Generative modeling of speech F_0 contours

H. Kameoka[1,2], K. Yoshizato[1], T. Ishihara[1], Y. Ohishi[2], K. Kashino[2], S. Sagayama[1*]

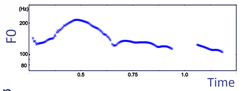
[1] The University of Tokyo, [2] NTT Communication Science Laboratories (*Currently with National Institute of Informatics)

Abstract: This paper reviews our ongoing work on generative modeling of speech fundamental frequency (F_0) contours for estimating prosodic features from raw speech data [3]-[6]. Proposed F_0 contour model is formulated by translating "Fujisaki model" into probabilistic model described as discrete-time stochastic process. The motivation behind this formulation is:

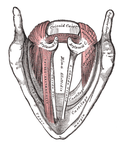
① to derive a general parameter estimation framework for Fujisaki model, allowing for introduction of powerful statistical methods

② to construct an automatically trainable version of Fujisaki model so that in future it can be incorporated into statistical-model-based text-to-speech synthesis systems

1. Introduction



- Fundamental frequency (F_0) contour
  - time course of frequency of vocal fold vibration (Manifestation of physical movement of thyroid cartilage)
  - contains various types of non-linguistic information
    - Speaker's identity, intention, attitude, mood, etc.
  - modeling and analyzing F_0 contours can be potentially useful for many speech applications in which prosodic information plays a significant role. (In speech synthesis, one challenge is to create a natural -sounding pitch contour for the utterance as a whole.)

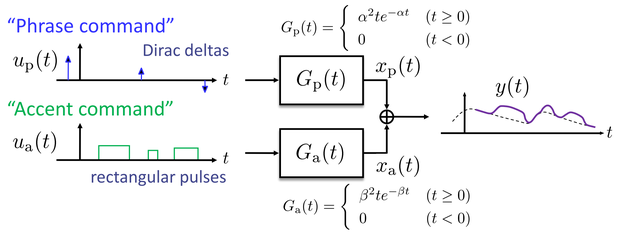


- "Fujisaki model" [1]
  - F_0 contour typically consists of two components:
    - **phrase components** (long term pitch variations over the duration of prosodic units), and
    - **accent components** (short term pitch variations in accented syllables).

    "Fujisaki model" is a well-founded mathematical model that describes F0 contour as the sum of these two components (See below for details).
  - A notable feature of this model is that the model parameters are associated with physiologically and linguistically meaningful quantities.
  - However, automatic estimation of Fujisaki model parameters from raw F_0 contour has been a difficult task...
- Aim of this paper
  - **translate Fujisaki model into a probabilistic model (stochastic process)**
  - motivation for this is twofold:
    - to provide general parameter estimation framework utilizing powerful statistical inference methods, and
    - to construct an automatically trainable version of Fujisaki model that can potentially be combined with statistical model for speech synthesis.

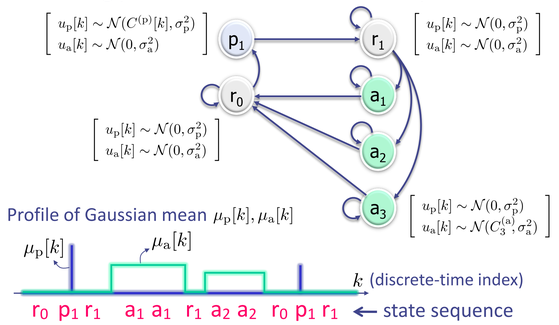
2. Original Fujisaki model [1]

- Fujisaki model assumes:
  - F_0 contour (log scale) = Phrase component + Accent component + Base value (lower bound for F_0)
    - contributions associated with forward-backward translation and rotation of thyroid cartilage, respectively
    - outputs of different 2nd order critically damped systems



3. Generative model of speech F_0 contours

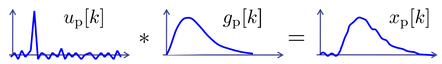
- Generative modeling of phrase and accent commands
  - Requirements
    - Phrase command is impulse
    - Accent command is rectangular pulse
    - Two commands do not overlap each other
  - Propose introducing path-restricted HMM for modeling phrase & accent command sequence



state transition: $s_k | s_{k-1} \sim \pi_{s_{k-1}, s_k}$

state emission: $\begin{bmatrix} u_p[k] \\ u_a[k] \end{bmatrix} | s_k \sim \mathcal{N}\left(\begin{bmatrix} \mu_p[k] \\ \mu_a[k] \end{bmatrix}, \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \sigma_a^2 \end{bmatrix} \right)$

