

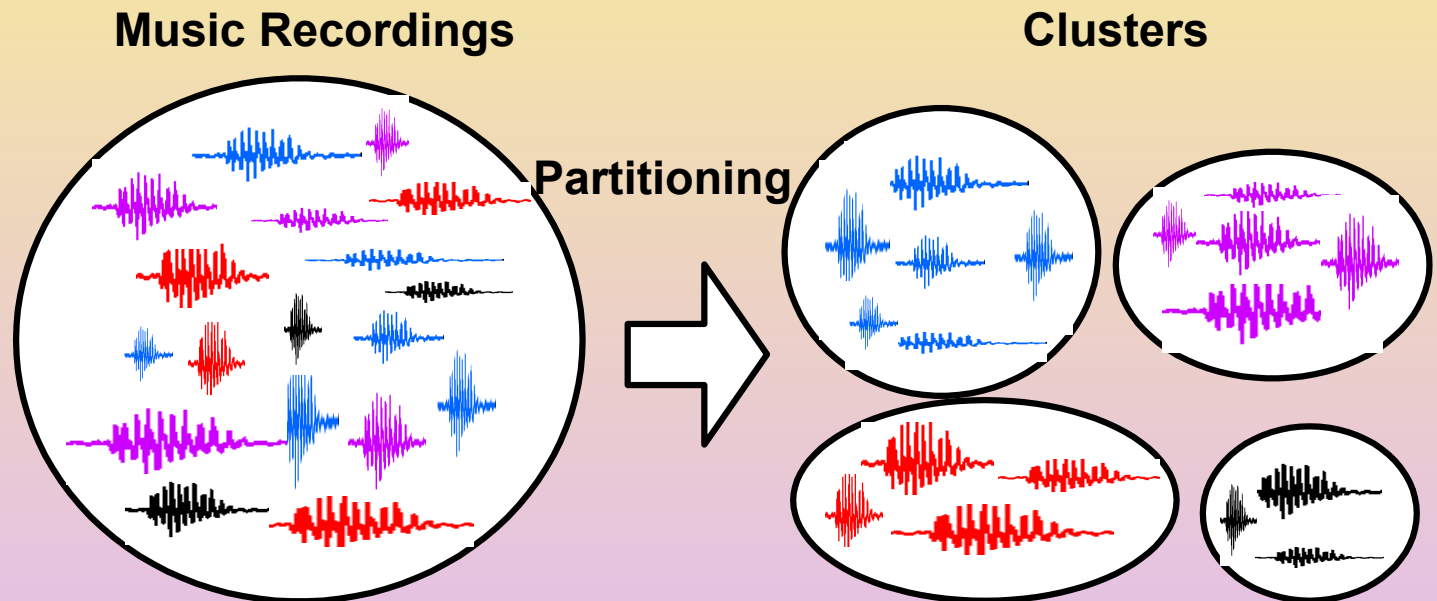
Blind Clustering of Popular Music Recordings Based on Singer Voice Characteristics

Wei-Ho Tsai

Institute of Information Science, Academia Sinica, Taiwan

The Task

- To cluster music recordings by singer



Applications

■ Music data indexing

- Organizing unlabeled or insufficiently well labeled music collections such as live concert recordings and bootlegs.

■ Karaoke services

- Efficiently organizing the customers' recordings.
- Personalization

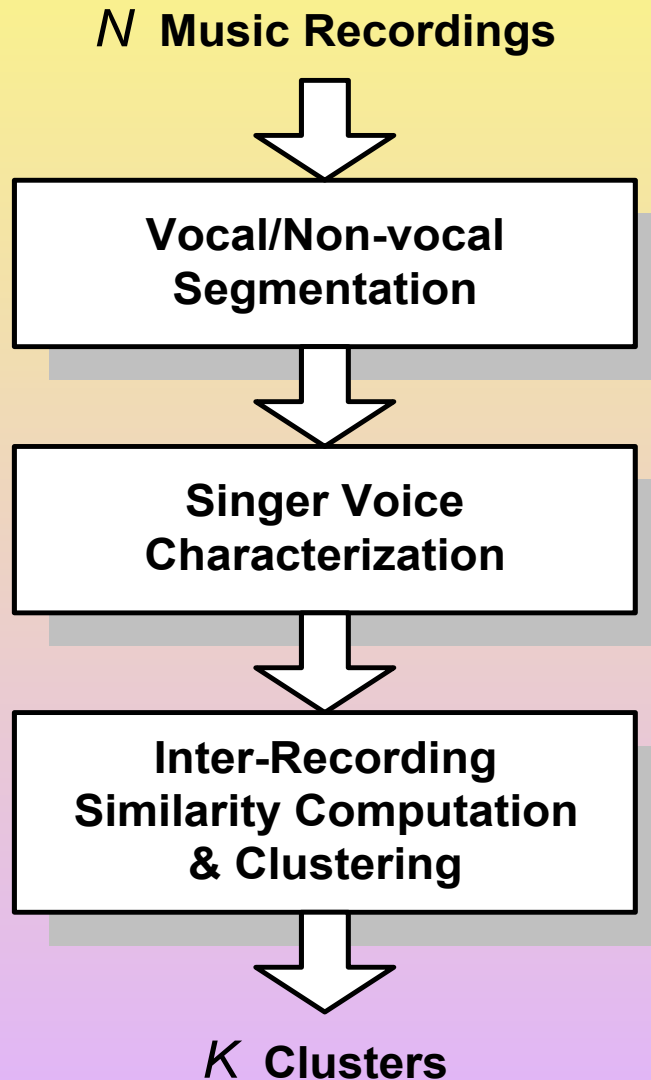
■ Music recommendation systems

- Suggesting music by singers with similar voices.

