# Musical Genre Classification

Wei-Ta Chu

2014/11/19

G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Trans. on Speech and Audio Processing, vol. 10, no. 5, 2002, pp. 293-302.

# Introduction



- The members of a particular genre share certain characteristics
- Automatic musical genre classification
  - Music information retrieval
  - Developing and evaluating features that can be used in similarity retrieval, classification, segmentation, and audio thumbnailing

# Related Work



- Audio classification has a long history originating from speech recognition
  - Classify audio signals into music, speech, and environmental sounds
  - Classify musical instrument sounds and sound effects
- The features they used are not adequate for automatic musical genre classification

# Feature Extraction



- Timbral Texture Features
  - spectral centroid, spectral rolloff, spectral flux, zero-crossing rate, MFCC, energy
- Rhythmic Content Features
- Pitch Content Features

# Spectral Centroid



- The center of gravity of the magnitude spectrum of short-time Fourier transform (STFT)

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]}$$

$M_t[n]$ is the magnitude of the Fourier transform at frame $t$ and frequency bin $n$

- A measure of spectral shape and higher centroid values correspond to “brighter” textures with high frequencies

# Spectral Rolloff



- The frequency $R_t$ such that

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n]$$

- A measure of the "skewness" of the spectral shape
- It is used to distinguish voiced from unvoiced speech and music. (unvoiced speech has a high proportion of energy contained in the high-freq. range of the spectrum)

# Spectral Flux



- Squared difference between the normalized magnitudes of successive spectral distributions

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2$$

$N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at frames $t$ and $t-1$

- A measure of the amount of local spectral change

# Zero-Crossing Rate



- A measure of the noisiness of the signal

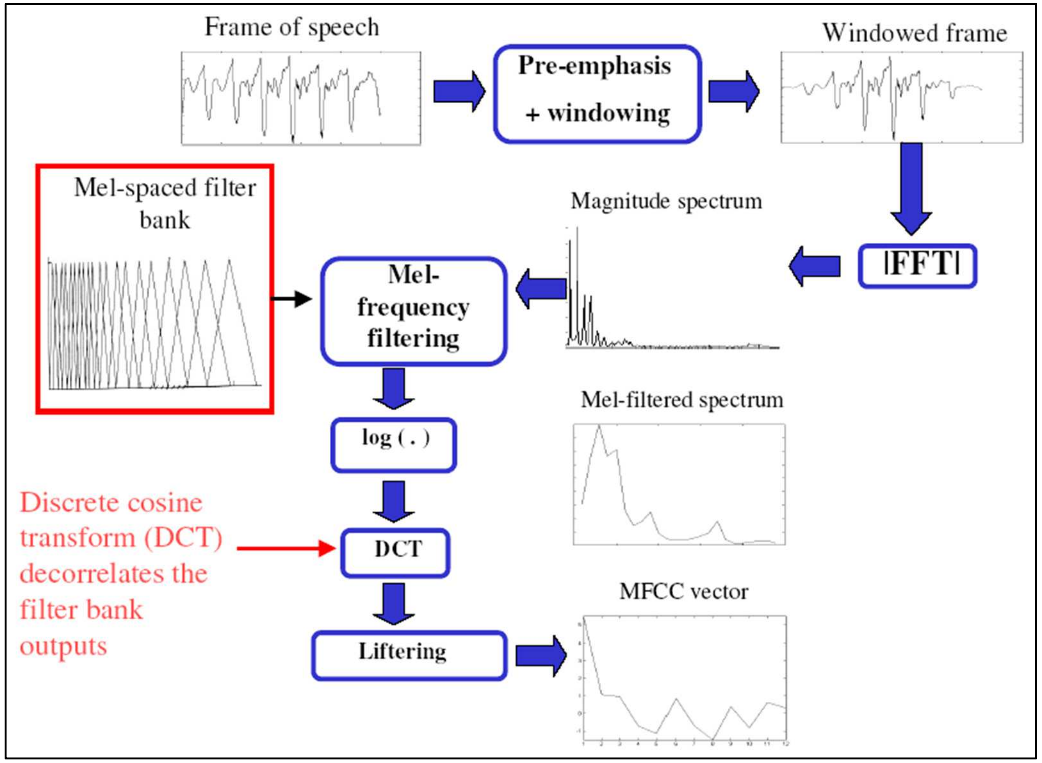
$$Z_t = \frac{1}{2}\sum_{n=1}^{N}\left|sign(x[n]) - sign(x[n-1])\right|$$

*sign* function is 1 for positive arguments and 0 for negative arguments
*x[n]* is the time domain signal for frame *t*

- Unvoiced speech has a low volume but a high ZCR

# Mel-Frequency Cepstral Coefficients (MFCC)

- First five coefficients provide the best genre classification performance




$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N}, \qquad 0 \le k < N$$

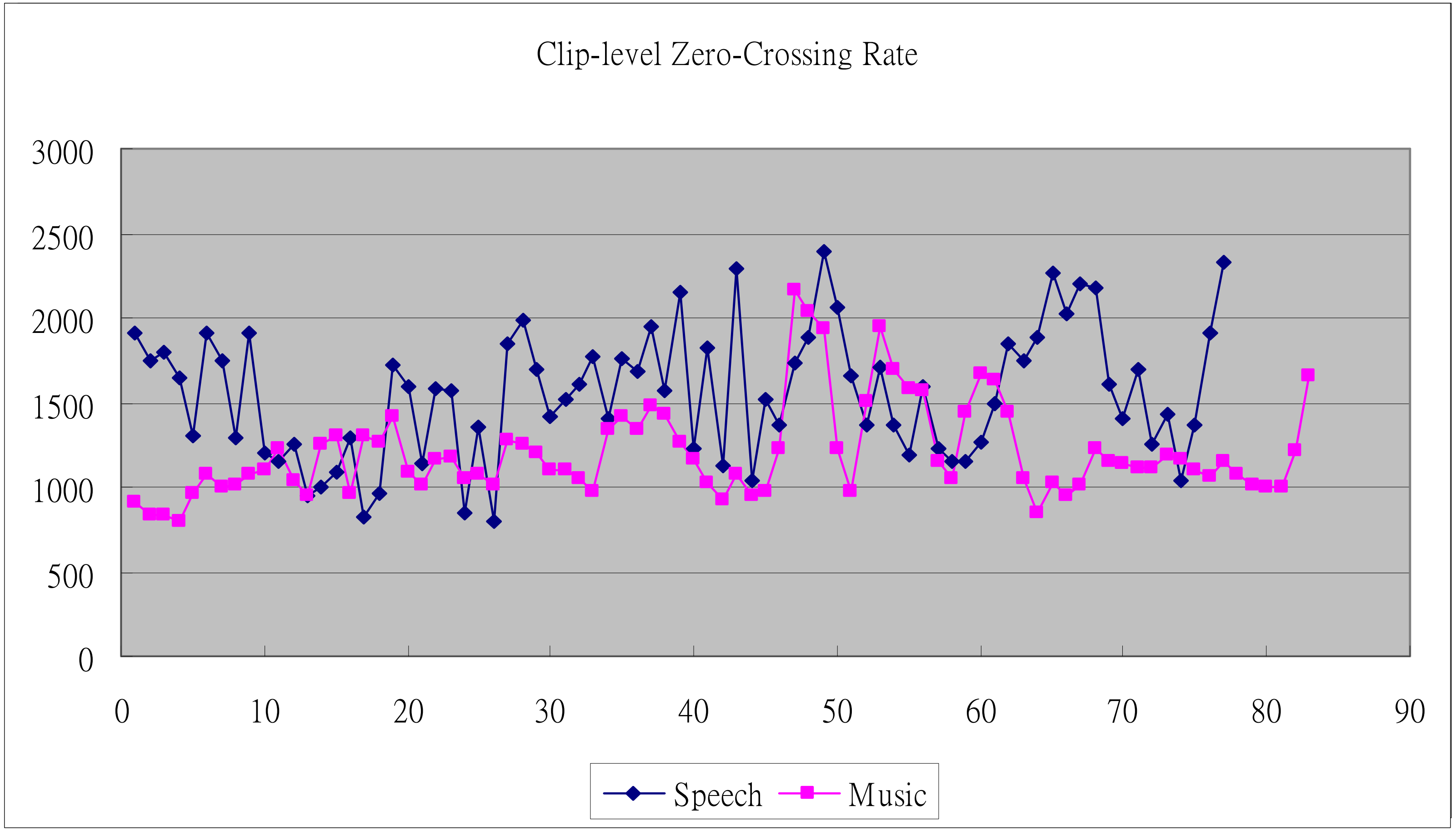
$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right], \qquad 0 < m \le M$$

