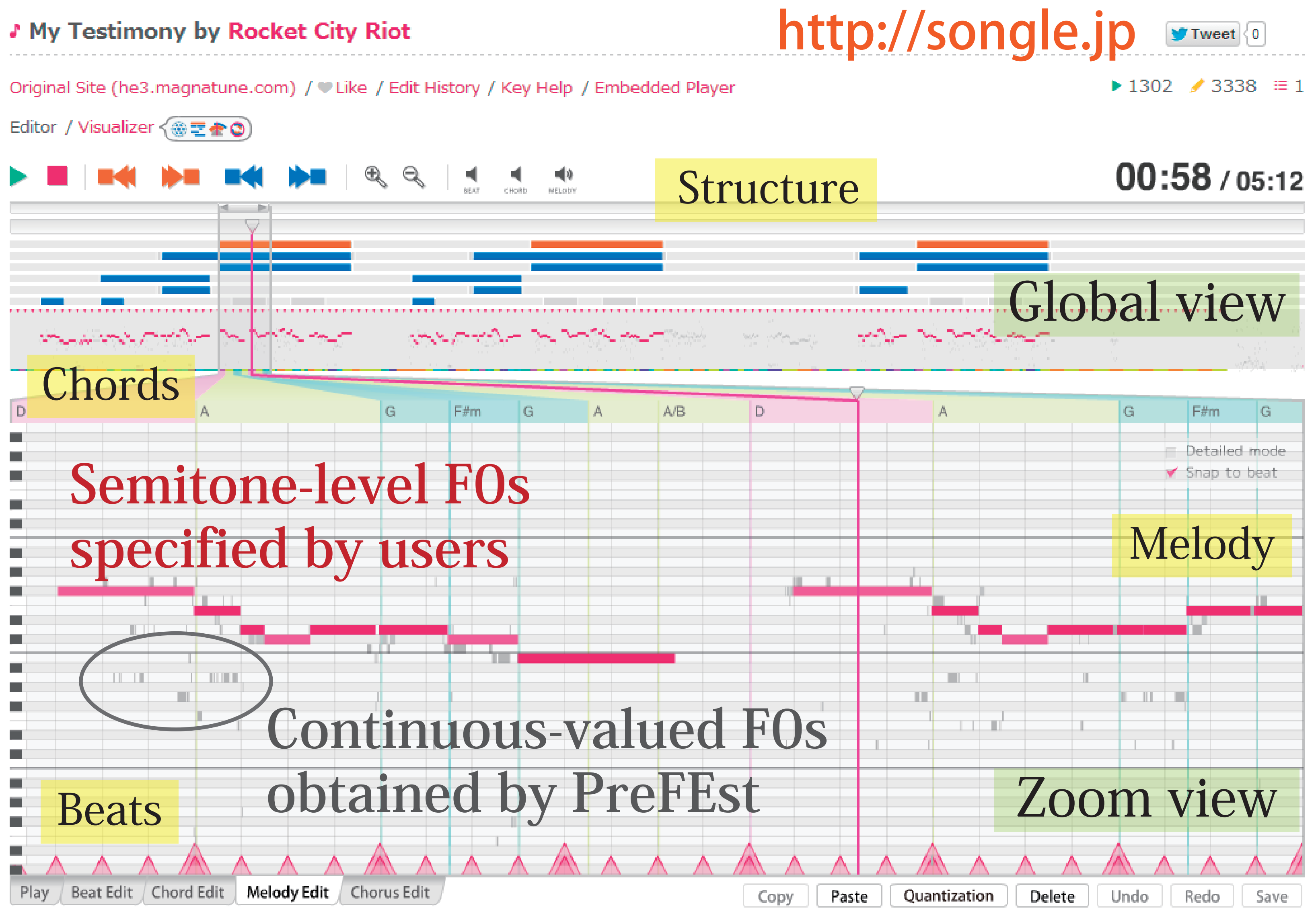
Cultivating Vocal Activity Detection for Musical Audio Signals in a Circulation-type Crowdsourcing Ecosystem