Major Challenges

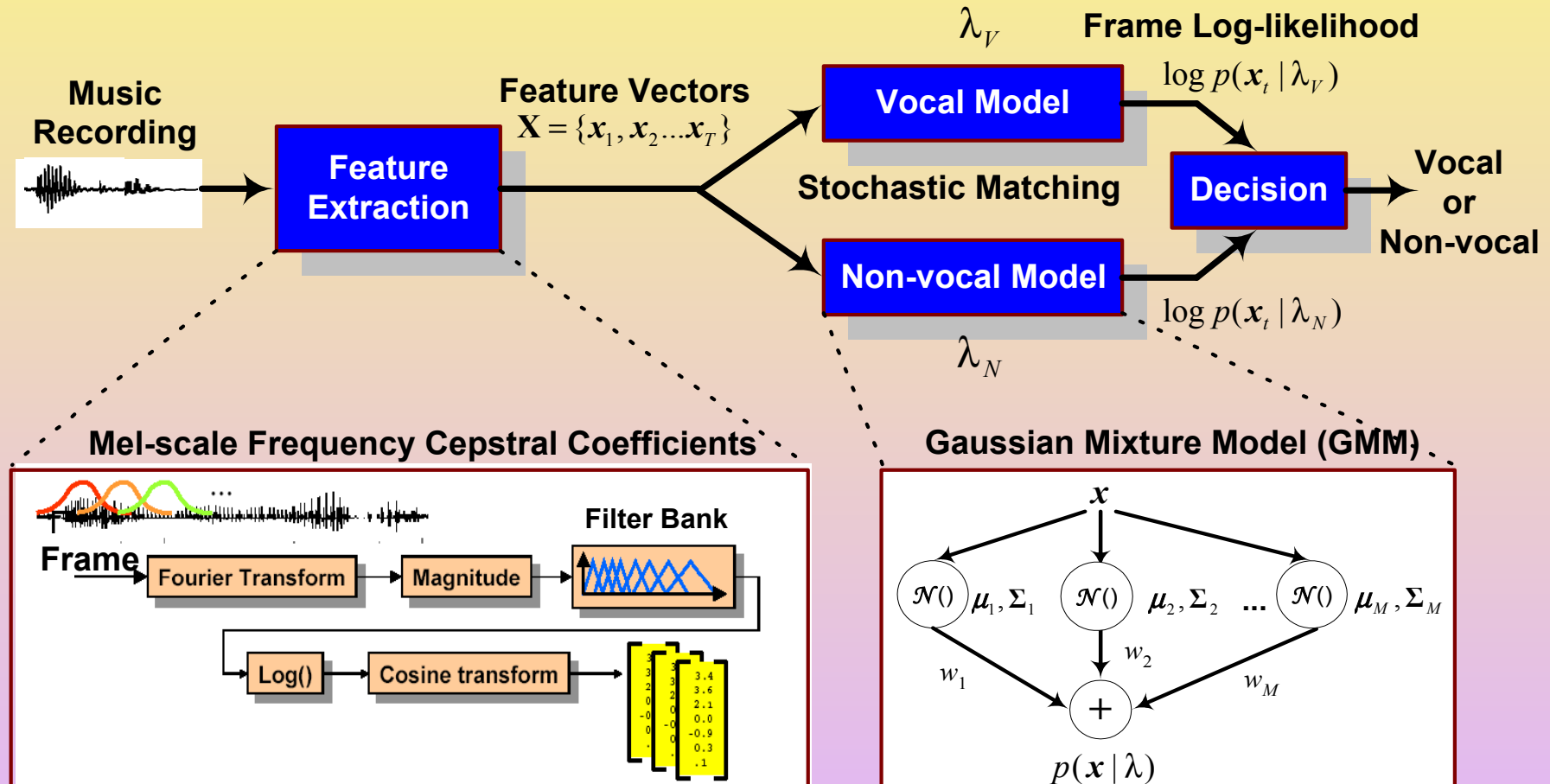
- **Singer's voices tend to be arbitrarily altered from time to time**
- **The vast majority of popular music contains background accompaniment during most or all vocal passages**

Method Overview



Vocal/Non-vocal Segmentation (I)

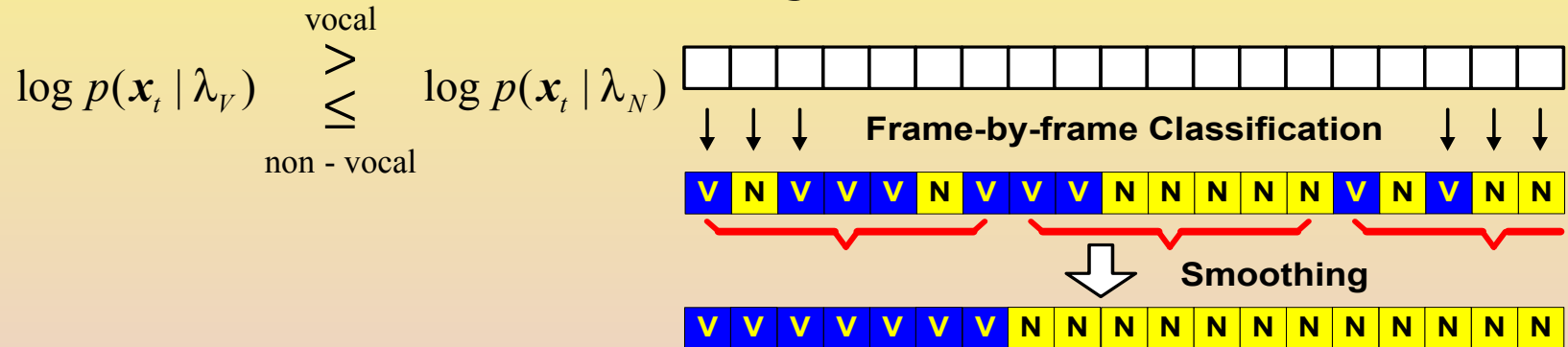
Block diagram



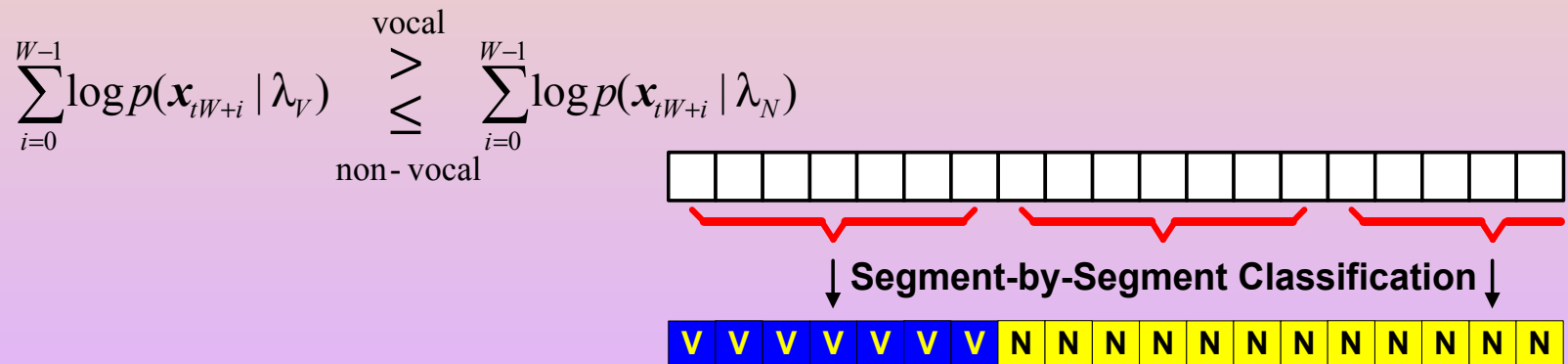
Vocal/Non-vocal Segmentation (II)