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos(\pi n(m-1/2)/M), \qquad 0 \le n < M$$

*M*: the number of filters
*N*: the size of the FFT

# Examples of Audio Features





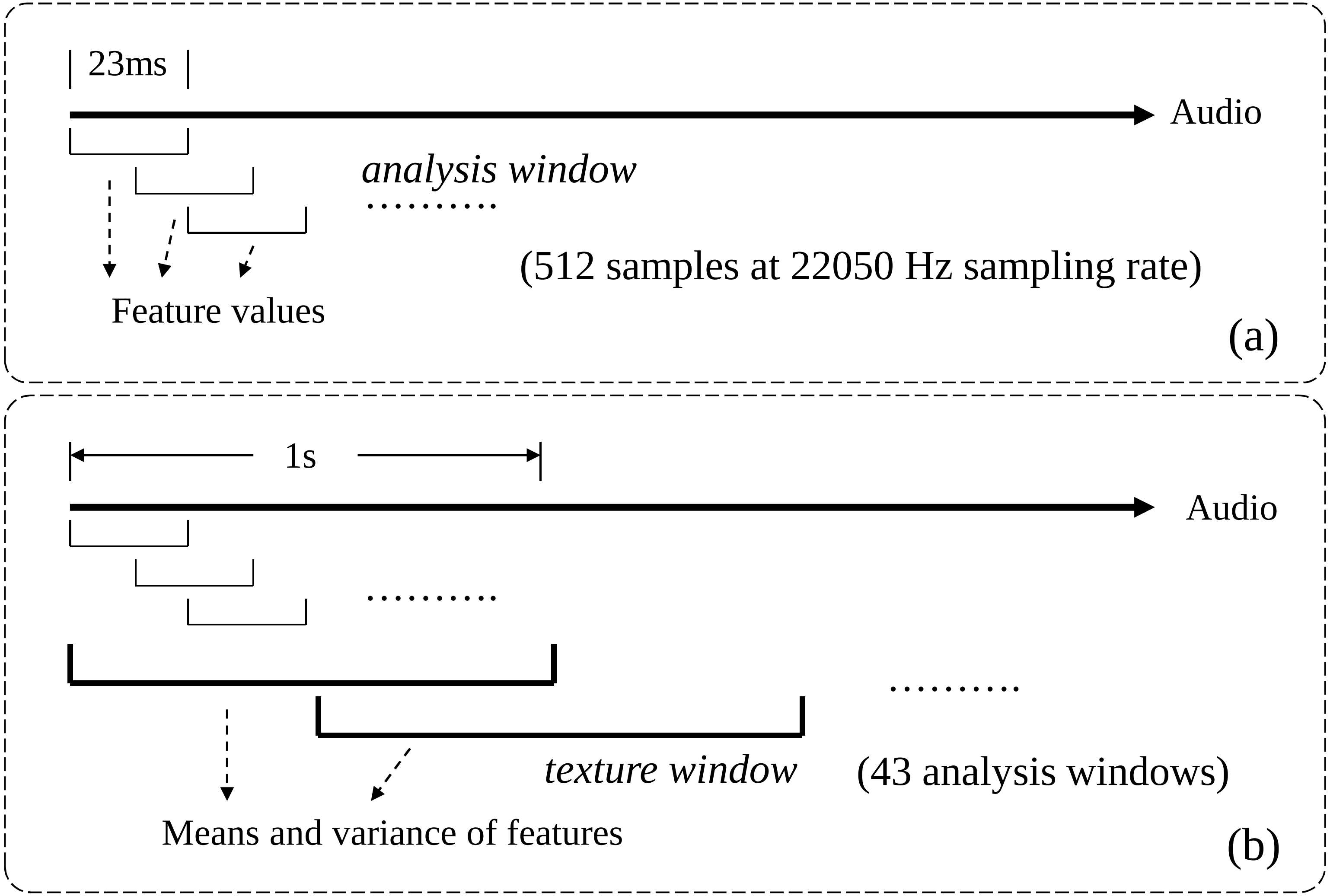
# Analysis and Texture Window (1/2)



- For short-time audio analysis, small audio segments are processed (*analysis window*).
- To capture the long term nature of sound “texture”, means and variances of features over a number of *analysis windows* are calculated (*texture windows*).
- For each texture window, multidimensional Gaussian distribution of features are estimated.

# Analysis and Texture Window (2/2)





# Low-Energy Feature



- Based on the texture window
- The percentage of analysis windows that have less energy than the average energy across the texture window.
- Ex: vocal music with silences have large low-energy value

# Rhythmic Content Features



- Characteristics: the regularity of the rhythm, the relation of the main beat to the subbeats, and the relative strength of subbeats to the main beat
- Steps of a common automatic beat detector
  - 1. Filterbank decomposition
  - 2. Envelop extraction
  - 3. Periodicity detection algorithm used to detect the lag at which the signal's envelope is most similar to itself
- Similar to pitch detection but with larger periods: approximately 0.5 to 1.5 s for beat vs. 2 ms to 50 ms for pitch

# Rhythmic Content Features



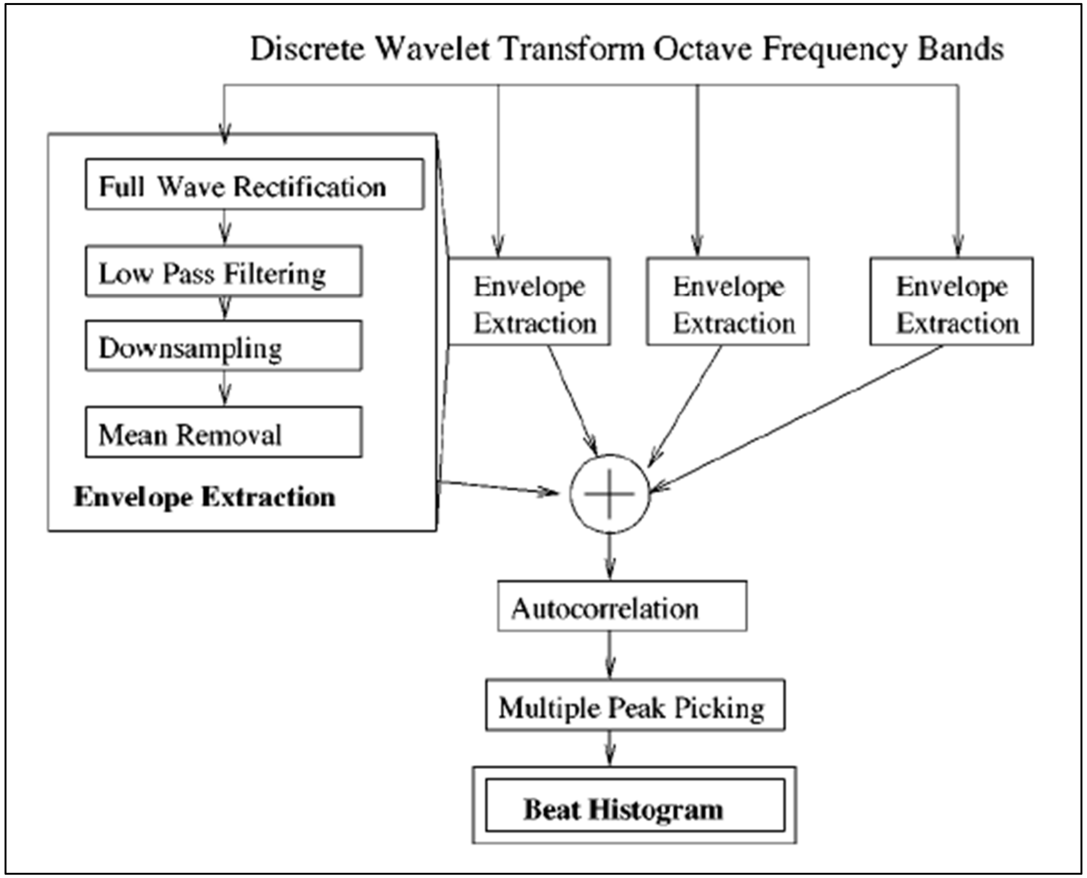
- Based on discrete wavelet transform (DWT)
  - Overcome the resolution problems (people percept differently in different freq. bands)
  - The DWT can be viewed as a computationally efficient way to calculate an octave decomposition of the signal in frequency.
  - DAUB4 filters are used.
- Find the rhythmic structure: detect the most salient periodicities of the signal

