MIXTURE OF GAUSSIAN PROCESS EXPERTS FOR PREDICTING SUNG MELODIC CONTOUR WITH EXPRESSIVE DYNAMIC FLUCTUATIONS

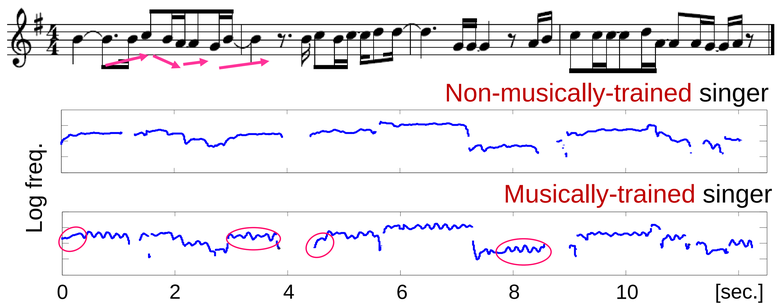
Yasunori Ohishi[1], *Daichi Mochihashi*[2], *Hirokazu Kameoka*[1], *Kunio Kashino*[1]

[1]NTT Communication Science Laboratories, NTT Corporation, Japan, [2]The Institute of Statistical Mathematics

The goal is to build a generative model that can characterize the singing style of a singer in the sung melodic contours, i.e., F_0 contours, and predict the F_0 contour that reflects the singing style for an arbitrary musical score, using a mixture of Gaussian process experts (MoGPEs). The experimental results showed that our model is promising for predicting the F_0 contour for a given musical score. An application of this work is singing style conversion which automatically converts a poor singing voice into an expressive singing voice.

Motivation

- Dynamic fluctuations in a sung melodic contour (F_0 contour)

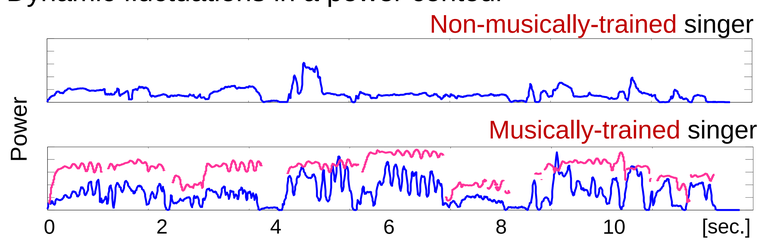


- *Vibrato*: quasi-periodic variation
- *Portamento*: gradual sliding of pitch

} Related to singing skill, style, and expression [1-3]

Musically-trained singer can skillfully control the F_0 fluctuations and expressively sing a melody based on the personal style

- Dynamic fluctuations in a power contour



- Rise and fall of power according to the pitch of musical notes
- Fluctuations correlated with *vibrato* in the F_0 contour

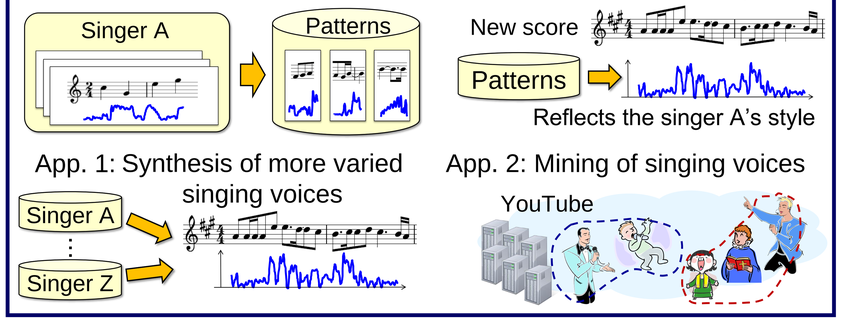
Musically-trained singer can also control the power of singing voice

Goal

- Prediction of the F_0 contour with the expressive dynamic fluctuations for an arbitrary musical note sequence