Decision

- Frame-based decision & smoothing

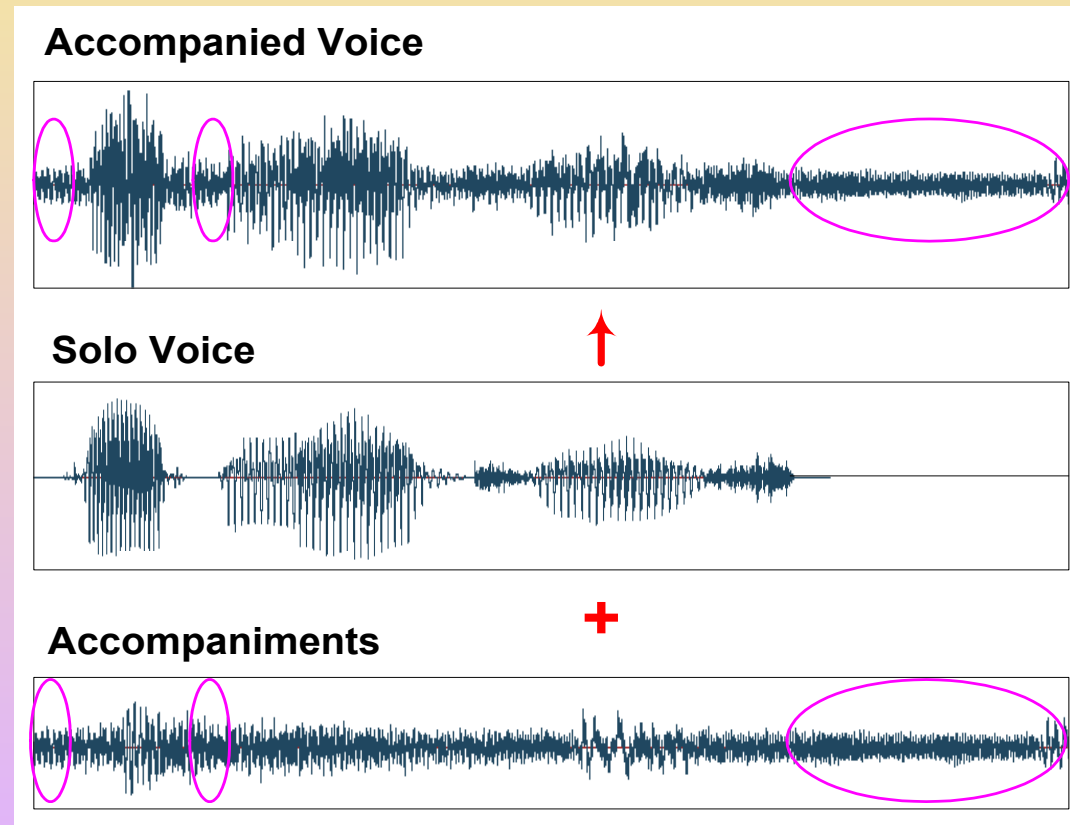


- Fixed-length-segment-based decision

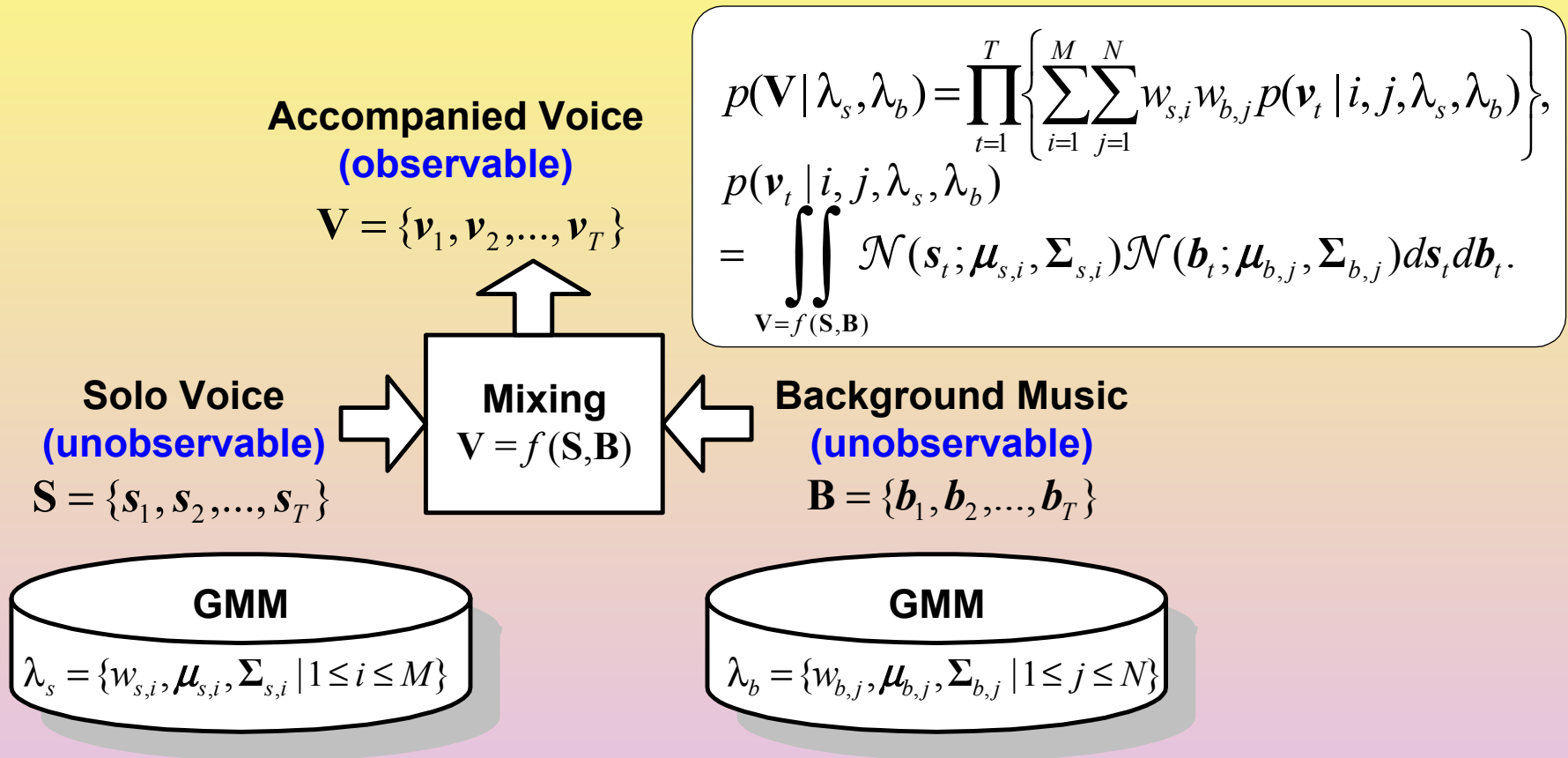


Cues For Singer Voice Characterization

- Substantial similarities exist between the instrumental regions and the accompaniment of the vocal signal
- Solo voice can be modeled via suppressing the background music estimated from the instrumental regions