Kazuyoshi Yoshii   Hiromasa Fujihara   Tomoyasu Nakano   Masataka Goto   (AIST, Japan)

Songle: A crowdsourcing-based Web service for active music listening

Four major types of musical elements can be visualized and sonificated on the Web-browser-based music player



♪ My Testimony by Rocket City Riot

http://songle.jp

Original Site (he3.magnatune.com) / Like / Edit History / Key Help / Embedded Player

Editor / Visualizer

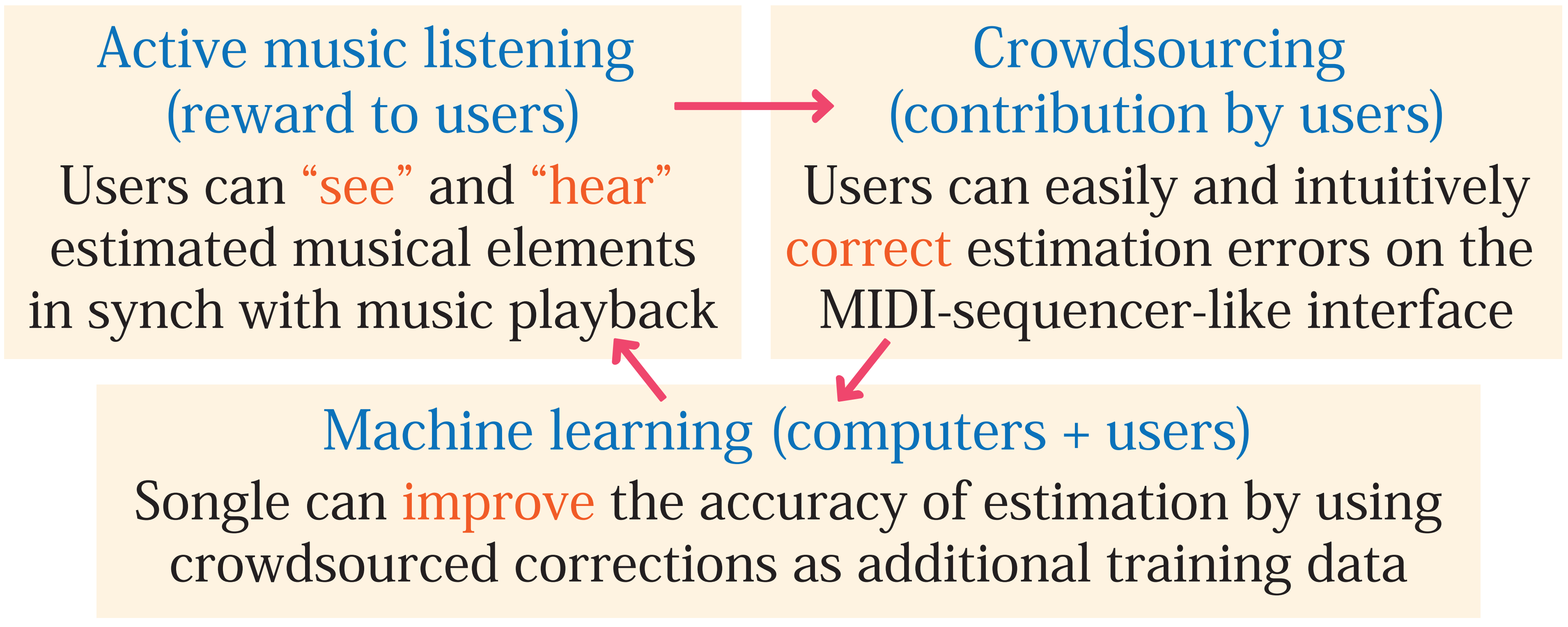
00:58 / 05:12

Play | Beat Edit | Chord Edit | Melody Edit | Chorus Edit

Copy | Paste | Quantization | Delete | Undo | Redo | Save

Melody edit mode: The trajectory of vocal F0s is visualized

Songle can automatically estimate the beats, chords, main melodies (vocal F0s and regions), and musical structures (repeated sections) of audio recordings (mp3, Youtube, etc.) existing on the Web