# Rhythmic Content Features



- Beat detection flowchart

**Beat:** the sequence of equally spaced phenomenal impulses which define a tempo for the music



# Octave



- 在數理上，每一個八度音程(Octave)正好對應於不同的振動模式，而兩個八度音程差的音在頻率上正好差上兩倍。例如：在第0個八度的La(記為A0)頻率為27.5 Hertz，則第1個八度的La(記為A1)頻率即為27.5*2=55.0 Hertz。在這每一個八度的音程中，又可再將其等分為12個頻率差相近的音，這分別對應於【C Db D Eb E F Gb G Ab A Bb B】，這樣的等分法就是所謂的十二平均律(Twelve-Tone Scale)。這當中每一個音符所對應的頻率，都可以藉由數學的方程式準確的算出

# Octave and Semi-tone



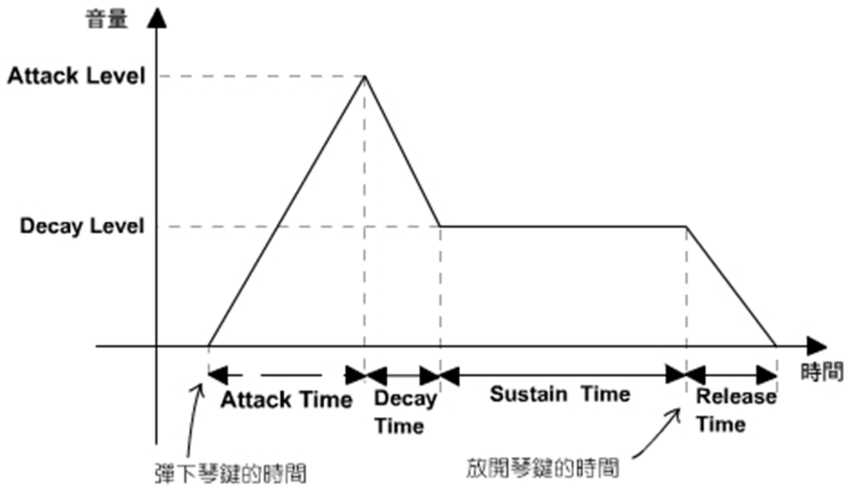
- There are 12 semitones in one octave, so a tone of frequency $f_1$ is said to be a semitone above a tone with frequency $f_2$ iff

$$f_1=2^{1/12}f_2=1.05946f_2$$

# Envelope



- 將一種音色波形的大致輪廓描繪出來，就可以表示出該音色在音量變化上的特性，而這個輪廓就稱為Envelope(波封)
- 一個波封可以用4種參數來描述，分別是Attack(起音)、Decay(衰減)、Sustain(延持)、與Release(釋音)，這四者也就是一般稱的"ADSR"。



# Envelop Extraction



- Full Wave Rectification

$$y[n] = \left| x[n] \right|$$

To extract the temporal envelope of the signal rather than the time domain signal itself

- Low-Pass Filtering (smoothing)

$$y[n] = (1-\alpha)x[n] + \alpha y[n-1], \quad \alpha = 0.99$$

To smooth the envelope

- Downsampling

$$y[n] = x[kn] \qquad k=16$$

Reduce the computation time

- Mean Removal

$$y[n] = x[n] - E[x[n]]$$

To make the signal centered to zero for the autocorrelation stage

# Enhanced Autocorrelation

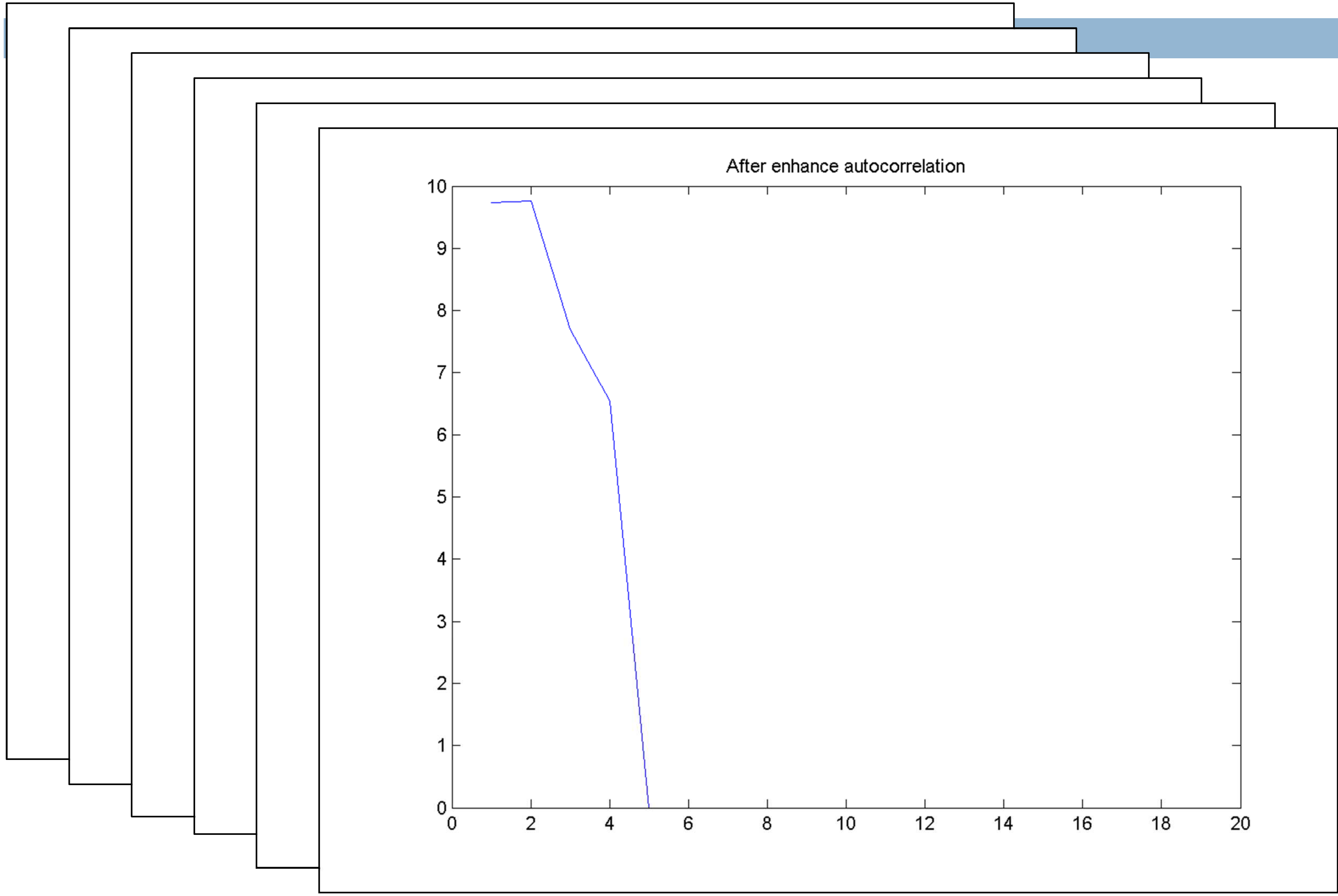


$$y[k] = \frac{1}{N} \sum_n x[n]x[n-k]$$

- The peaks of the autocorrelation function correspond to the time lags where the signal is most similar to itself
- The time lags correspond to beat periodicities

