# Context-Dependent Additive Log F0 Model for HMM-Based Speech Synthesis

*Heiga Zen, Norbert Braunschweiler*

{heiga.zen,norbert.braunschweiler}@crl.toshiba.co.uk

*Speech Technology Group, Toshiba Research Europe Ltd. Cambridge Research Laboratory, 208 Cambridge Science Park, Cambridge, UK*

TOSHIBA
Leading Innovation

## 1. Introduction

### HMM-based speech synthesis [Yoshimura;'99]

· Typical system
- Spectrum, excitation, & duration are modeld by HMMs
- Speech param. trajectories are generated from HMMs
- Spectrum: (Mel) Cepstrum or LSP
- Excitation: Log F0 with V/UV

· Problems
- Spectrum & F0 are final results of speech production
  ⇒ There are unobservable features & structures
- Spectrum: Articulatory features
  * Articulators determine the resonance characteristics
  ⇒ Acoustic & articulatory joint modeling [Ling;'08]
- Excitation (F0): Additive structure [Fujisaki;'08]
  * Composed as superposition of multi-level F0 contours
  * Standard HMM cannot capture this additive nature

### Additive acoustic models

· Gaussian additive model [Nankaku;'08]
- Additive components are modeled by Gaussians
- Not applicable yet for F0 due to huge computation

· Addtively boosted HMMs [Qian;'08]
- Sequentially boosting HMMs to minimize F0 RMSE
- Unable to extract underlying additive nature

· Bias additive model (proposed)
- Additive components are modeled by bias terms
- Computationally less expensive than convolutional one

## 2. Definition of bias additive model

1. Observations: generated as the sum of additive comps.

$$o_t = \sum_{i=1}^{P} o_t^{(i)} \begin{cases} \text{Component 1} & o_t^{(1)} \\ + & \\ \vdots & \\ + & \\ \text{Component } P & o_t^{(P)} \end{cases}$$

$t$ : frame
$o_t$ : observation at $t$
$o_t^{(i)}$ : $i$-th additive component

2. 1,...,$P$−1-th comps. are bias, $P$-th comp. is Gaussian

$$o_t^{(i)} \sim \mathcal{N}\left(\lambda_{m_t}^{(i)} \mu_{m_t}^{(i)}, 0\right) = \lambda_{m_t}^{(i)} \mu_{m_t}^{(i)}$$

$$o_t^{(P)} \sim \mathcal{N}\left(0, \sigma_{m_t}^2\right)$$

$m_t$ : model at $t$
$\mu_{m_t}^{(i)}$ : bias term
$\lambda_{m_t}^{(i)}$ : scaling factor
$\sigma_{m_t}^2$ : variance

⇓

Observation distribution

$$o_t \sim \mathcal{N}\left(\nu_{m_t}, \sigma_{m_t}^2\right), \quad \nu_{m_t} = \sum_{i=1}^{P-1} \lambda_{m_t}^{(i)} \mu_{m_t}^{(i)}$$

Component distributions

$$\begin{cases} \text{Component 1} & o_t^{(1)} = \lambda_{m_t}^{(1)} \mu_{m_t}^{(1)} \\ + & + \\ \vdots & \vdots \\ + & + \\ \text{Component } P-1 & \vdots \\ + & + \\ \text{Component } P & o_t^{(P)} \sim \mathcal{N}\left(0, \sigma_{m_t}^2\right) \end{cases}$$

⇒ Obs. distribution is almost identical to that of CAT [Gales;'00]

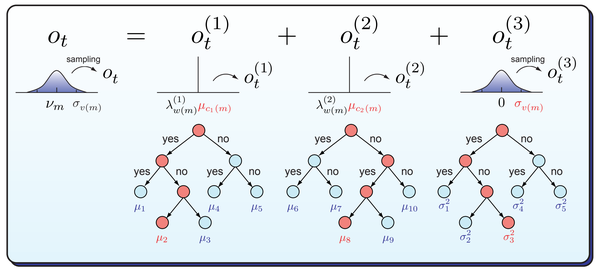
## 3. Context-dependent modeling

### Cluster-dependent decision trees

· Context-dependent acoustic modeling
- Vast # of possible contexts in ASR & TTS
- Almost impossible to cover all possible contexts
  ⇒ Decision tree-based context clustering [Odell;'95]