- Generative modeling of F_0 contour
  - Relationship between $u_p[k]$ and $x_p[k]$



$$\underbrace{\begin{bmatrix} u_p[1] \\ u_p[2] \\ \vdots \\ u_p[K] \end{bmatrix}}_{\boldsymbol{u}_p} = \underbrace{\begin{bmatrix} f_0^{(p)} & & & O \\ f_1^{(p)} & f_0^{(p)} & & \\ f_2^{(p)} & f_1^{(p)} & f_0^{(p)} & \\ & \ddots & \ddots & \ddots \\ O & & f_2^{(p)} & f_1^{(p)} & f_0^{(p)} \end{bmatrix}}_{\boldsymbol{F}_p} \underbrace{\begin{bmatrix} x_p[1] \\ x_p[2] \\ \vdots \\ x_p[K] \end{bmatrix}}_{\boldsymbol{x}_p}$$

  - Relationship between $u_a[k]$ and $x_a[k]$

$$\boldsymbol{u}_a = \boldsymbol{F}_a \boldsymbol{x}_a$$

  - Probability density function of $x_p[k]$ and $x_a[k]$

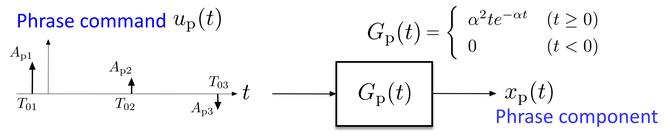
$$\boldsymbol{x}_p | s_1, \ldots, s_K \sim \mathcal{N}(\boldsymbol{F}_p^{-1}\boldsymbol{\mu}_p, \boldsymbol{F}_p^{-1}\boldsymbol{\Sigma}_p\boldsymbol{F}_p^{-\mathsf{T}})$$
$$\boldsymbol{x}_a | s_1, \ldots, s_K \sim \mathcal{N}(\boldsymbol{F}_a^{-1}\boldsymbol{\mu}_a, \boldsymbol{F}_a^{-1}\boldsymbol{\Sigma}_a\boldsymbol{F}_a^{-\mathsf{T}})$$

  - Probability density function of $y[k]$

$$\boldsymbol{y} | \boldsymbol{s} \sim \mathcal{N}(\boldsymbol{F}_p^{-1}\boldsymbol{\mu}_p + \boldsymbol{F}_a^{-1}\boldsymbol{\mu}_a + \mu_b \mathbf{1}, \boldsymbol{F}_p^{-1}\boldsymbol{\Sigma}_p\boldsymbol{F}_p^{-\mathsf{T}} + \boldsymbol{F}_p^{-1}\boldsymbol{\Sigma}_p\boldsymbol{F}_p^{-\mathsf{T}} + \boldsymbol{\Sigma}_b)$$

Discretized Fujisaki model

- Discretization of phrase control mechanism



  - 2nd order critically damped filter (Laplace domain): $\alpha^2/(s+\alpha)^2$
  - Inverse filter (Laplace domain): $(s+\alpha)^2/\alpha^2$

Backward difference approximation $s = \dfrac{1 - z^{-1}}{t_0}$ (Sampling period)

  - Inverse filter (z-domain):

$$\frac{z^{-2} - 2(\alpha t_0 + 1)z^{-1} + (\alpha t_0 + 1)^2}{\alpha^2 t_0^2}$$

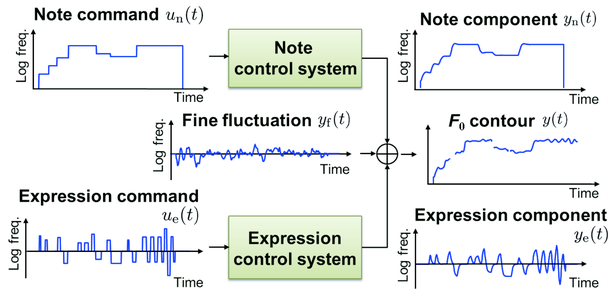
$$u_p[k] = f_0^{(p)} x_p[k] + f_1^{(p)} x_p[k-1] + f_2^{(p)} x_p[k-2]$$
where $f_0^{(p)} = (\psi - 1)^2$, $f_1^{(p)} = -2\psi(\psi - 1)$, $f_2^{(p)} = \psi^2$

Constrained all-pole model governed by single parameter $\psi = 1 + 1/(\alpha t_0)$

- Discretization of accent control mechanism
  - follows exactly the same as above

4. Singing voice F_0 contour modeling [6]

- Propose describing singing voice F_0 contour based on a model structurally similar to Fujisaki model



- This model can be translated into a probabilistic model in the same way as above.