# Example




After enhance autocorrelation
10
9
8
7
6
5
4
3
2
1
0
0
2
4
6
8
10
12
14
16
18
20

# Peak Detection and Histogram Calculation

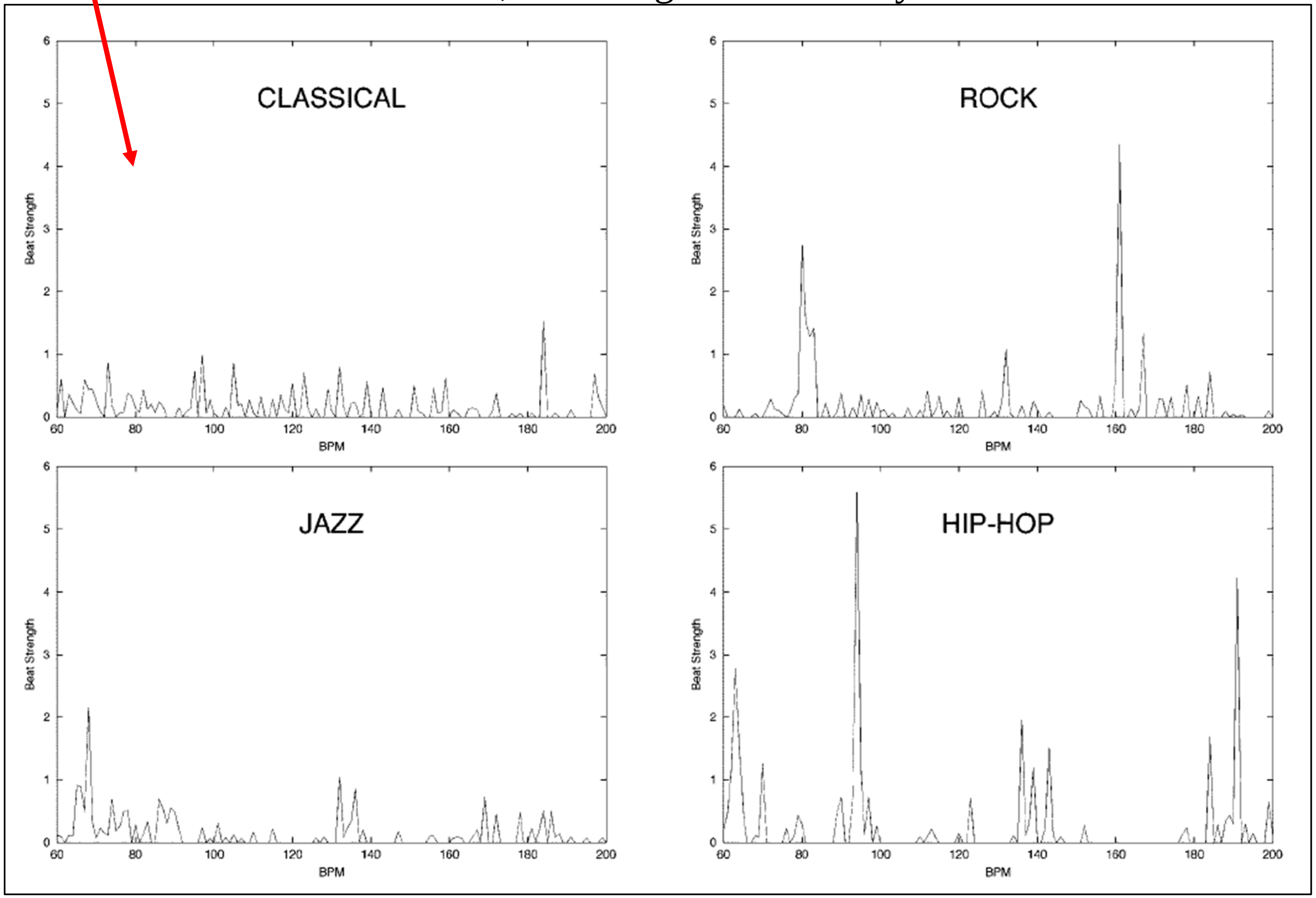


- The first three peaks of the enhanced autocorrelation function are selected and added to a beat histogram (BH).
- The bins of BH correspond to beats-per-minute (bpm) from 40 to 200 bpm.
- For each peak, the peak amplitude is added to the histogram.
  - Peaks having high amplitude (where the signal is highly similar) are weighted more strongly

# Beat Histogram



Multiple instruments of the orchestra, no strong self-similarity



# Beat Histogram Features



- A0, A1: relative amplitude (divided by the sum of amplitudes) of the first and second histogram peak
- RA: ratio of the amplitude of the second peak divided by the amplitude of the first peak
- P1, P2: period of the first and second peaks in bpm
- SUM: overall sum of the histogram (indication of beat strength)

# Introduction of Pitch



- Pitch (音高): 構成樂音的最基本要素在於音高，也就是聲音的頻率。
- 在樂理上，樂音音符可分為七個基本音，即【Do Re Me Fa Sol La Si】，以美式的符號則記為【C D E F G A B】而第八個音則稱為高八度的Do。

# Pitch Content Feature



- The signal is decomposed into two frequency bands (below and above 1000 Hz)
- Envelope extraction is performed for each frequency band.
- The envelopes are summed and an enhanced autocorrelation function is computed.
- The prominent peaks correspond to the main pitches for that short segment of sound.

# Beat and Pitch Detection



- The process of beat detection resembles pitch detection with larger periods.
- For beat detection, a window of 65536 samples at 22050 Hz is used.
- For pitch detection, a window of 512 samples is used.

Autocorrelation: $y[k] = \frac{1}{N} \sum_n x[n]x[n-k]$

different range of *k*

# Pitch Histogram



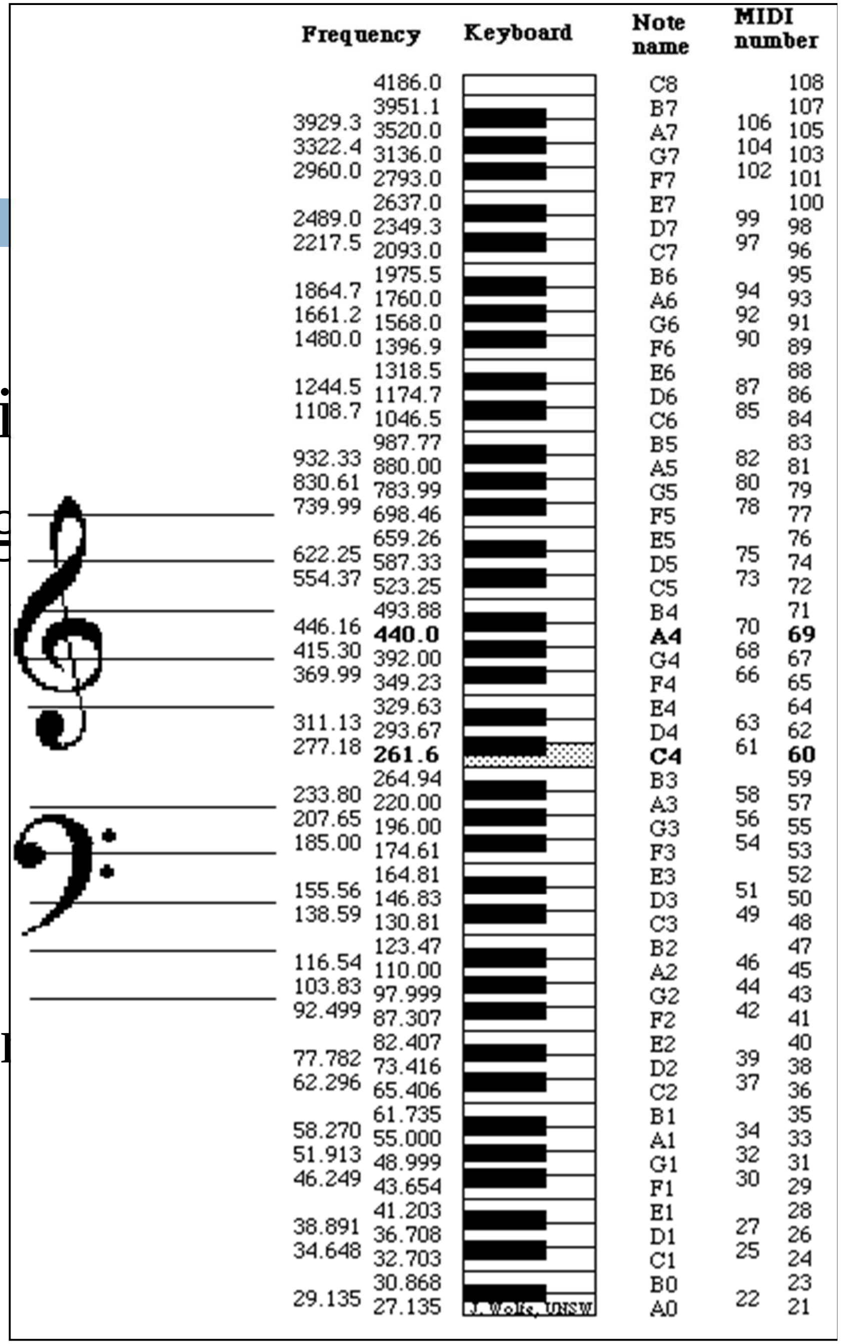
- For each analysis window, the
  are accumulated into a pitch hi
- The frequencies corresponding
  peak are converted to musical

$$n = 12 \times \log_2 \frac{f}{440} + 69$$

*f* is the frequency in Hertz
*n* is the histogram bin (MIDI note

http://www.phys.unsw.edu.au/~jw/notes.html

$$\left.\begin{matrix}\textbf{69}\\ \\ \textbf{70}\end{matrix}\right\}\ \textbf{semitone}$$



# Folded and Unfolded PH



- In the folded case (FPH)

  $c = n \bmod 12$

  $c$ is the folded histogram bin
  $n$ is the unfolded histogram bin

- The folded version (FPH) contains information regarding the pitch classes or harmonic content of the music. The unfolded version (UPH) contains information about the pitch range of the piece.