Solo Voice Modeling (I)



- λ_b can be approximately estimated using the instrumental regions
- Our aim is to find an optimal solo voice model λ_s such that

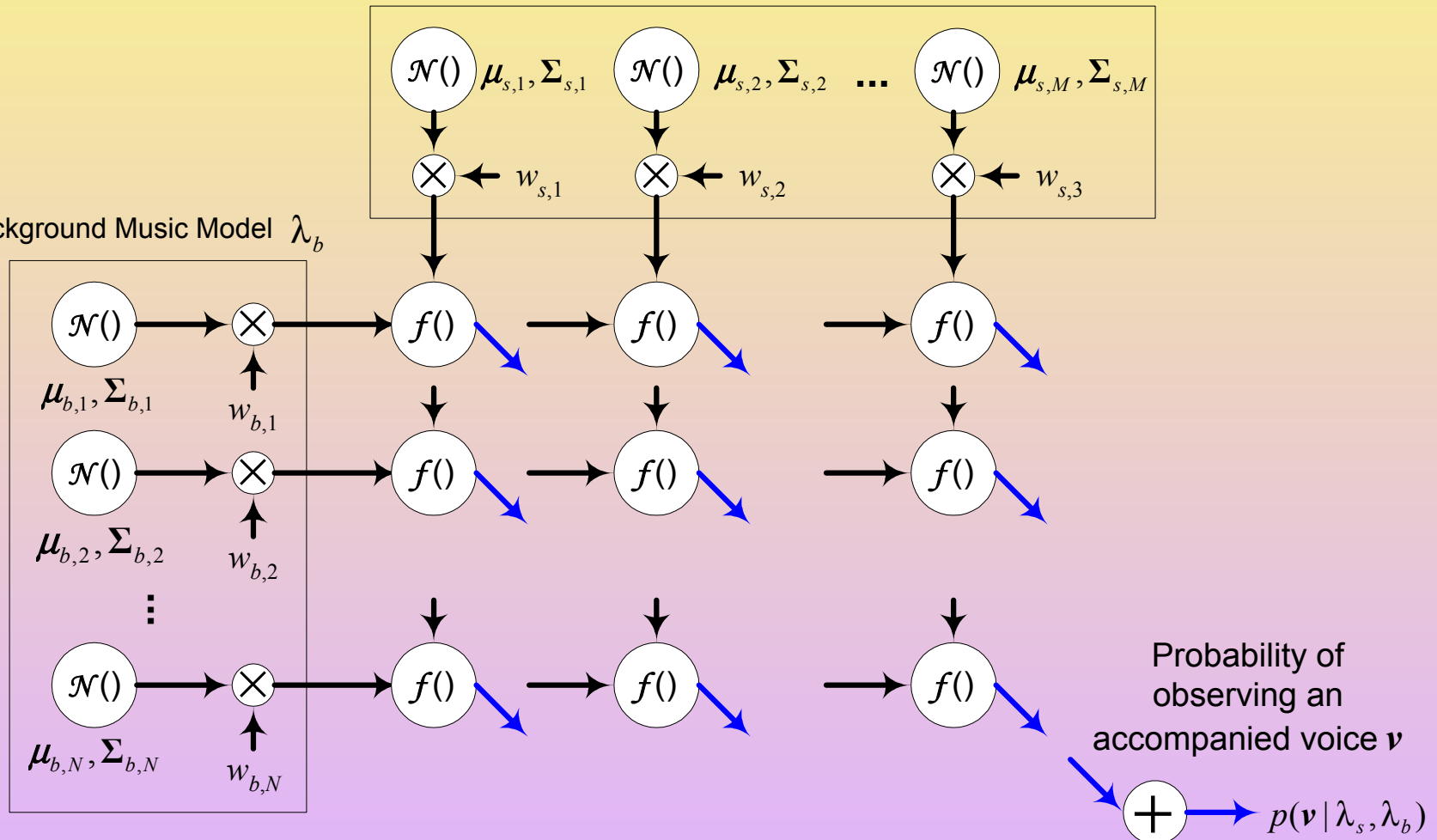
$$\lambda_s^* = \arg \max_{\lambda_s} p(\mathbf{V} | \lambda_s, \lambda_b).$$

Solo Voice Modeling (II)

Accompanied Voice Generation

Solo Voice Model λ_s

Background Music Model λ_b



Solo Voice Modeling (III)

■ Parameter estimation via Expectation-Maximization

- Defining an auxiliary function

$$Q(\lambda_s, \hat{\lambda}_s) = \sum_{t=1}^T \sum_{i=1}^I \sum_{j=1}^J p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b) \log p(i, j, \mathbf{v}_t | \hat{\lambda}_s, \lambda_b),$$

where $p(i, j, \mathbf{v}_t | \hat{\lambda}_s, \lambda_b) = w_{s,i} w_{b,j} p(\mathbf{v}_t | i, j, \hat{\lambda}_s, \lambda_b),$

$$p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b) = \frac{w_{s,i} w_{b,j} p(\mathbf{v}_t | i, j, \lambda_s, \lambda_b)}{\sum_{m=1}^I \sum_{n=1}^J w_{s,m} w_{b,n} p(\mathbf{v}_t | m, n, \lambda_s, \lambda_b)}.$$

- Letting $\nabla Q(\lambda_s, \hat{\lambda}_s) = 0$ for each parameter to be re-estimated, we have

$$\hat{w}_{s,i} = \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^J p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b),$$

$$\hat{\mu}_{s,i} = \frac{\sum_{t=1}^T \sum_{j=1}^N p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b) \cdot E\{s_t | \mathbf{v}_t, i, j, \lambda_s, \lambda_b\}}{\sum_{t=1}^T \sum_{j=1}^N p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b)},$$

$$\hat{\Sigma}_{s,i} = \frac{\sum_{t=1}^T \sum_{j=1}^J p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b) \cdot E\{s_t s_t' | \mathbf{v}_t, i, j, \lambda_s, \lambda_b\}}{\sum_{t=1}^T \sum_{j=1}^J p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b)} - \mu_{s,i} \mu_{s,i}',$$

Solo Voice Modeling (IV)

■ Re-estimation formulas for cepstral features

- Suppose V is a cepstral feature, and S and B are additive in the time domain, then $v_t = \log[\exp(s_t) + \exp(b_t)]$. We approximate $v_t \approx \max(s_t, b_t)$.