(1) Learning of the F_0 patterns for musical note sequences

(2) Prediction of the F_0 contour for a new musical note sequence

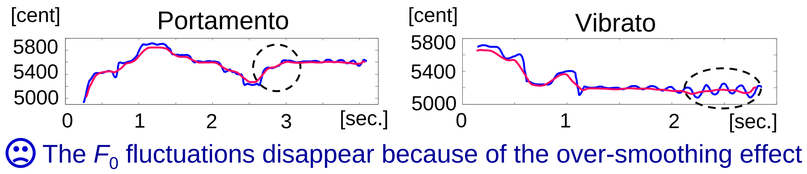


App. 1: Synthesis of more varied singing voices

App. 2: Mining of singing voices

Previous works

- Hidden Markov model (HMM) [4-6]
  - Difficult to characterize the continuous and rapid changes in the fluctuations, because its hidden-state space is discrete
  - Context-dependent decision-tree clustering causes an over-smoothing effect of the F_0 contours

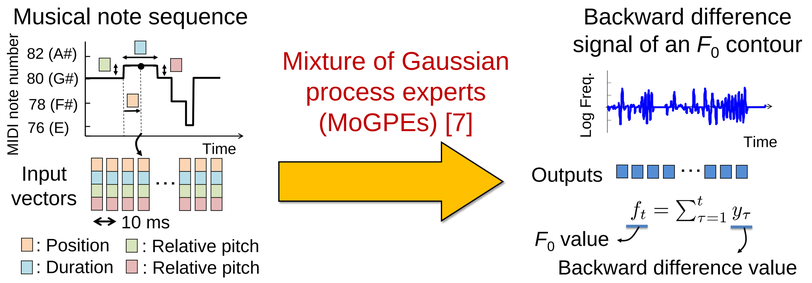


☹ The F_0 fluctuations disappear because of the over-smoothing effect

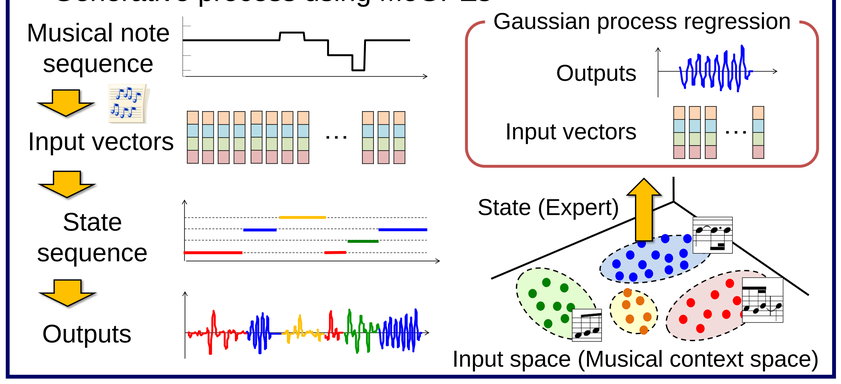
Characterize the F_0 fluctuations using Gaussian process regression which is a sophisticated machine learning algorithm

Statistical modeling of F_0 contour

- Use the singing voice signals synchronized with melodies



- Generative process using MoGPEs



Gaussian process regression (GPR) [8]

- Probability distribution of a contour

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K} + \eta^2 \boldsymbol{I})$$

Outputs (Contour) $\boldsymbol{y}$ — Mean $\boldsymbol{0}$ — Covariance matrix $\boldsymbol{K} + \eta^2 \boldsymbol{I}$

Calculation: Outputs, Input vectors

Kernel matrix

20	10	5	3
10	20	⋱	⋮
5	⋱	⋱	10
3	⋯	10	20

$+\eta^2$ Noise

Unit matrix

1	0	0	0
0	1	⋱	⋮
0	⋱	⋱	0
0	⋯	0	1

⇒ Kernel matrix: control the dynamics of a contour

- Multiple kernel learning [9]

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, w^2(\psi_1 \boldsymbol{K}_1 + \psi_2 \boldsymbol{K}_2 + \ldots, +\psi_M \boldsymbol{K}_M) + \eta^2 \boldsymbol{I})$$