5. MAP estimation

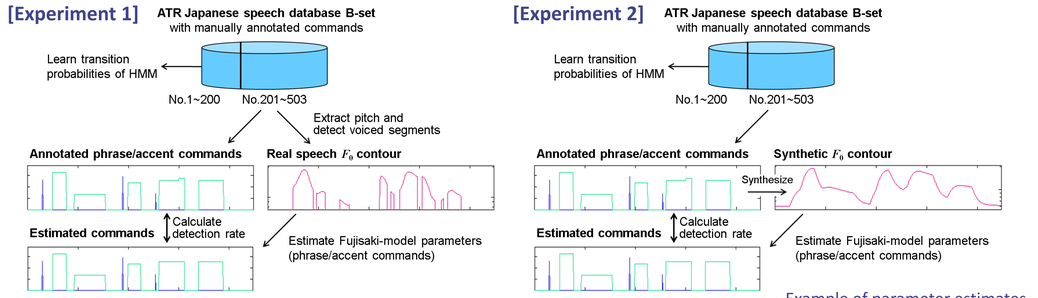
- Maximum A Posteriori estimation of Fujisaki model parameters (namely, state sequence of present HMM)

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}}{\operatorname{argmax}} \log p(\boldsymbol{s}|\boldsymbol{y}) = \underset{\boldsymbol{s}}{\operatorname{argmax}} \log p(\boldsymbol{y}|\boldsymbol{s}) p(\boldsymbol{s})$$

(observed F_0 contour)

  - Likelihood function: $p(\boldsymbol{y}|\boldsymbol{s}) = \mathcal{N}(\boldsymbol{F}_p^{-1}\boldsymbol{\mu}_p + \boldsymbol{F}_a^{-1}\boldsymbol{\mu}_a + \mu_b \mathbf{1}, \boldsymbol{F}_p^{-1}\boldsymbol{\Sigma}_p\boldsymbol{F}_p^{-\mathsf{T}} + \boldsymbol{F}_p^{-1}\boldsymbol{\Sigma}_p\boldsymbol{F}_p^{-\mathsf{T}} + \boldsymbol{\Sigma}_b)$
  - Prior distribution: $p(\boldsymbol{s}) = \pi_{s_1} \prod_{k=2}^{K} \pi_{s_{k-1}, s_k}$
- Developed Iterative algorithm for estimating $\hat{\boldsymbol{s}}$ based on EM algorithm
  - See [3]-[6] for details.

5. Evaluation using real/synthetic data



- Results

[Experiment 1]

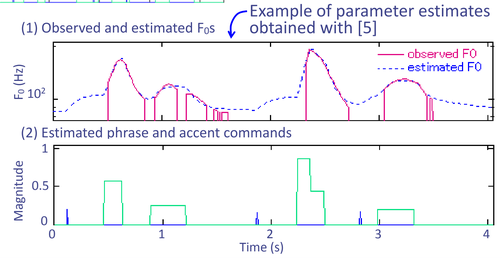
	DR (%)	RMSE
Conventional [2]	68.8	0.1719
Proposed1 [4]	69.7	0.1372
Proposed2 [5]	69.5	0.0611

[Experiment 2]

	DR (%)
Conventional [2]	72.6
Proposed1 [4]	76.8
Proposed2 [5]	83.4

DR: Detection rate
"RMSE": RMSE between observed F_0 contour and optmized model

Example of parameter estimates obtained with [5]



References

[1] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," in Vocal Physiology: Voice Production, Mechanisms and Functions, O. Fujimura, Ed. Raven Press, 1988.

[2] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002), 2002, pp. 509–512.

[3] H. Kameoka, J. Le Roux, and Y. Ohishi, "A statistical model of speech F0 contours," in Proceedings of the 2010 ISCA Workshop on Statistical and Perceptual Audition (SAPA 2010), 2010, pp. 43–48.

[4] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Statistical approach to fujisaki-model parameter estimation from speech signals and its quantitative evaluation," in Proceedings of Speech Prosody 2012, 2012, pp. 175–178.

[5] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Hidden markov convolutive mixture model for pitch contour analysis of speech," in Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), 2012.

[6] Y. Ohishi, H. Kameoka, D. Mochihashi, K. Kashino, "A stochastic model of singing voice F0 contours for characterizing expressive dynamic components," in Proc. Interspeech 2012, 2012.