- It can be shown that

$$p(v_t | i, j, \lambda_s, \lambda_b) = \mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2) \Phi\left(\frac{v_t - \mu_{b,j}}{\sigma_{b,j}}\right) + \mathcal{N}(v_t; \mu_{b,j}, \sigma_{b,j}^2) \Phi\left(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}\right), \quad \Phi(\tau) = \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw.$$

$$E\{s_t | v_t, i, j, \lambda_s, \lambda_b\} = p(s_t = v_t | i, j, \lambda_s, \lambda_b) \cdot v_t + (1 - p(s_t = v_t | i, j, \lambda_s, \lambda_b)) \cdot E\{s_t | s_t < v_t, i, j, \lambda_s, \lambda_b\}$$

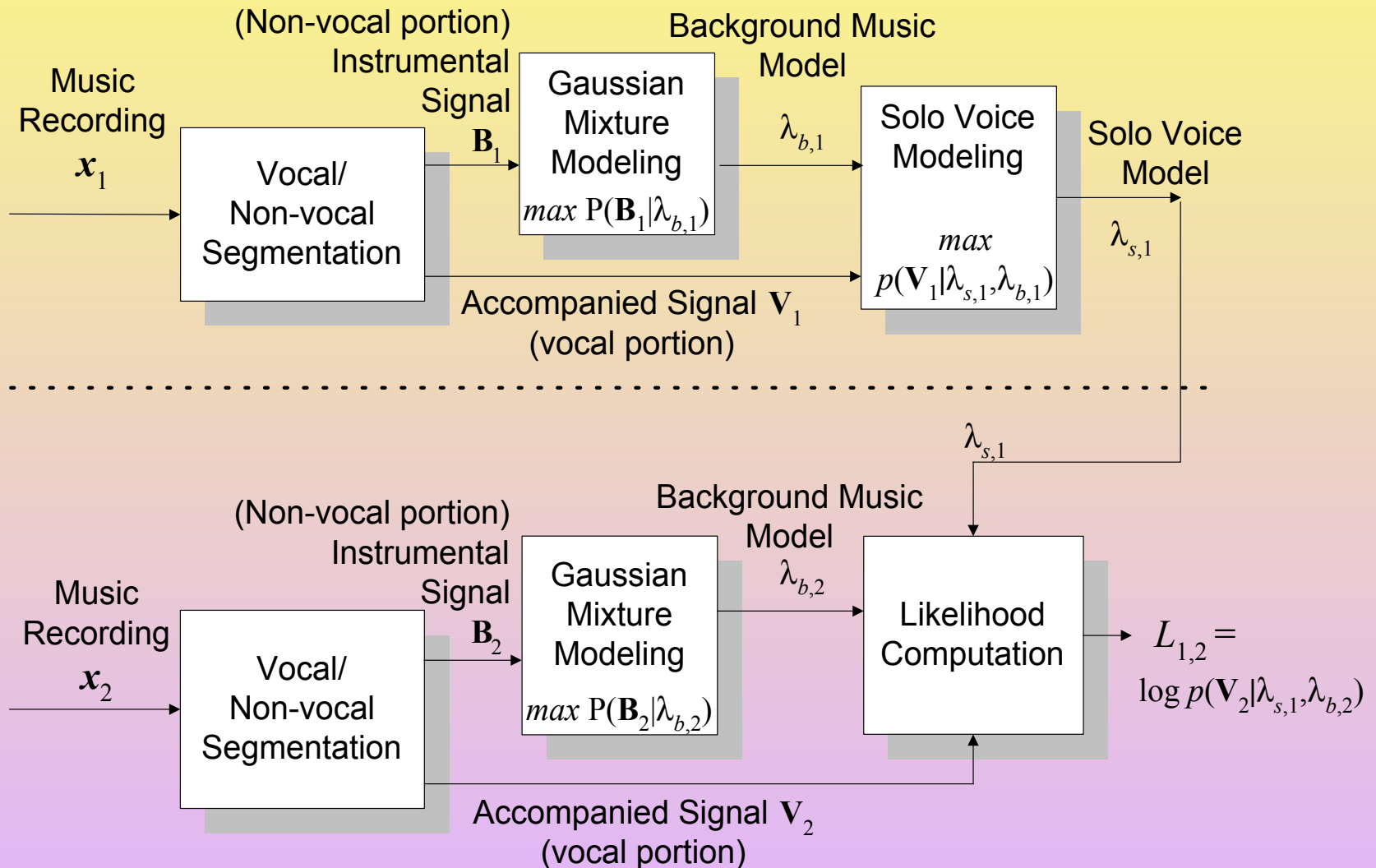
$$E\{s_t^2 | v_t, i, j, \lambda_s, \lambda_b\} = p(s_t = v_t | i, j, \lambda_s, \lambda_b) \cdot v_t^2 + (1 - p(s_t = v_t | i, j, \lambda_s, \lambda_b)) \cdot E\{s_t^2 | s_t < v_t, i, j, \lambda_s, \lambda_b\}$$

$$p(s_t = v_t | i, j, \lambda_s, \lambda_b) = \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2) \Phi\left(\frac{v_t - \mu_{b,j}}{\sigma_{b,j}}\right)}{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2) \Phi\left(\frac{v_t - \mu_{b,j}}{\sigma_{b,j}}\right) + \mathcal{N}(v_t; \mu_{b,j}, \sigma_{b,j}^2) \Phi\left(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}\right)},$$