Active music listening (reward to users)

Users can "see" and "hear" estimated musical elements in synch with music playback

Crowdsourcing (contribution by users)

Users can easily and intuitively correct estimation errors on the MIDI-sequencer-like interface

Machine learning (computers + users)

Songle can improve the accuracy of estimation by using crowdsourced corrections as additional training data

Contribution-amplified mechanism (smarter than Wikipedia)

If users correct estimation errors of a few songs, Songle learns to analyze all other songs more accurately

Vocal activity detection based on crowdsourcing and machine learning

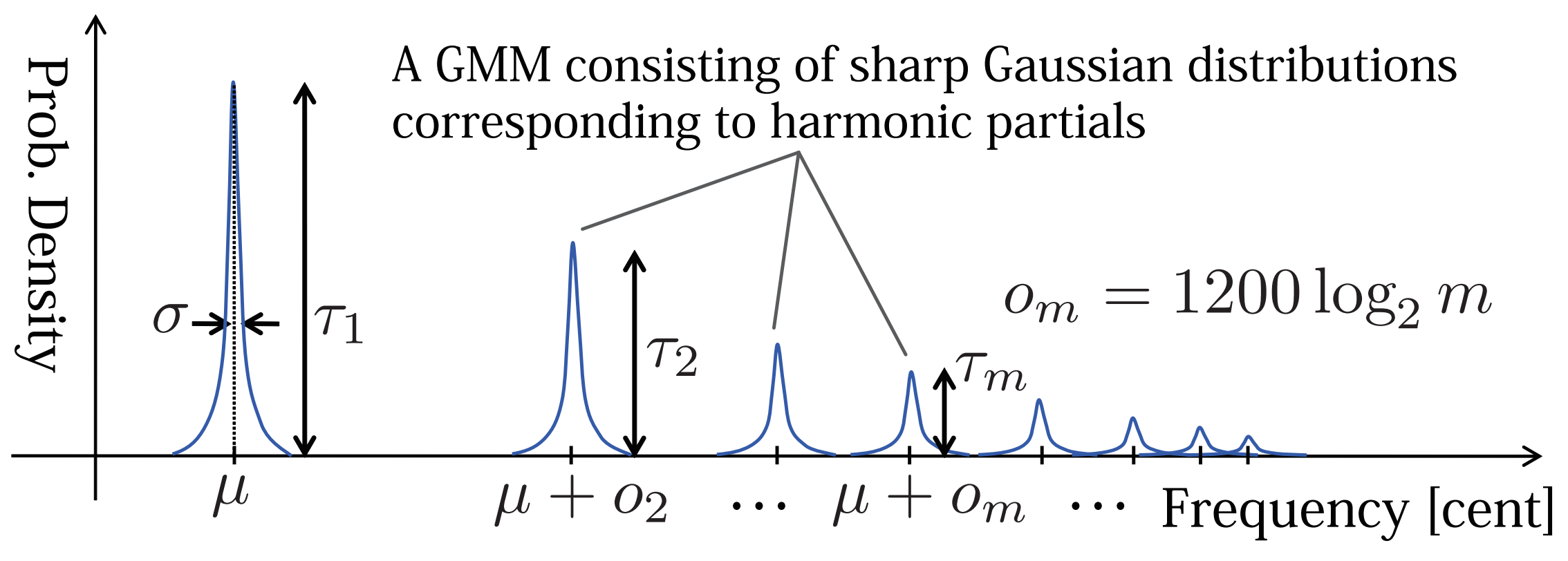
Main-melody information corrected by users is leveraged for improving VAD based on main-melody extraction

Songle currently uses a promising VAD method [Fujihara 2010] consisting of the following three steps:

