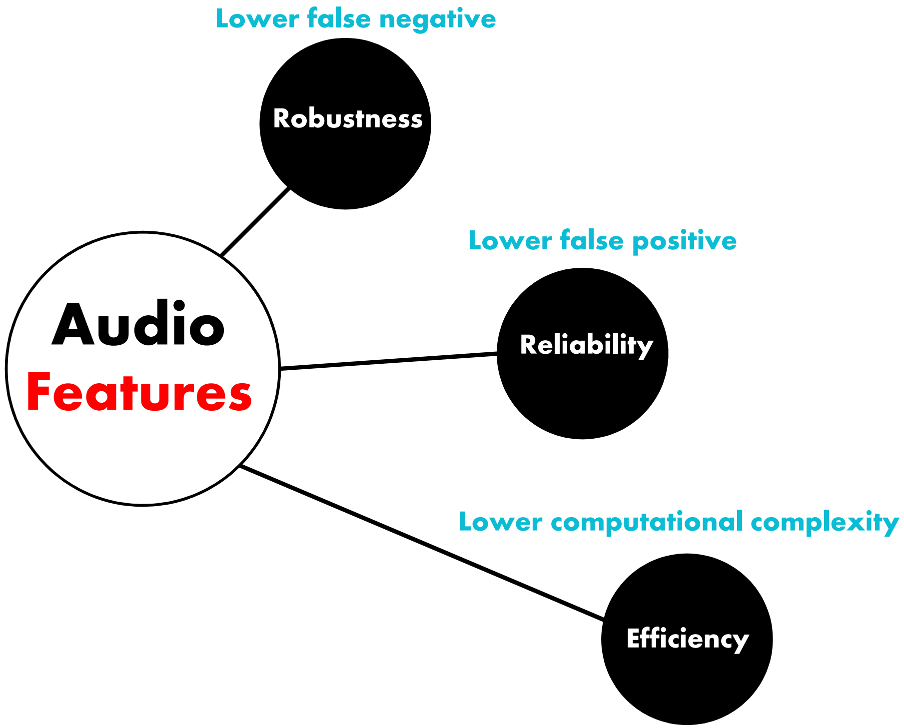
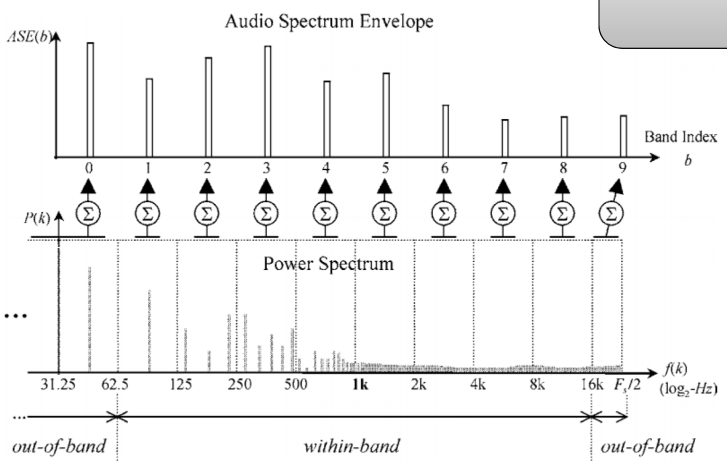
Audio & Speech Technology

[4] Audio Features



Audio Spectral Envelope

$$ASE[b] = \sum_{k \in Band[b]} P[k]$$



$$P[k] = |S[k]|^2$$

$$S[k] = Spectrum$$

MPEG-7 Audio and. Beyond. Audio Content Indexing and. Retrieval. Hyoung-Gook Kim, et. al.

Audio Spectral Flatness

$$ASF[b] = \frac{\sqrt[N_b]{\prod_{k \in Band[b]} P[k]}}{\frac{1}{N_b} \sum_{k \in Band[b]} P[k]}$$

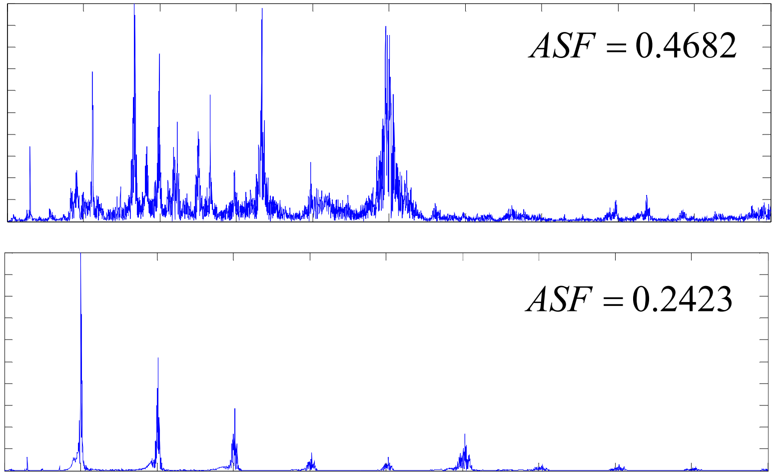
➡ Geometric Mean

➡ Arithmetic Mean

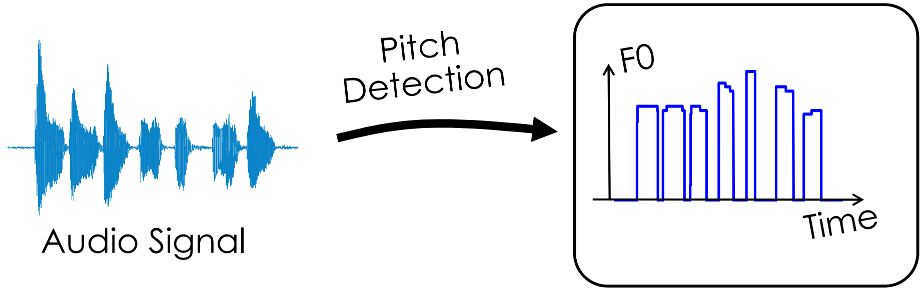
N_b = No. of frequencies in each band

$$\sqrt[N_b]{\prod_{k \in Band[b]} P[k]} = \exp\left(\frac{1}{N_b} \sum_{k \in Band[b]} \ln P[k] \right)$$