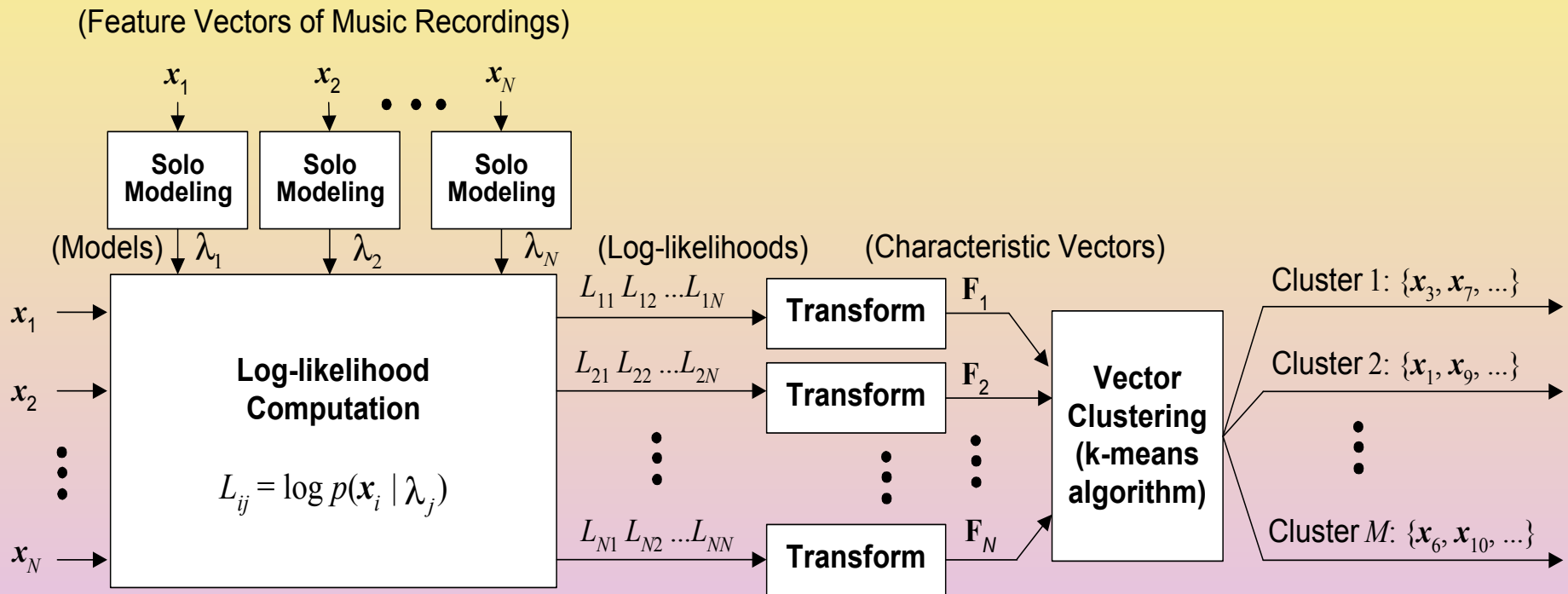
$$E\{s_t | s_t < v_t, i, j, \lambda_s, \lambda_b\} = \mu_{s,i} - \sigma_{s,i} \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)}{\Phi\left(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}\right)}.$$

$$E\{s_t^2 | s_t < v_t, i, j, \lambda_s, \lambda_b\} = \mu_{s,i}^2 + \sigma_{s,i}^2 - (\mu_{s,i} + v_t) \sigma_{s,i} \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)}{\Phi\left(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}\right)}.$$

Inter-recording Likelihood Computation

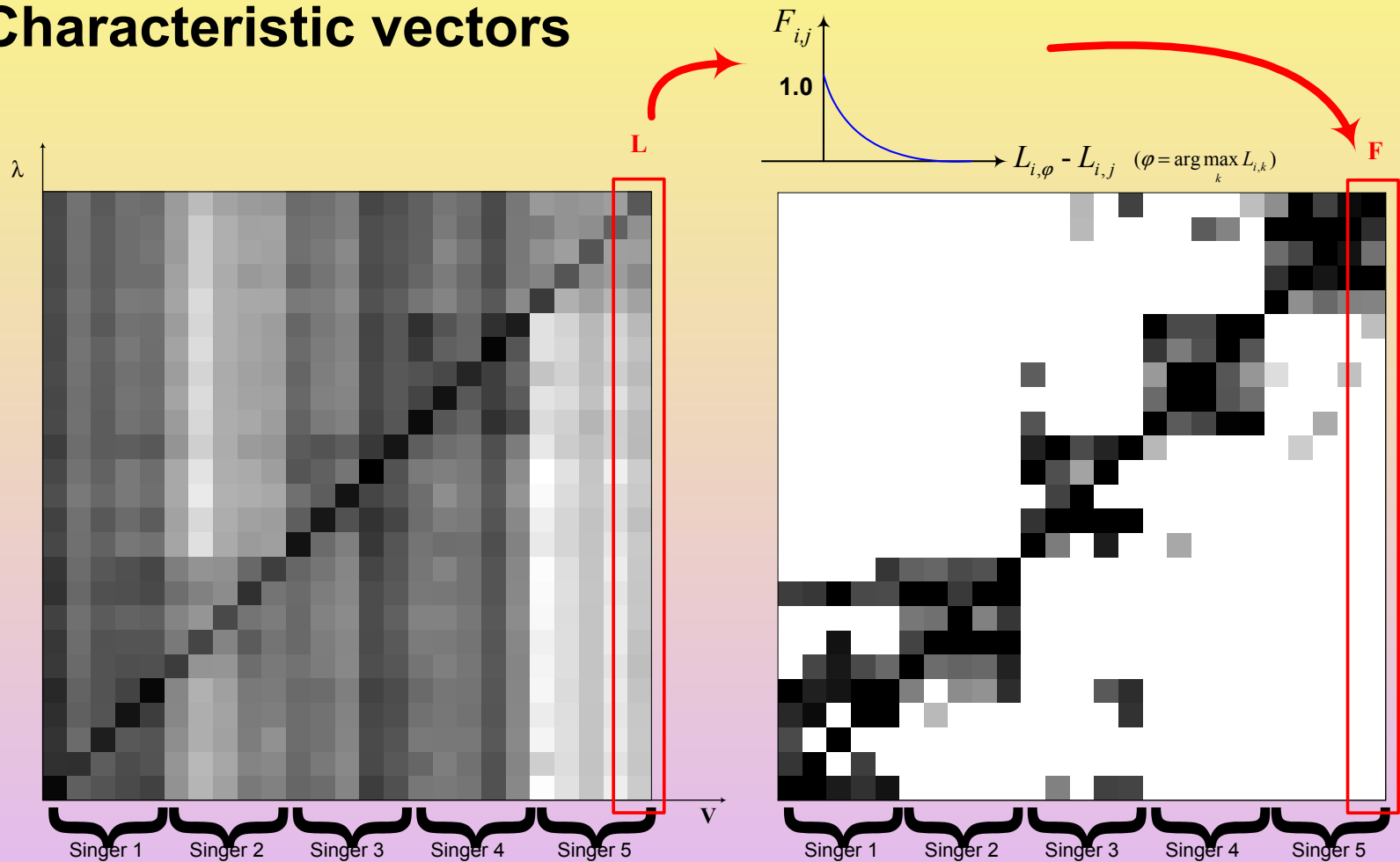


Singer-based Clustering (I)



Singer-based Clustering (II)

■ Characteristic vectors



- Converting into a problem of conventional vector clustering.