Estimating F0s of main melody

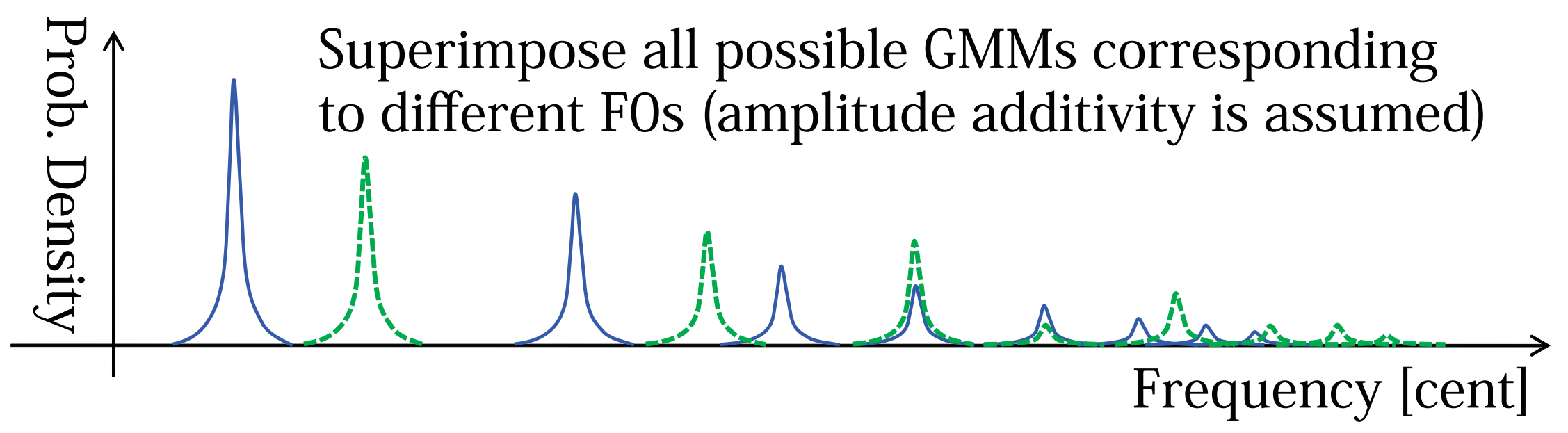
A predominant-F0 estimation method called PreFEst [Goto 2004] fits a probabilistic model to an observed amplitude spectrum

A probabilistic model of a single sound



$$p(x|\mu, \boldsymbol{\tau}) = \sum_{m=1}^{M} \tau_m \mathcal{N}\left(x|\mu + 1200 \log_2 m, \sigma^2\right)$$

A probabilistic model of multiple sounds



Prob. Density
Superimpose all possible GMMs corresponding to different F0s (amplitude additivity is assumed)
Frequency [cent]