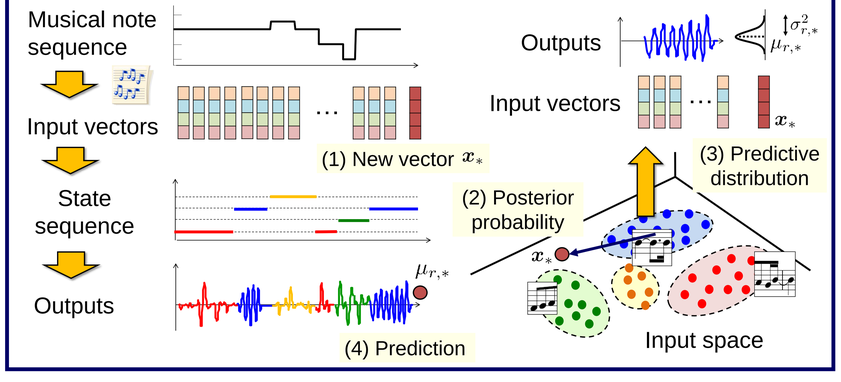
Gaussian kernel (Smooth function) — Periodic kernel (Periodic function)

⇒ Characterize the dynamics of an contour using weights $\psi_1, \psi_2, \ldots, \psi_M$

Parameter estimation and prediction

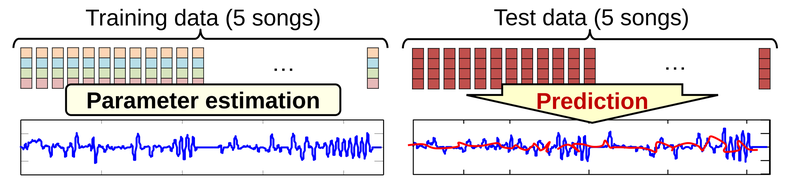
MCMC-EM algorithm [10]: estimate a state sequence, parameters of GPRs, and parameters of Gaussian distributions in the input space

Prediction: use the mean value of the predictive distribution

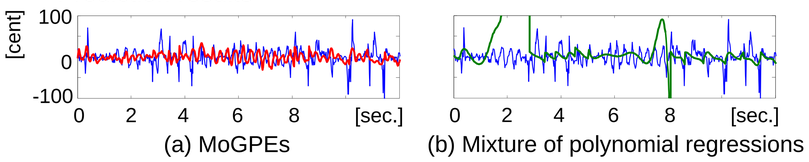


Evaluation

- Predictive performance of the output for a new input vector
  - Used the F_0 contours and the MIDI signals of melodies annotated manually in the AIST annotation [11]
  - Training data: Song No. 38, 39, 42, 44, and 45 (649.3 sec.)
  - Test data: Song No. 46, 64, 72, 74, and 76 (625.6 sec.)



- Prediction results



- Evaluation measure: root mean square error (RMSE)

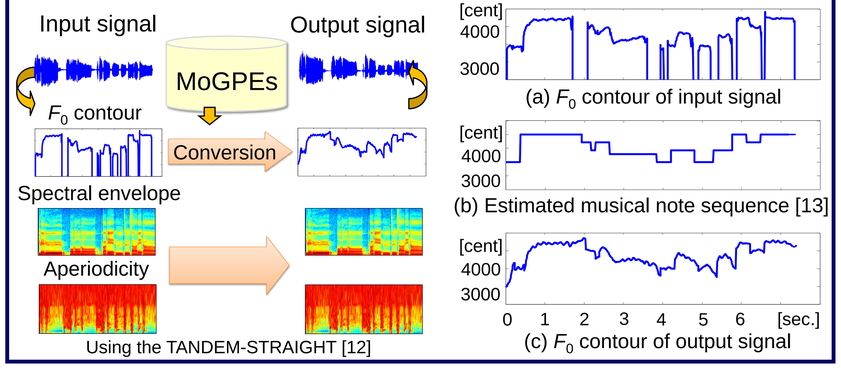
$$\mathrm{RMSE} = \sqrt{\sum_{t=1}^{T_t} (y_t - \mu_{*,t})^2 / T_t}$$

y_t : Actual output $\quad T_t$: Signal length

$\mu_{*,t}$: Predictive mean

Num. of experts	5	10	20	30	40	50
MoPRs	65.6	71.8	59.7	73.1	79.6	56.6
MoGPEs	**26.4**	**25.0**	**24.0**	**23.9**	**23.1**	**22.3**