Determining The Number Of Clusters

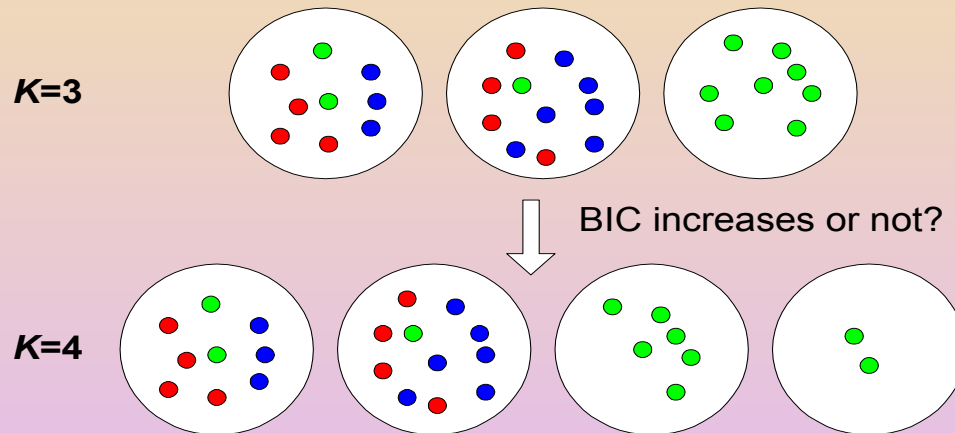
■ Bayesian Information Criterion (BIC)

- Choosing one among a set of candidate models $\{\Lambda_1, \Lambda_2, \dots, \Lambda_K\}$ can best represent a given data set \mathcal{D}

$$\text{BIC}(\Lambda_i) = \log p(\mathcal{D} | \Lambda_i) - \frac{1}{2} \gamma d_i \log |\mathcal{D}|,$$

d_i : no. of free parameters in model Λ_i
 $|\mathcal{D}|$: size of the data set \mathcal{D}
 γ : a penalty factor

■ Viewing a K -clustering as a candidate model



- An appropriate number of clusters can be determined by $K^* = \arg \max_{1 \leq K \leq M} \text{BIC}(K).$

Experimental Results (I)

■ Music data

- 416 tracks from Mandarin pop music CDs
- Sub-set DB-1: 200 tracks
 - 10 male & 10 female singers; 10 different songs/singer
 - Used for the performance evaluation of the singer-based clustering
- Sub-set DB-2: 216 tracks
 - 8 male & 13 female singers; none of the singers appeared in DB-1
 - Used for the training of the vocal and non-vocal models

■ Vocal/non-vocal segmentation results

- Assessment method

$$\text{Frame accuracy (in\%)} = \frac{\text{\# correctly - classified frames}}{\text{\# total frames}} \times 100\%$$

- The performance achieved with the frame-based decision and the segment-based decision were, respectively, 76.8% and 77.6% frame accuracy

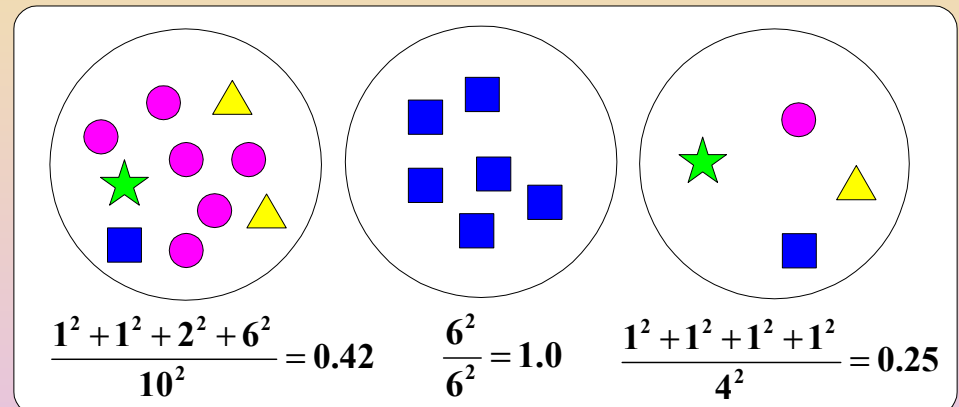
Experimental Results (II)

■ Clustering assessment method

– Cluster purity

$$\rho_k = \sum_{p=1}^P \frac{n_{kp}^2}{n_k^2},$$

- ρ_k is the purity of the cluster k , n_k the total no. of recordings in the cluster k , and n_{kp} the no. of recordings in the cluster k that were performed by singer p