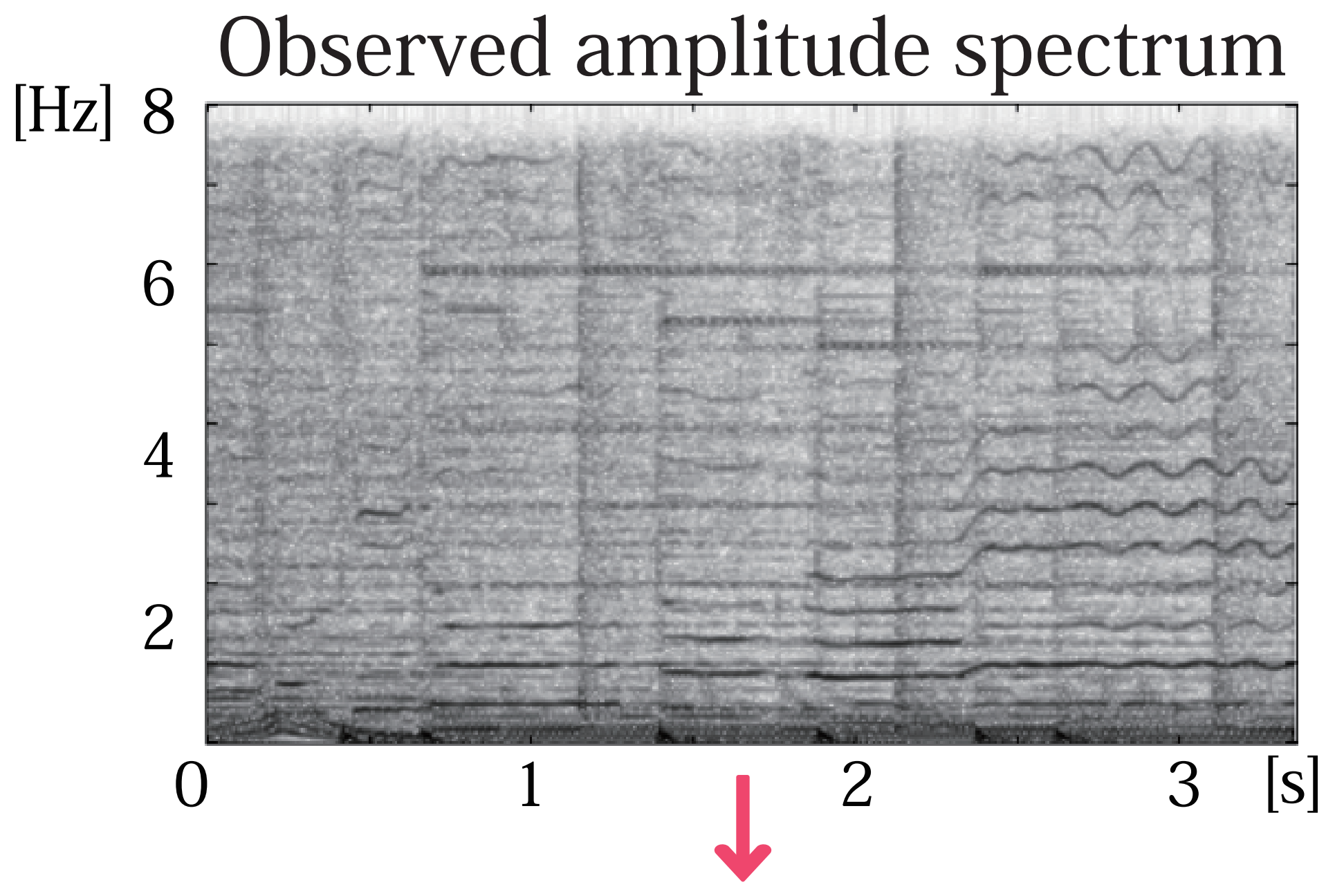
$$p(x|\boldsymbol{\tau}, p(\mu)) = \int \boldsymbol{p(\mu)} p(x|\mu, \boldsymbol{\tau}) d\mu \xrightarrow{} F0 = \operatorname*{argmax}_{\mu} p(\mu)$$

Probability density function of the F0 (F0's PDF)