· Standard CAT [Gales;'00]
- Assumes all clusters have the same tree
- This assumption is unnecessary
  ⇒ Each cluster (bias & variance) can have its own tree

· Bias additive model (proposed)
- Each cluster (bias & variance) has different tree




- All bias trees are built simultaneously
  ⇒ Can extract the underlying additive structure
- All variances are tied while building bias trees
  ⇒ Reduces computational complexity

## 4. Parameter estimation

### EM algorithm-based ML estimation

$$\mathcal{Q}(\Lambda, \Lambda') = -\frac{1}{2} \sum_{m,t} \gamma_{m,t} \left( \log \left| \sigma_{v(m)}^2 \right| + \frac{(o_t - \nu_m)^2}{\sigma_{v(m)}^2} \right) + C$$

$$\nu_m = \sum_{i=1}^{P-1} \lambda_{w(m)}^{(i)} \mu_{c_i(m)} \qquad \boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_N]^\top$$

· By equating the first partial deriv. of $\mathcal{Q}$-function w.r.t. $\boldsymbol{\mu}$ to $\mathbf{0}$

$$\begin{bmatrix} g_{11} & \cdots & g_{1N} \\ \vdots & \ddots & \vdots \\ g_{N1} & \cdots & g_{NN} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix} = \begin{bmatrix} k_1 \\ \vdots \\ k_N \end{bmatrix}$$

$$g_{n_1 n_2} = \sum_{\substack{m,t,i,j \\ c_i(m)=n_1 \\ c_j(m)=n_2}} \frac{\gamma_{m,t}}{\sigma_{v(m)}^2} \lambda_{w(m)}^{(i)} \lambda_{w(m)}^{(j)} \qquad k_{n_1} = \sum_{\substack{m,t,i \\ c_i(m)=n_1}} \frac{\gamma_{m,t}}{\sigma_{v(m)}^2} \lambda_{w(m)}^{(i)} o_t$$

⇒ All bias terms can be determined simultaneously

· Rank defficient ⇒ Least square solution

· Other parameters (variances & scaling factors) can be estimated in the same way as the standard CAT

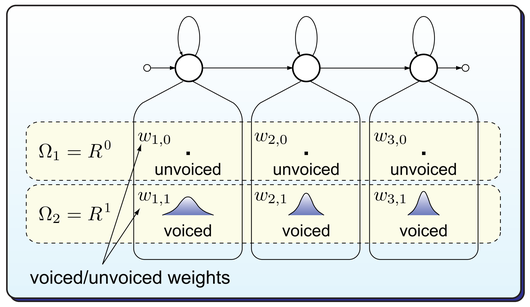
## 5. Application to log F0 modelling

### Application of additive acoustic models

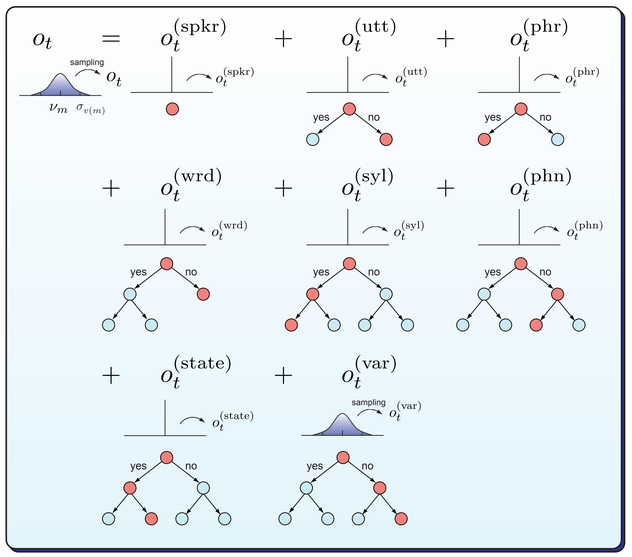
· Effectiveness of additive acoustic model
- Depends on the nature of observations
- If observations have the context-dependent additive nature
  ⇒ Additive acoustic model works effectively