# Modified FPH



- The FPH is mapped to a circle of fifths histogram so that adjacent histogram bins are spaced a fifth apart rather than a semitone

$$c' = (7 \times c) \bmod 12$$

五度音程：三個全音加上一個半音的距離
G→全音→A→全音→B→半音→C→全音→D

- The distances between adjacent bins after mapping are better suited for expressing tonal music relations
- Jazz or classical music tend to have a higher degree of pitch change than rock or pop music.

# Pitch Histogram Features

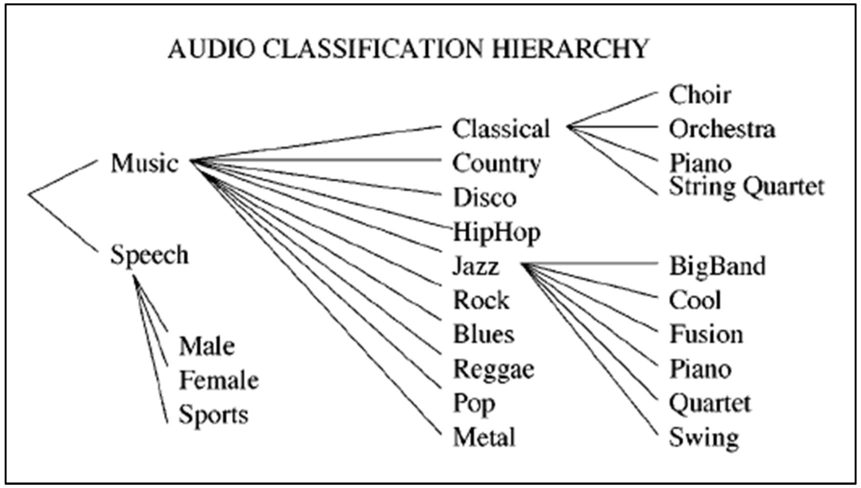


- FA0: amplitude of maximum peak of the folded histogram.
- UP0, FP0: period of the maximum peak of the unfolded and folded histograms
- IPO1: pitch interval between the two most prominent of the folded histogram (main tonal interval relation)
- SUM: the overall sum of the histogram

# Evaluation



- Classification
  - Simple Gaussian classifier
  - Gaussian mixture model
  - K-nearest neighbor classifier
- Datasets
  - 20 musical genres and 3 speech genres
  - 100 excerpts each with 30 sec
  - Taken from radio, CD, and mp3. The files were stored as 22050 Hz, 16-bit, mono audio files.



# Experiments



- Use a single-vector to represent the whole audio file.
- The vector consists of timbral texture features (9(FFT)+10(MFCC)=19-dim), rhythmic content features (6-dim), and the pitch content features (5-dim)
- 10-fold cross validation (90% training and 10% testing each time)

# Results

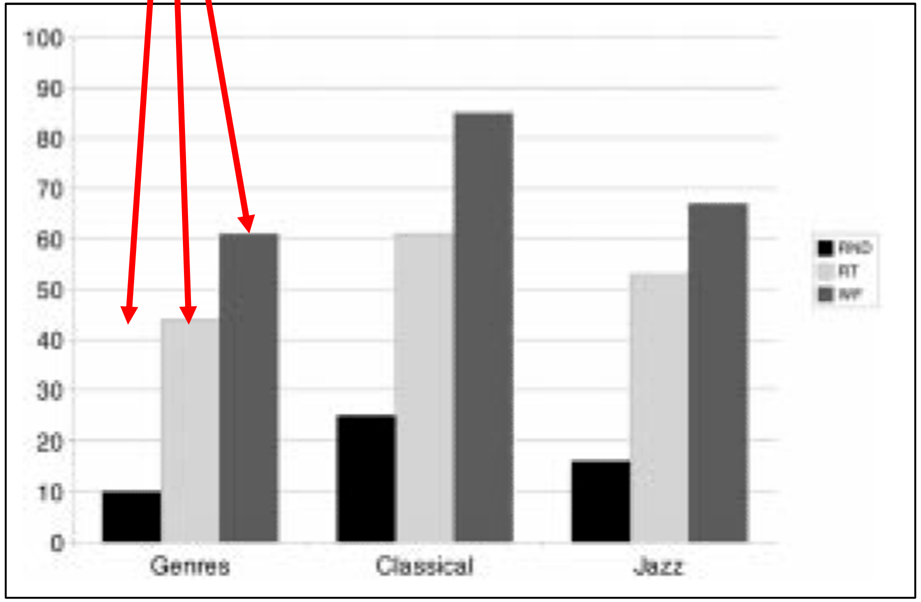


- RT GS: for real-time classification per frame using only timbral texture feature
- GS: simple Gaussian

**Random, RT GS, and GMM(3)**

TABLE I
CLASSIFICATION ACCURACY MEAN AND STANDARD DEVIATION

| | Genres(10) | Classical(4) | Jazz(6) |
|---|---|---|---|
| Random | 10 | 25 | 16 |
| RT GS | 44 ± 2 | 61 ± 3 | 53 ± 4 |
| GS | 59 ± 4 | 77 ± 6 | 61 ± 8 |
| GMM(2) | 60 ± 4 | 81 ± 5 | 66 ± 7 |
| GMM(3) | 61 ± 4 | 88 ± 4 | 68 ± 7 |
| GMM(4) | 61 ± 4 | 88 ± 5 | 62 ± 6 |
| GMM(5) | 61 ± 4 | 88 ± 5 | 59 ± 6 |
| KNN(1) | 59 ± 4 | 77 ± 7 | 57 ± 6 |
| KNN(3) | 60 ± 4 | 78 ± 6 | 58 ± 7 |
| KNN(5) | 56 ± 3 | 70 ± 6 | 56 ± 6 |



# Other Classification Results



- The STFT-based feature set is used for the music/speech classification
  - 86% accuracy
- The MFCC-based feature set is used for the speech classification
  - 74% accuracy

# Detailed Performance



TABLE II
GENRE CONFUSION MATRIX

| | cl | co | di | hi | ja | ro | bl | re | po | me |
|---|---|---|---|---|---|---|---|---|---|---|
| cl | **69** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| co | 0 | **53** | 2 | 0 | 5 | 8 | 6 | 4 | 2 | 0 |
| di | 0 | 8 | **52** | 11 | 0 | 13 | 14 | 5 | 9 | 6 |
| hi | 0 | 3 | 18 | **64** | 1 | 6 | 3 | 26 | 7 | 6 |
| ja | 26 | 4 | 0 | 0 | **75** | 8 | 7 | 1 | 2 | 1 |
| ro | 5 | 13 | 4 | 1 | 9 | **40** | 14 | 1 | 7 | 33 |
| bl | 0 | 7 | 0 | 1 | 3 | 4 | **43** | 1 | 0 | 0 |
| re | 0 | 9 | 10 | 18 | 2 | 12 | 11 | **59** | 7 | 1 |
| po | 0 | 2 | 14 | 5 | 3 | 5 | 0 | 3 | **66** | 0 |
| me | 0 | 1 | 0 | 1 | 0 | 4 | 2 | 0 | 0 | **53** |

**cl: classical**
**co: country**
**di: disco**
**hi: hiphop**
**ja: jazz**
**ro: rock**
**bl: blues**
**re: reggae**
**po: pop**
**me: mental**

26% of classical music is wrongly classified as jazz music

- The matrix shows that the misclassifications of the system are similar to what a human would do.

