1. Introduction

HMM-based speech synthesis [Yoshimura;’99]
- Typical system
  - Spectrum, excitation, & duration are modeled 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]
  * Consideration of contextual nature

Additive components
- Gaussian additive model [Nankaku;’08]
  - Additive components are modeled by Gaussians
  - Not applicable yet for F0 due to huge computation
- Additively boosted HMMs [Qian;’08]
  - Sequentially boosting HMMs to minimize F0 RMSE
- Bias additive model (proposed)
  - Additive components are modeled by bias terms
  - Computationally less expensive than convolutional ones

2. Definition of bias additive model

1. Observations: generated as the sum of additive comps.
   \[ \mathbf{a}_t = \sum_{i=1}^{P-1} \mathbf{g}_{i}^{(P)}(\mathbf{a}_{i,\mu}) + \mathbf{g}_{P}^{(P)}(\mathbf{a}_{P,\mu}) + \mathbf{v}_t \]

2. \( \mathbf{a}^{(P)} \sim N(0, \sigma_{a}^{2}) \)

3. By equating the first partial deriv. of Q-function w.r.t. \( \theta \)

4. Component distributions

5. All bias terms can be determined simultaneously

6. Rank deficient \( \Rightarrow \) Least square solution

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

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
    \( \Rightarrow \) Decision tree-based context clustering [Odell;’95]

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

- Bias additive model (proposed)
  - Each cluster (bias & variance) has different tree
    \( \Rightarrow \) Can extract the underlying additive structure

4. Parameter estimation

EM algorithm-based ML estimation

\[ Q(\lambda, \Lambda) = \frac{1}{2} \sum_{n=1}^{N} \log |\Sigma_{n}| + \frac{(\mathbf{a}_{n} - \mu_{n})^{T} \Sigma_{n}^{-1} (\mathbf{a}_{n} - \mu_{n})}{2} + C \]

5. Application to log F0 modeling

Additive acoustic models

- Effectiveness of additive acoustic model
  - Depends on the nature of observations
  - If observations have the context-dependent additive nature
    \( \Rightarrow \) 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)

6. Experiment

Example of mean vector sequence

Additive components

Baseline | SyllState | All
--- | --- | ---
\( \hat{a}_{1} \) | 1 | 4
\( \hat{a}_{2} \) | 4 | 1
\( \hat{a}_{3} \) | 11 | 48
\( \hat{a}_{4} \) | 3 | 52
\( \hat{a}_{5} \) | 1,230 | 553
\( \hat{a}_{6} \) | 1,314 | 628
\( \hat{a}_{7} \) | 2,607 | 2,075
\( \hat{a}_{8} \) | 2,607 | 2,075
\( \hat{a}_{9} \) | 5,214 | 5,628
\( \hat{a}_{10} \) | 5,372 |

Preference scores

Baseline | SyllState | No pref.
--- | --- | ---
32.6 | 43.5 | 23.9
41.8 | 32.7 | 25.4

7. References

Yoshimura;’99 - “Simultaneous modeling of spectrum, pitch and...” Eurospeech ’99
Log;’08 - “Articulatory control of HMM-based parametric speech...” Interspeech ’08
Fujisaki;’04 - “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
Sakai;’04 - “F0 modeling with multi-layer additive modeling based...” ICSJ SSWS, ’04
Tokuda;’02 - “Multi-space probability distribution HMM...” IEICE Trans. Inf. Syst., ’02