· Logarithmic fundamental frequency (log F0) contours
- log F0 has an additive nature [Fujisaki;'08]
- Various additive F0 models have been proposed
  * Fujisaki's model [Fujisaki;'08]
  * Multi-layer additive model [Sakai;'04]

· HMM-based speech synthesis
- log F0 contours are modeled by MSD-HMMs [Tokuda;'02]
- MSD consists of 1 continuous space & 1 discrete space
- Continuous space (voiced) is modeled by Gaussian
- Discrete space (unvoiced) is modeled by discrete distribution
- Proposed model is integrated to MSD-HMMs (voiced frames)




· Application to log F0 modeling
- Additive structure is defined based on linguistic knowledge
  ⇒ Spkr., utt., phr., word, syl., phone, state, & variance
- Decision trees of all layers are built simultaneously
- Questions related to each layer pnly are applied




## 6. Experiment

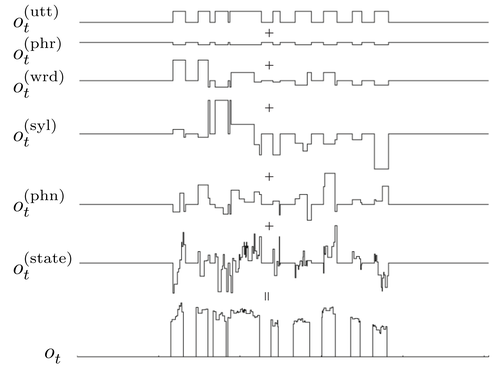
### Conditions

| | |
|---|---|
| Database | Toshiba internal database, 1 female speaker |
| Training data | 250 utterances (manually corrected labels & F0) |
| Test data | 60 sentences from 10 domains |
| Sampling freq. | 16 kHz |
| Analysis win. | 5-ms shift |
| Feature vec. | 0∼39 order Mel-cepstral coefficients, Δ & ΔΔ |
| Topology | 5-state, left-to-right no-skip |

### Constructed decision trees

| Additive components | # of leaf nodes | | |
|---|---|---|---|
| | Baseline | Syl+State | All |
| $o_t^{(\text{spkr})}$ | – | 1 | 1 |
| $o_t^{(\text{utt})}$ | – | – | 11 |
| $o_t^{(\text{phr})}$ | – | – | 48 |
| $o_t^{(\text{wrd})}$ | – | – | 52 |
| $o_t^{(\text{syl})}$ | – | 1,230 | 553 |
| $o_t^{(\text{phn})}$ | – | – | 1,314 |
| $o_t^{(\text{state})}$ | 2,607 | 1,376 | 628 |
| $o_t^{(\text{var})}$ | 2,607 | 3,075 | 2,765 |
| Total | 5,214 | 5,628 | 5,372 |

### Example of mean vector sequence




### Preference scores

| Baseline | Syl+State | No pref. |
|---|---|---|
| 32.6 | 43.5 | 23.9 |

| Baseline | All | No pref. |
|---|---|---|
| 41.8 | 32.7 | 25.4 |

## 7. References

Yoshimura;'99 - "Simultaneous modeling of spectrum, pitch and...," Eurospeech '99.
Ling;'08 - "Articulatory control of HMM-based parametric speech...," Interspeech '08.
Fujisaki;'08 - "In search of models in speech communication research," Interspeech '08.
Nankaku;'08 - "Acoustic modeling with contextual additive structure for...," ICASSP '08.
Qian;'08 - "Generating natural F0 trajectory with additive trees," Interspeech '08.
Gales;'00 - "Cluster adaptive training of hidden Markov models," IEEE Trans. SAP, '00.
Odell;"95 - "The use of contexts in large vocaburary," PhD thesis, Cambridge Univ., '95.
Sakai;'04 - "F0 modeling with multi-layer additive modeling based...," ISCA SSW5, '04.
Tokuda;'02 - "Multi-space probability distribution HMM," IEICE Trans. Inf. Syst., '02.