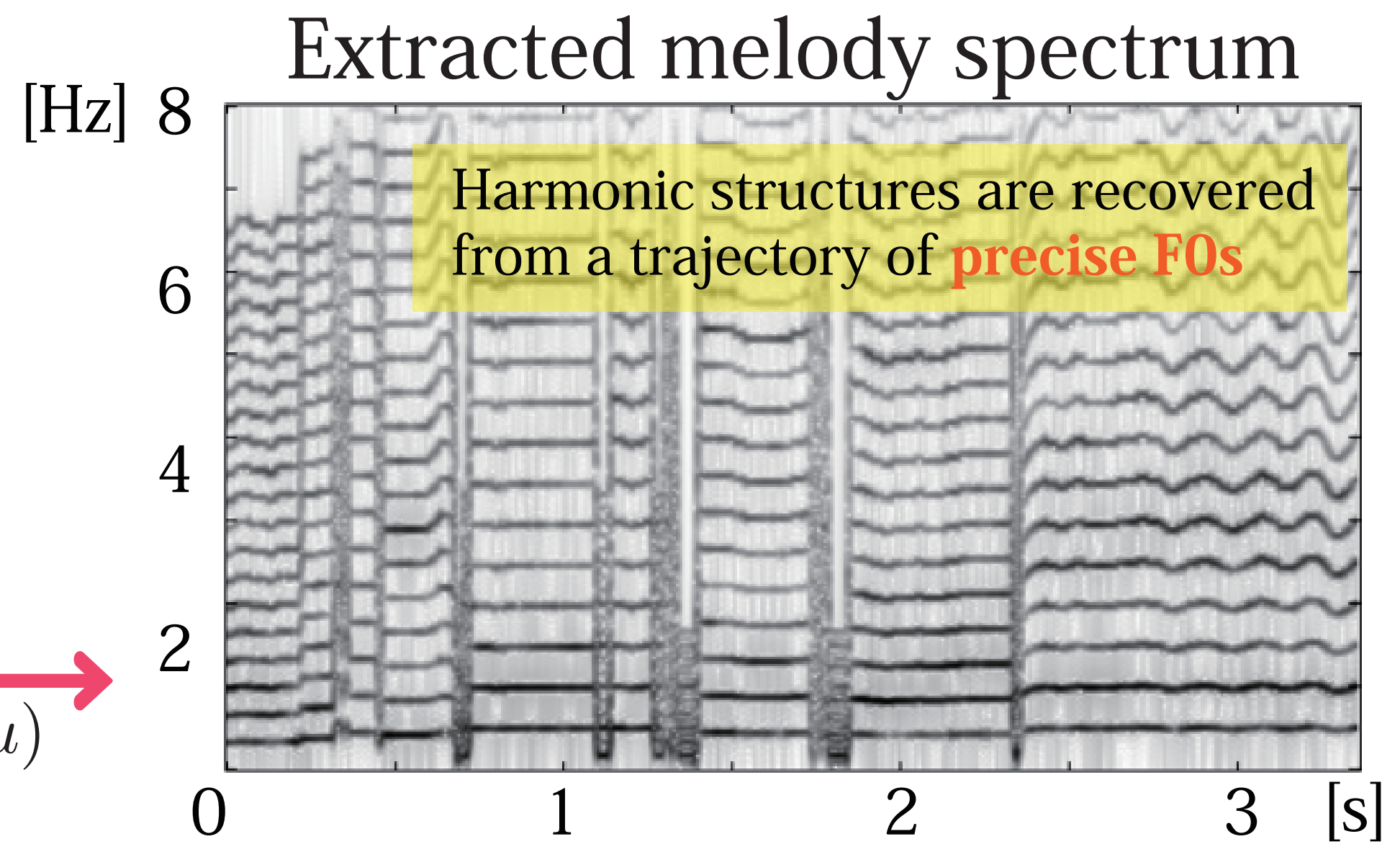
Extracting main melody

Melody signals are resynthesized from predominant harmonic structures by using a sinusoidal synthesis method