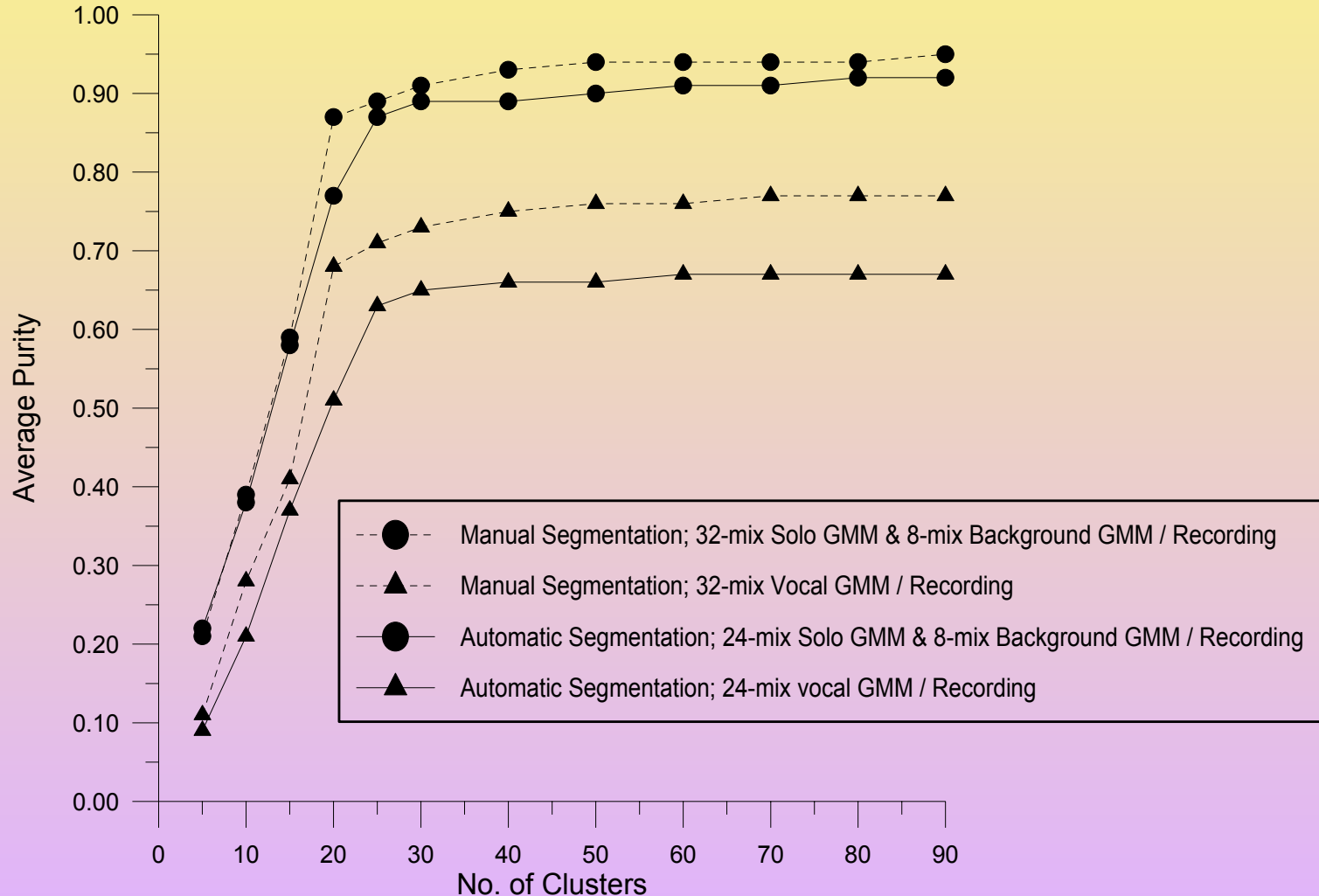
– Average purity

$$\bar{\rho} = \frac{1}{M} \sum_{k=1}^K n_k \rho_k,$$

- M is the total no. of recordings, and K the no. of clusters

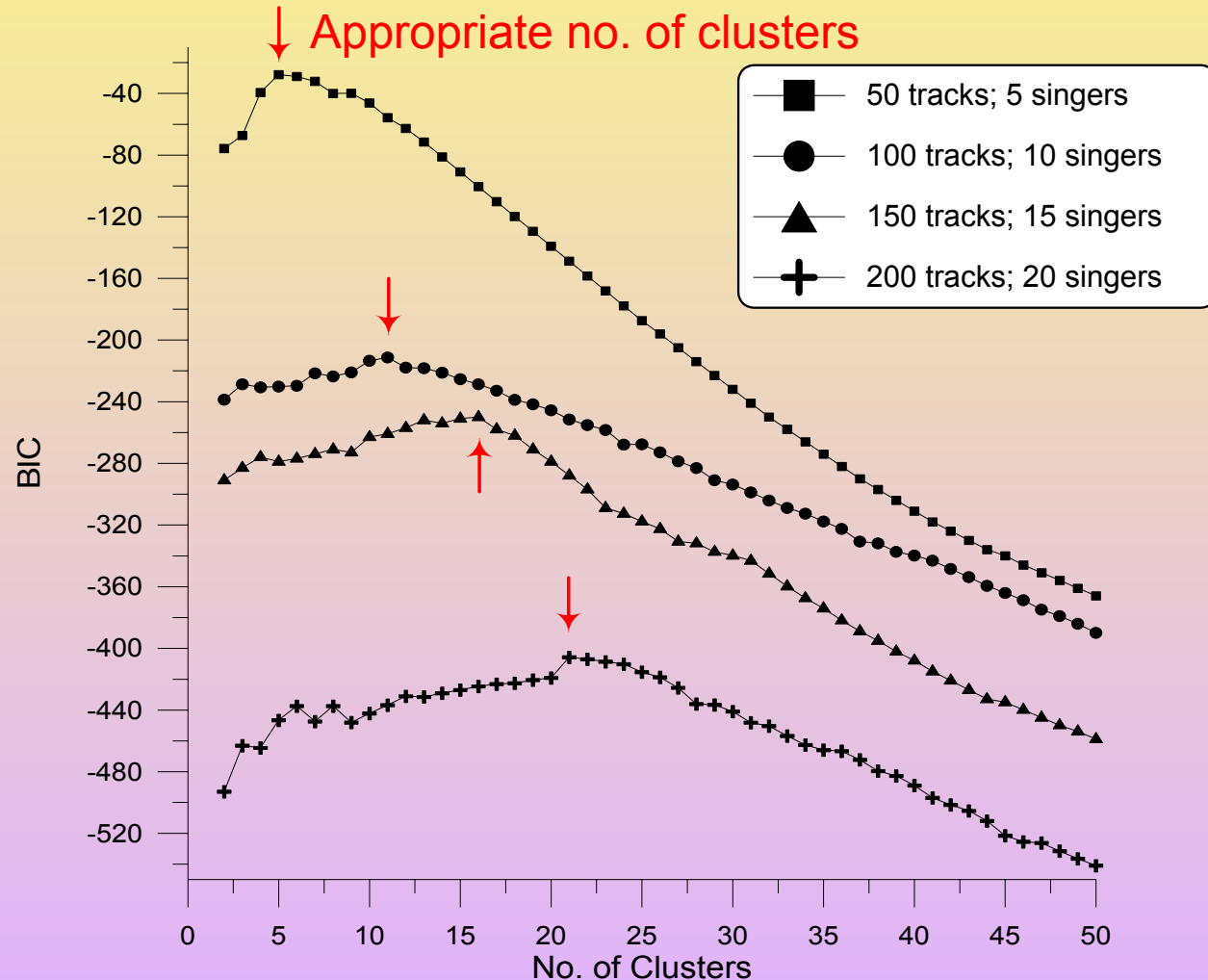
Experimental Results (III)

■ Result of the clustering for 200 tracks (20 singers \times 10 songs)



Experimental Results (IV)

■ Results of automatically determining the no. of clusters



Summary

■ We have

- Separated vocal from non-vocal segments of music;
- Isolated singers' vocal characteristics from the background music;
- Clustered music recordings by singer.

■ We will

- Handle a wider variety of music data including duets, trios, chorus, background vocals, or music with multiple simultaneous or non-simultaneous singers.