Rock music has worst accuracy because of its broad nature

# Performance on Classical and Jazz



TABLE III
JAZZ CONFUSION MATRIX

| | BBand | Cool | Fus. | Piano | 4tet | Swing |
|---|---|---|---|---|---|---|
| BBand | **42** | 2 | 1 | 0 | 6 | 1 |
| Cool | 21 | **67** | 5 | 4 | 23 | 10 |
| Fus. | 28 | 16 | **88** | 0 | 38 | 22 |
| Piano | 1 | 0 | 0 | **80** | 0 | 0 |
| 4tet | 4 | 5 | 2 | 0 | **19** | 5 |
| Swing | 4 | 10 | 4 | 16 | 14 | **62** |

BBand: bigband
Cool: cool
Fus.: fusion
Piano: piano
4tet: quartet (四重奏)
Swing: swing

TABLE IV
CLASSICAL CONFUSION MATRIX

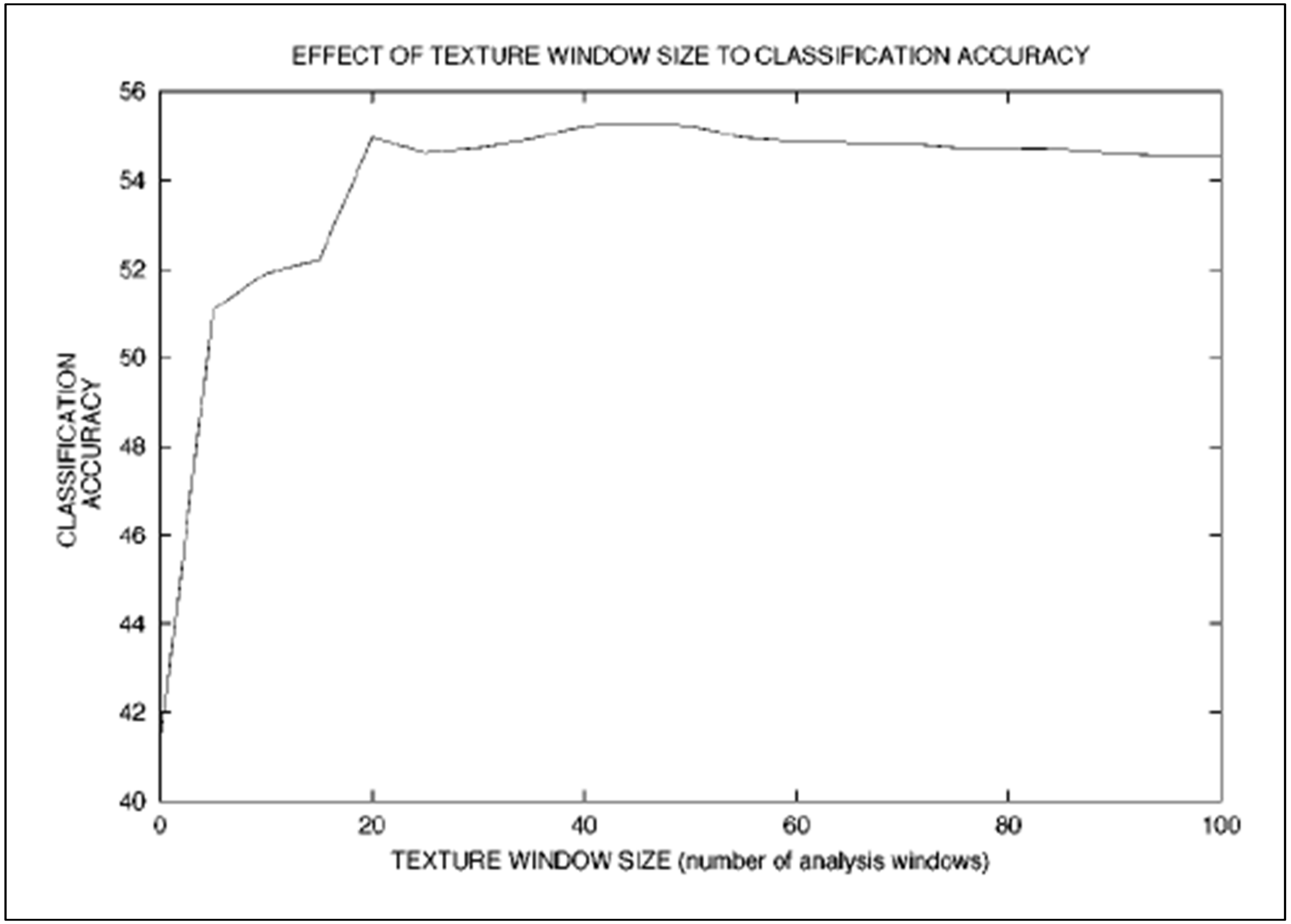
| | Choir | Orch. | Piano | Str.4tet |
|---|---|---|---|---|
| Choir | **99** | 7 | 7 | 3 |
| Orch. | 0 | **58** | 2 | 7 |
| Piano | 0 | 9 | **86** | 4 |
| Str.4tet | 1 | 26 | 5 | **86** |

Choir: choir
Orch.: orchestra
Piano: piano
Str.4tet: String Quarter
(弦樂四重奏)

# Importance of Texture Window Size



- 40 analysis windows was chosen



# Importance of Individual Feature Sets



- Pitch histogram features and beat histogram features perform worse than the timbral-texture features (STFT, MFCC)

TABLE V
INDIVIDUAL FEATURE SET IMPORTANCE

| | Genres | Classical | Jazz |
|---|---|---|---|
| RND | 10 | 25 | 16 |
| PHF(5) | 23 | 40 | 26 |
| BHF(6) | 28 | 39 | 31 |
| STFT(9) | 45 | 78 | 58 |
| MFCC(10) | 47 | 61 | 56 |
| FULL(30) | 59 | 77 | 61 |

The rhythmic and pitch content feature sets seem to play a less important role in the classical and jazz dataset classification

**It's possible to design genre-specific feature sets.**

# Human Performance for Genre Classification



- Ten genres used in previous study: blues, country, classical, dance, jazz, latin, pop, R&B, rap, and rock
- 70% correct after listening to 3 sec
- Although direct comparison of these results is not possible, it’s clear that the automatic performance is not far away from the human performance.

# Conclusion



- Three feature sets are proposed: timbral texture, rhythmic content, and pitch content features
- 61% accuracy has been achieved
- Possible improvements:
  - Information from melody and singer voice
  - Expand the genre hierarchy both in width and depth
  - More exploration of pitch content features
- MARSYAS: http://webhome.cs.uvic.ca/~gtzan/

43

# Audio Effects Detection

Wei-Ta Chu

2014/11/19

R. Cai, L. Lu, and H.-J. Zhang, “Highlight sound effects detection in audio stream,” Proc. of ICME, 2003, pp. 37-40.

# Introduction



- Model and detect three sound effects: laughter, applause, and cheer
- Sound effect detection must handle the following cases:
  - Model more particular sound classes
  - Recall the expected sound effects only and ignore others
- Characteristics:
  - High recall and precision
  - Extensibility: it should be easy to add or remove sound effect models for new requirements.

# Audio Feature Extraction



- All audio streams are 16-bit, mono-channel, and down-sampled to 8kHz. Each frame is of 200 samples (25 ms), with 50% overlaps.
- Features
  - Short-time energy
  - Average ZCR
  - Sub-band energies
  - Brightness and bandwidth
  - 8 order MFCC
- These features form a 16-dimensional feature vector for a frame. To describe the variance btw frames, the gradient feature of adjacent frames is also considered, and is concatenated to the original vector. Thus we have a 32-dim feature vector for each frame.