Observed amplitude spectrum

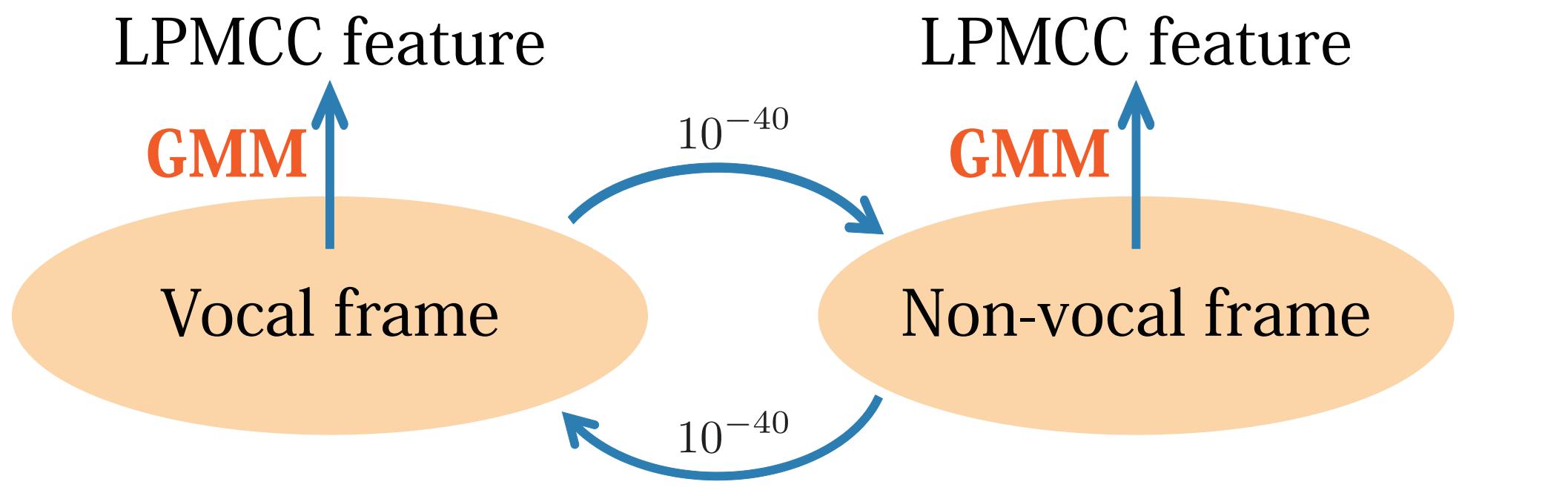


Extracted melody spectrum



Detecting vocal/non-vocal regions

Acoustic features are extracted from melody signals and discriminated into vocal/non-vocal classes by using a hidden Markov model (HMM)



$$\hat{S} = \operatorname*{argmax}_{S} \sum_{t} \left(\log p(\boldsymbol{x}_t|s_t) + \log p(s_{t+1}|s_t)\right)$$

We aim to leverage crowdsourced data
vocal F0s (quantized at a semitone level)
vocal regions
for re-training the vocal/non-vocal GMMs

Technical issue:
How to recover raw "precise" F0s from quantized "semitone-level" F0s?

