Disclaimer

- The material provided in this document is not my original work and is a summary of some one else’s work(s).
- A simple Google search of the title of the document will direct you to the original source of the material.
- I do not guarantee the accuracy, completeness, timeliness, validity, non-omission, merchantability or fitness of the contents of this document for any particular purpose.
- Downloaded from najeebkhan.github.io

Speech Parameter Generation Algorithms for HMM with Mixture Gaussian Distribution

Presented by

Najeeb

July 10th 2014

Incorporation of State Duration Density

- Let $p_k(d_k)$ be the probability of being exactly d_k frames at state k, then the probability of state sequence q can be written as

$$P(q \mid \lambda) = \prod_{k=1}^{K} p_k(d_k) \qquad (5)$$

- where K is the total number of states visited during T frames

$$\sum_{k=1}^{K} d_{qk} = T \qquad (6)$$

Incorporation of State Duration Density

- The logarithm of $P(O,Q|\lambda, T)$ can be written as

$$\log P(O,Q\mid\lambda) = W_d \log P(q\mid\lambda) + \log P(s\mid q,\lambda) + \log P(O\mid Q,\lambda)$$

- Where W_d is a scaling factor for the score on state durations
- If W_d is set high, the state sequence q = (q1, q2, . . . , qT) is determined only by $P(q|\lambda, T)$

Incorporation of State Duration Density

- The state duration density is modeled by a single Gaussian pdf, q which maximizes P(q|λ, T) under the constraint (6) is given by

$$d_k = m_k + \rho.\sigma_k^2 \qquad 1 \leq \mathrm{k} \leq \mathrm{K}$$

$$\rho = \left(T - \sum_{k=1}^{K} m_k \right) \Big/ \sum_{k=1}^{K} \sigma_k^2$$

Incorporation of State Duration Density

- The mixture sequence i = $(i_1, i_2, \ldots, i_T)$ is determined in such a way that

$$\log w_{q_t i_t} - \frac{1}{2}\log\left|U_{q_t i_t}\right|$$

is maximized.

SPEECH PARAMETER GENERATION BASED ON MAXIMUM LIKELIHOOD CRITERION

- For a given continuous mixture HMM λ, we derive an algorithm for determining speech parameter vector sequence

$$O = [o_1, o_2, \dots, o_T]'$$

- Such that

$$P(O \mid \lambda) = \sum_{all\ Q} P(O, Q \mid \lambda)$$

is maximized where

$$Q = \{(q_1, i_1), (q_2, i_2), \dots, (q_T, i_T)\}$$

SPEECH PARAMETER GENERATION BASED ON MAXIMUM LIKELIHOOD CRITERION

- We assume that the speech parameter vector o_t consists of the static feature vector c_t and dynamic feature vectors $\Delta c_t, \Delta^2 c_t$

$$c_t = [c_t(1), c_t(2), \ldots, c_t(M)]'$$

$$\Delta c_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) c_{t+\tau} \qquad (1)$$

$$\Delta^2 c_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(\tau) c_{t+\tau} \qquad (2)$$

Maximizing P(O|Q, λ) with respect to O

- The logarithm of P(O|Q, λ) can be written as

$$\log P(O \mid Q, \lambda) = -\frac{1}{2} O^T U^{-1} O + O^T U^{-1} M + K$$

- Where

$$U^{-1} = diag[U_{q1,i1}^{-1}, U_{q2,i2}^{-1}, \ldots, U_{qT,iT}^{-1}]$$

$$M = [\mu_{q1,i1}^T, \mu_{q2,i2}^T, \ldots, \mu_{qT,iT}^T]$$

- P(O|Q, λ) is maximized when O = M without the conditions (1), (2)

$$O = M \Rightarrow Max\ P(O \mid Q, \lambda)$$

Maximizing P(O|Q, λ) with respect to O

- Conditions (1),(2) can be arranged in a matrix form

$$O = WC \quad (3)$$

- Where

$$C = [c_1, c_2, \ldots, c_T]^T$$

$$W = [w_1, w_2, \ldots, w_T]^T \quad w_t = [w_t^{(1)}, w_t^{(2)}, w_t^{(3)}]$$

$$w_t^{(n)} = [0_{M\times M}, \ldots, 0_{M\times M}, w^{(n)}(-L_-^{(n)})I_{M\times M},$$
$$\ldots, w^{(n)}(0)I_{M\times M}, \ldots, w^{(n)}(L_+^{(n)})I_{M\times M}, \ldots,$$
$$0_{M\times M}, \ldots, 0_{M\times M}]^T, \quad \mathrm{n}=0,1,2$$

Maximizing P(O|Q, λ) with respect to O

- Under the condition (3), maximizing P(O|Q, λ) with respect to O is equivalent to that with respect to C

$$\frac{\partial \log P(WC \mid Q, \lambda)}{\partial C} = 0$$

$$\frac{\partial \left[-\frac{1}{2} [WC]^T U^{-1} [WC] + [WC]^T U^{-1} M + K \right]}{\partial C} = 0$$

$$W^T U^{-1} W C = W^T U^{-1} M^T \quad (4)$$

Pitch Pattern Generation

- To obtain an F0 parameter sequence
  - Voiced and unvoiced regions are determined based on space weights at each state
  - Then F0 values are obtained in the same manner to spectral parameter sequence within voiced regions

Pitch Pattern Generation

- The dynamic features are

$$\Delta c_t = \frac{1}{2}(c_{t+1} - c_{t-1})$$

$$\Delta^2 c_t = \frac{1}{4}(c_{t+2} - 2c_t + c_{t-2})$$

$$\delta^l p_t = \frac{1}{14}(-3p_{t-3} - 2p_{t-2} - p_{t-1} + 6p_t)$$

$$\delta^r p_t = \frac{1}{14}(3p_{t+3} + 2p_{t+2} + p_{t+1} - 6p_t)$$

- If there are more than one unvoiced frames among frames required for calculation of δ^l_{pt} or δ^r_{pt}, one or both of them were handled as unvoiced since unvoiced frames do not have values of log F0, and therefore, δ^l_{pt} or δ^r_{pt} cannot be calculated

Pitch Pattern Generation

- What is the need for space weights?