# Sound Effect Modeling



- HMMs can describe the time evolution between states using the transition probability matrix.
- A complete connected HMM is used for each sound effect, with the 4 continuous Gaussian mixtures modeling each state.
- Training data: 100 pieces of samples segmented from audio-track. Each piece is about 3s-10s and totally about 10 min training data for each class.
- A clustering algorithm is used to determine the state numbers of HMM.
  - 2 for applause, and 4 for cheer and laughter

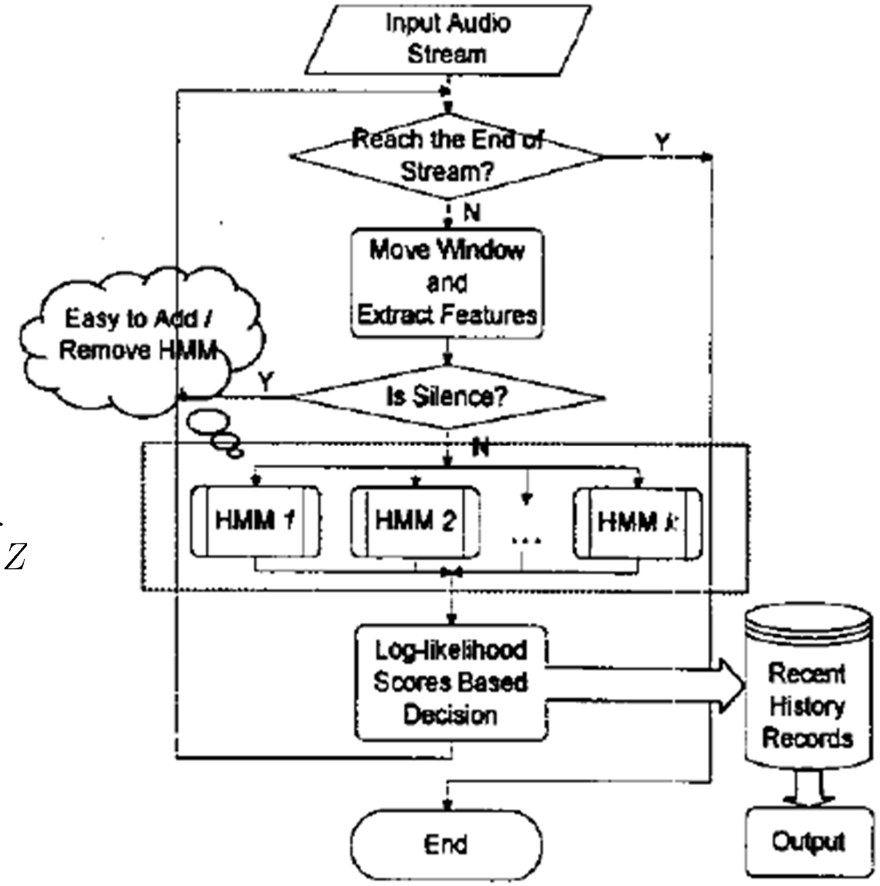
# Sound Effect Detection



- 1s moving window with 50% overlapping
- Each data window is further divided into 25ms frames with 50% overlapping
- Silence window is skipped

$$AverageSTE < \delta_E \ \& \ AverageZCR < \delta_Z$$

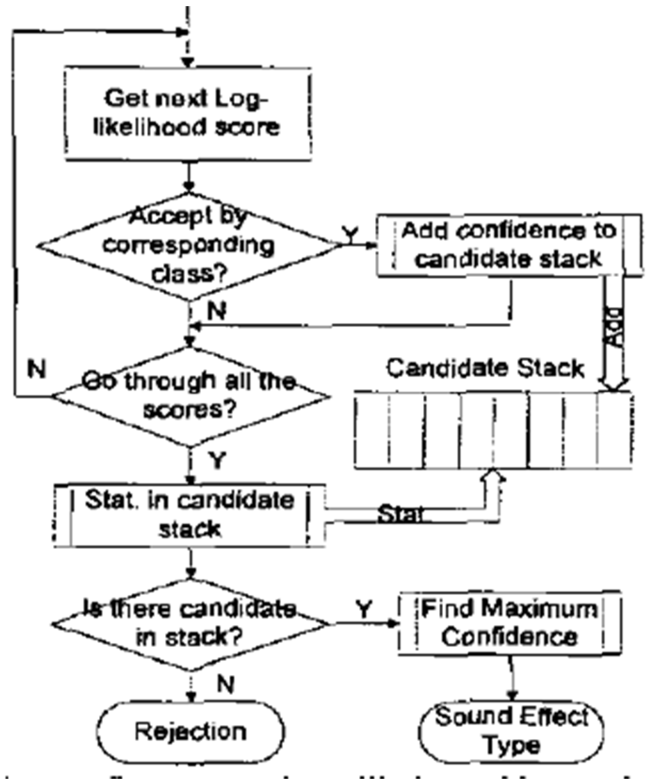
- Non-silence window is compared against each sound effect model to get likelihood score $P(\boldsymbol{O}|\lambda)$ or $\log P(\boldsymbol{O}|\lambda)$



# Log-Likelihood Scores Based Decision Method



- Unlike audio classification, we can't simply classify the sliding window into the class which has the maximum log-likelihood score.
- Each log-likelihood score is examined to see if the window data is "accepted" by the corresponding sound effect.
- Optimal decision based on Bayesian decision theory



# Log-Likelihood Scores Based Decision Method



- Cost function

$$C = P(C_j)P(\overline{C}_j|C_j)C_{\overline{C}_j|C_j} + P(\overline{C}_j)P(C_j|\overline{C}_j)C_{C_j|\overline{C}_j}$$

$C_{\overline{C}_j|C_j}$ is the cost of false rejection

$C_{C_j|\overline{C}_j}$ is the cost of false acceptance

- To minimize the cost, use Bayesian decision rule

$$\frac{p(s_j|C_j)}{p(s_j|\overline{C}_j)} \geq R_j$$

(likelihood ratio)

$s_j$ is the log-likelihood score under HMM of sound effect $C_j$

$p(s_j|C_j)$ and $p(s_j|\overline{C}_j)$ are probability distributions of log-likelihood scores of the samples within and outside $C_j$, respectively

(likelihood function)

# Log-Likelihood Scores Based Decision Method



- Bayesian threshold:

$$R_j = \frac{C_{C_j|\overline{C}_j}}{C_{\overline{C}_j|C_j}} \times \frac{P(\overline{C}_j)}{P(C_j)}$$

  - The priori probabilities are estimated based on the database.
  - The cost of FR is set larger than that of FA, given that a high recall ratio is more important for summarization and highlight extraction

# Likelihood Function



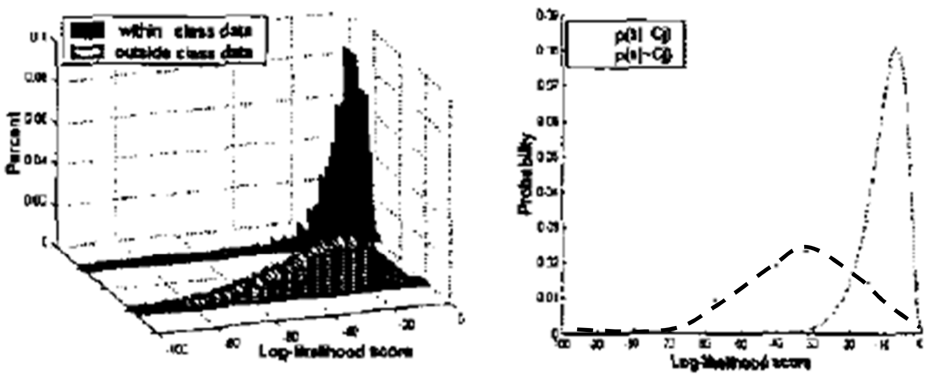
$p(s_j|C_j)$ and $p(s_j|\overline{C}_j)$ are estimated from the database.

- The distribution of samples within and outside the sound effect applause.
- To approximate these distributions (asymmetric), it's more reasonable to use negative Gamma distribution

$$p(s|C_j) = -\frac{1}{\beta^{\alpha}\Gamma(\alpha)} s^{\alpha-1} e^{s/b}$$

$$\alpha = \mu^2/\sigma^2$$

$$\beta = \sigma^2/\mu$$



(a) (b)