Estimating F0s with prior knowledge

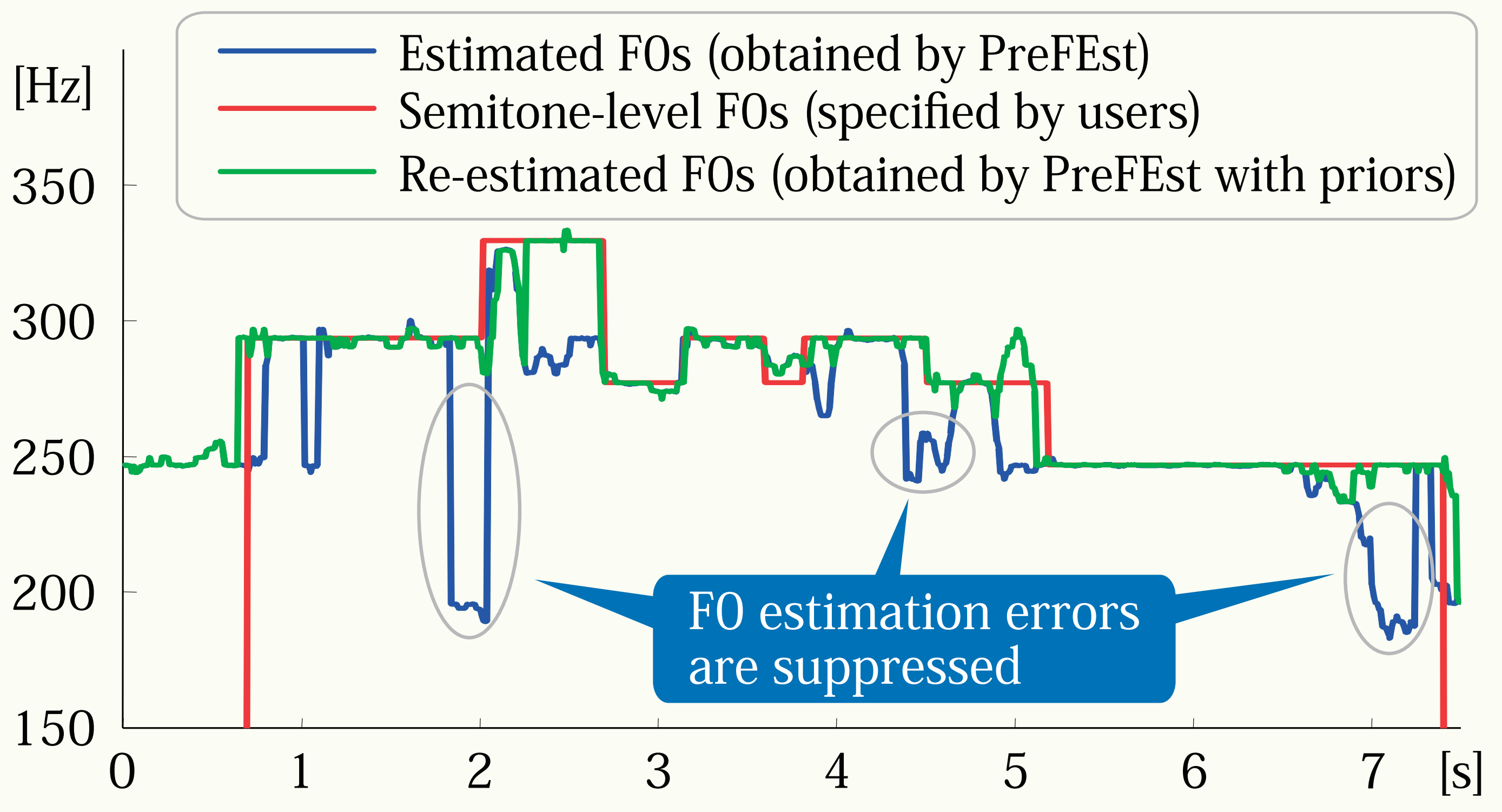
PreFEst can take into account a rough F0 estimate

The true F0 is assumed to be Gaussian distributed around the semitone-level F0 (prior knowledge)

$$p(p(\mu)) \propto \exp\left(-\beta_\mu \mathcal{D}_{\mathrm{KL}}(p_0(\mu)|p(\mu))\right)$$

$$p_0(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

→ **Maximum-a-posteriori (MAP) estimation**



F0 estimation errors are suppressed

Experimental evaluation using Songle data

Baseline **66.6%**

Prediction Annotation	Vocal	Non-vocal
Vocal	**12,347 [s]**	4,086
Non-vocal	2,252	**268**

Vocal/non-vocal GMMs were trained from 100 pieces of RWC Music Database: Popular Music [Goto 2002]

The VAD method was evaluated for 100 pieces on Songle that have been most frequently annotated

Vocal/non-vocal GMMs were trained from 100 RWC pieces + 90 Songle pieces and the VAD method was evaluated for the rest 10 pieces (10-fold cross validation)

Without F0 estimation **67.6%**

Prediction Annotation	Vocal	Non-vocal
Vocal	**12,452**	3,981
Non-vocal	2,152	**368**

With F0 estimation **69.6%**

Prediction Annotation	Vocal	Non-vocal
Vocal	**12,505**	3,928
Non-vocal	1,827	**693**

Note that non-vocal frames available for evaluation were much fewer than vocal frames available for evaluation (issue of the Songle editing interface)

We plan to incorporate this crowdsourcing-based self-improvement framework into various kinds of music analysis such as beat tracking, chord recognition, and auto tagging