Singing style conversion



Appendix

Joint distribution

$$p(\{\boldsymbol{x}_t, y_t\}_{t=1}^T, \{z_t\}_{t=1}^T, \{\theta_r^{\boldsymbol{x}}\}_{r=1}^R, \{\theta_r^{\boldsymbol{x}}\}_{r=1}^R | \Omega) = \prod_{r=1}^R \left[p(\theta_r^{\boldsymbol{x}}|\Omega) p(X_r|\theta_r^{\boldsymbol{x}}) p(\boldsymbol{y}_r|X_r, \theta_r^{\mathrm{GP}}, \Omega) \right] \times p(\{z_t\}_{t=1}^T|\Omega)$$

$$\begin{cases} p(\theta_r^{\boldsymbol{x}}|\Omega) = \mathcal{N}(\boldsymbol{\mu}_r; \boldsymbol{m}_0, \boldsymbol{\Sigma}_r/\beta_0) \mathcal{W}(\boldsymbol{\Sigma}_r^{-1}; \boldsymbol{W}_0, \nu_0), \\ p(\boldsymbol{y}_r|X_r, \boldsymbol{\theta}_r^{\mathrm{GP}}, \Omega) = \mathcal{GP}\left(\boldsymbol{y}_r; \boldsymbol{0}, \boldsymbol{K}_r + \eta_r^2 \boldsymbol{I}_r\right), \end{cases} \quad \begin{array}{l} p(X_r|\theta_r^{\boldsymbol{x}}) = \mathcal{N}(X_r; \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) \\ p(\{z_t\}_{t=1}^T|\Omega) = \dfrac{\Gamma(\alpha)}{\Gamma(\alpha+T)} \displaystyle\prod_{r=1}^R \dfrac{\Gamma(T_r + \alpha/R)}{\Gamma(\alpha/R)} \end{array}$$

Predictive distribution

$$p(y_*|\mathcal{D}, \boldsymbol{x}_*, \Theta, \Omega) = \sum_{r=1}^R p(y_*|\boldsymbol{y}_r, X_r, \boldsymbol{x}_*, z_* = r, \theta_r^{\mathrm{GP}}) p(z_* = r|\boldsymbol{x}_*, \theta_r^{\boldsymbol{x}}) = \mathcal{N}(y_*; \mu_*, \sigma_*^2)$$

$$\mu_* = \sum_{r=1}^R c_r \boldsymbol{k}_{r,*}^{\mathrm{T}} (\boldsymbol{K}_r + \eta_r^2 \boldsymbol{I}_r)^{-1} \boldsymbol{y}_r, \quad \sigma_*^2 = \sum_{r=1}^R c_r^2 (k_r(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_{r,*}^{\mathrm{T}} (\boldsymbol{K}_r + \eta_r^2 \boldsymbol{I}_r)^{-1} \boldsymbol{k}_{r,*}), \quad c_r = p(z_* = r|\boldsymbol{x}_*, \theta_r^{\boldsymbol{x}})$$

References

[1] J. Sundberg, Northern Illinois University Press, 1987.
[2] T. Saitou *et al.*, *Speech Communication*, vol. 45, no. 3-4, pp. 405–417, 2005.
[3] L. Regnier, Ph.D. thesis, IRCAM / UPMC in Paris, France, 2013.
[4] T. Yoshimura *et al.*, in *Proc. EUROSPEECH 1999*.
[5] H. Zen *et al.*, *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
[6] T. Nose *et al.*, in *Proc. INTERSPEECH 2013*, pp. 378–382.
[7] E. Meeds *et al.*, in *Proc. NIPS 2006*.
[8] C. E. Rasmussen *et al.*, MIT Press, Cambridge, Mass, USA, 2006.
[9] K. Yoshii *et al.*, in *Proc. ICASSP 2013*, pp. 463–467.
[10] C. Andrieu *et al.*, *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.
[11] M. Goto, in *Proc. ISMIR 2006*.
[12] H. Kawahara *et al.*, in *Proc. ICASSP 2008*, pp. 3933–3936.
[13] Y. Ohishi *et al.*, in *Proc. INTERSPEECH 2012*.