**Figure 3.** (a) log-likelihood scores distributions; (b) Approximate (a) with Gamma distribution

# Decision

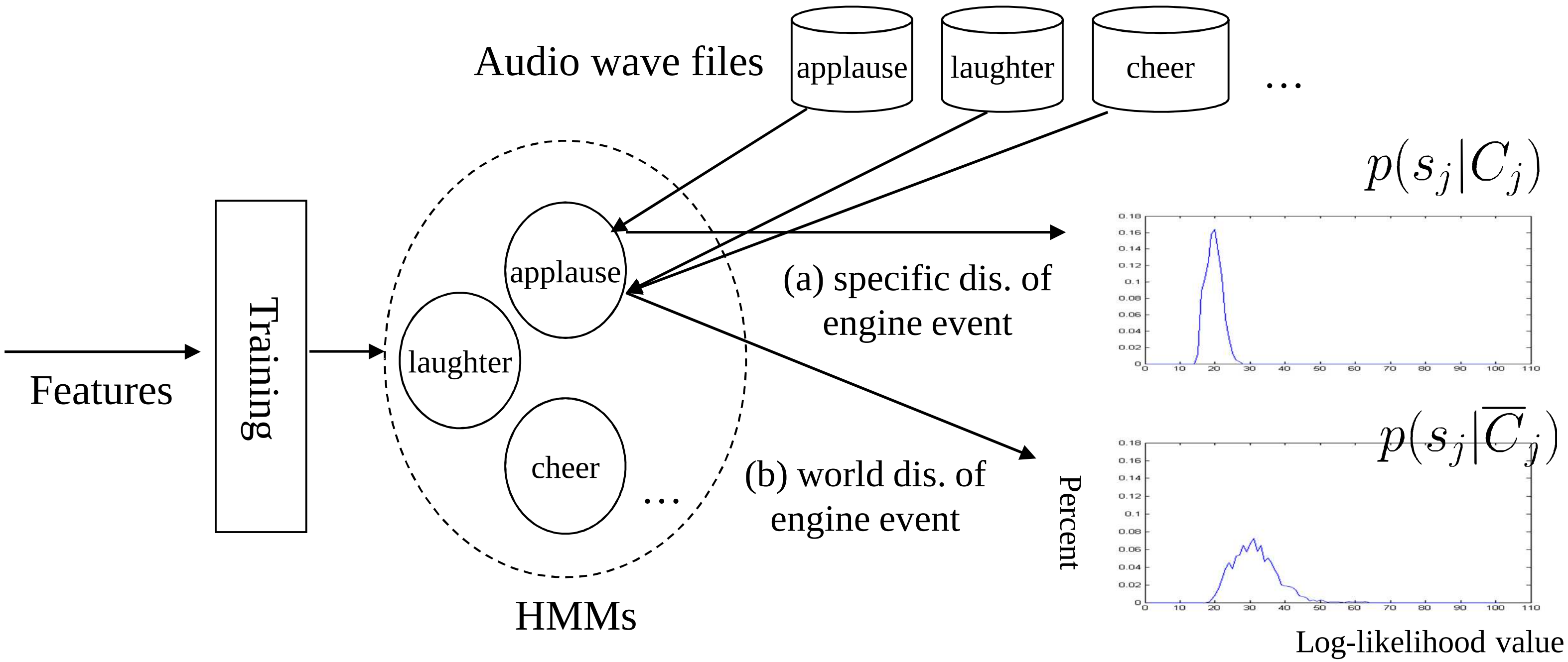


- Abnormal scores are pruned first
  - Score whose distance to $\mu$ are larger than $2\sigma$ are abnormal
- The windows that confirms to $\frac{p(s_j|C_j)}{p(s_j|\overline{C}_j)} \geq R_j$ are considered to be "accepted" by a sound effect.
- If it is accepted by a sound effect, the corresponding likelihood score is considered as confidence.
- It is classified into the *i*th sound effect if

$$i = \arg\max_j p(s_j|C_j)$$

# Overall






The confidence score of an audio segment:
(based on likelihood ratio)

$$\frac{p(s_j|C_j)}{p(s_j|\overline{C}_j)} \geq R_j$$

# Sound Effect Attention Model



- Audio attention model is constructed to describe the saliency of each sound effect
- Based on energy and confidence in sound effects

$$\overline{E} = E_{avr} / Max\_E_{avr}$$

$$\overline{P}_j = \exp(s_j - Max\_s_j)$$

$E_{avr}$ and $s_j$ denote the average energy and log-likelihood score under model $j$ of an audio segment

- The attention model for class *j* is defined as

$$M_j = \overline{E} \times \overline{P}_j$$

# Sound Effect Attention Model



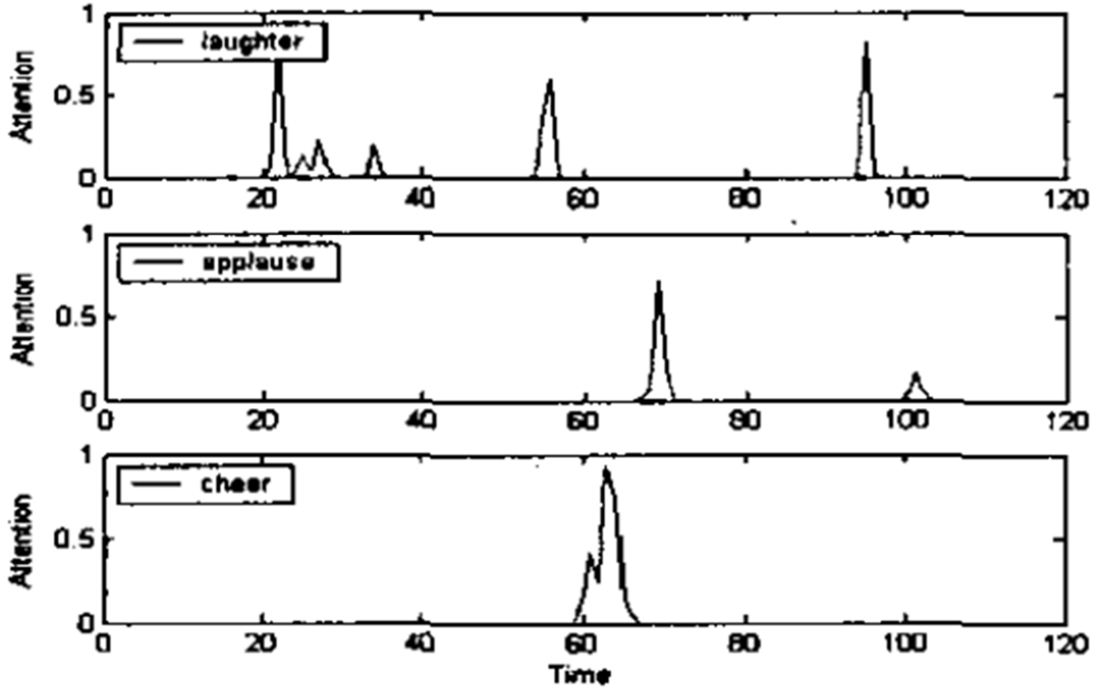



**Figure 4.** Sound effect attention model curves

# Experiments



- The testing database is about 2 hours videos, including NBC's TV show (30 min), CCTV's TV show (60 min), and table tennis (30 min).
- Two kind of distribution curves – Gaussian and Gamma – are compared.
  - Gamma distribution increase the precision by 9.3%, while just affects the recall ratio by 1.8%.

**Table I.** Performance on different *p.d.f.* distribution

| *p.d.f.* | Sound Effect | Recall | Precision |
|---|---|---|---|
| *Gaussian* | laughter | 0.959 | 0.791 |
| | applause | 0.933 | 0.668 |
| | cheer | 0.907 | 0.906 |
| *Gamma* | laughter | 0.927 | 0.879 |
| | applause | 0.910 | 0.850 |
| | cheer | 0.907 | 0.916 |

# Experiments



- Average recall is 92.95% and average precision is 86.88%.
- Higher recall can meet the requirements for highlights extraction and summarization.
- In table tennis, reporters' exciting voice would be detected as laughter. Moreover, sound effects are often mixed with music, speech, and other environment sounds.

**Table 2.** Performance of the algorithm

| Video | Sound Effect | Recall | Precision |
|---|---|---|---|
| Hollywood Square | laughter | 0.927 | 0.879 |
| | applause | 0.910 | 0.850 |
| | cheer | 0.907 | 0.916 |
| Lucky 52 | laughter | 0.956 | 0.813 |
| | applause | 0.894 | 0.826 |
| | cheer | 0.910 | 0.917 |
| Table Tennis Championship | laughter | 0.977 | 0.778 |
| | applause | 0.956 | 0.945 |
| | cheer | 0.957 | 0.946 |

# References




- G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Trans. on Speech and Audio Processing, vol. 10, no. 5, 2002, pp. 293-302.
- R. Cai, L. Lu, and H.-J. Zhang, “Highlight sound effects detection in audio stream,” Proc. of ICME, 2003, pp. 37-40.
- L. Lu, R. Cai, and A. Hanjalic, “Towards a unified framework for content-based audio analysis,” Proc. of ICASSP, vol. 2, 2005, pp. 1069-1072.
- M.A. Bartsch and G.H. Wakefield, “Audio thumbnailing of popular music using chroma-based representations,” IEEE Trans. on Multimedia, vol. 7, no. 1, 2005, pp. 96-104.