Audio Spectral Flatness



Audio Fundamental Frequency

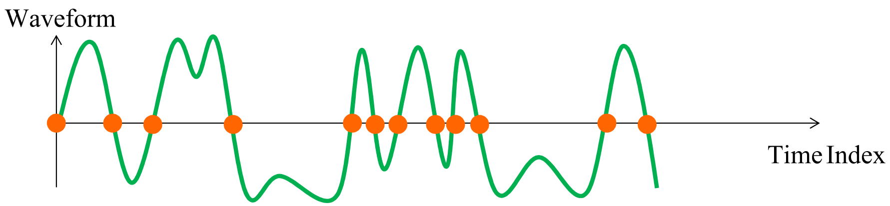


Zero Crossing Rate

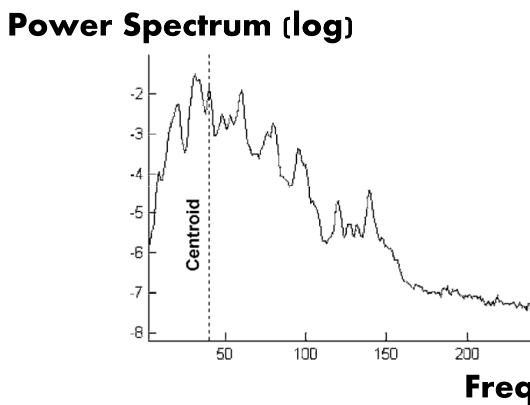
$$ZCR = \frac{F_S}{2N} \sum_{n=1}^{N-1} \left| sign(s[n]) - sign(s[n-1]) \right|$$

$N =$ No. samples in $s[i]$

$F_S =$ Sampling Frequency



Spectral Centroid



$$SC = \frac{\sum_k f[k]P[k]}{\sum_k P[k]}$$

Harmonic Ratio

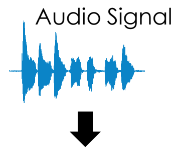
$$R[m]=\frac{\sum_{n\in Frame} s[n]s[n-m]}{\sqrt{\sum_{n\in Frame} s^2[n]\ \sum_{n\in Frame} s^2[n-m]}}\Rightarrow HR=\max_{m\geq M_0} R[m]$$

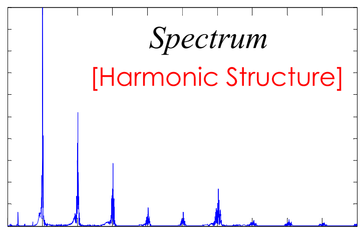
R = Autocorrelation Function

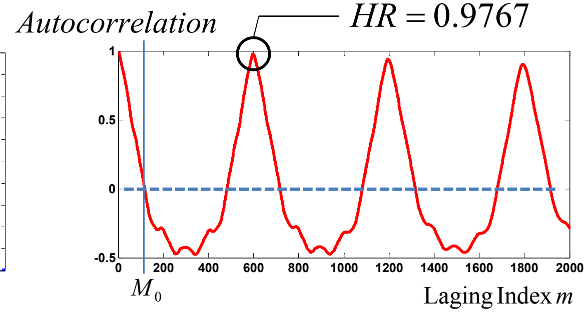
m = Lagging index

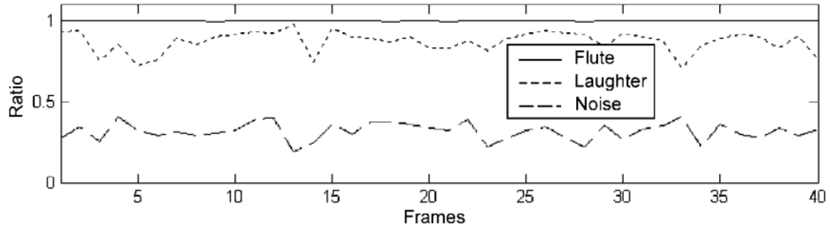
M_0 = Position of the first zerocrossing of autocorrelation R

Harmonic Ratio

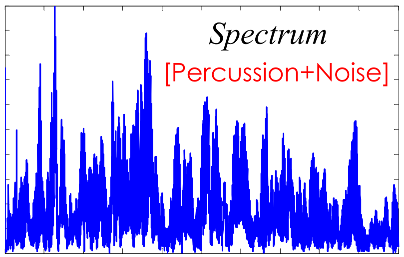


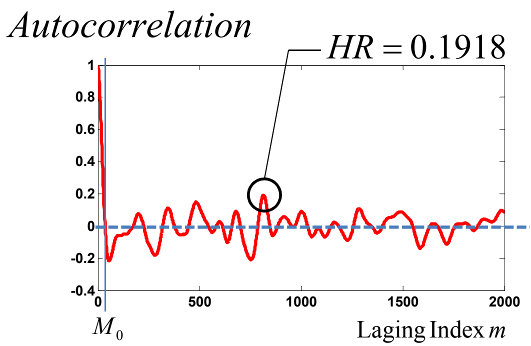




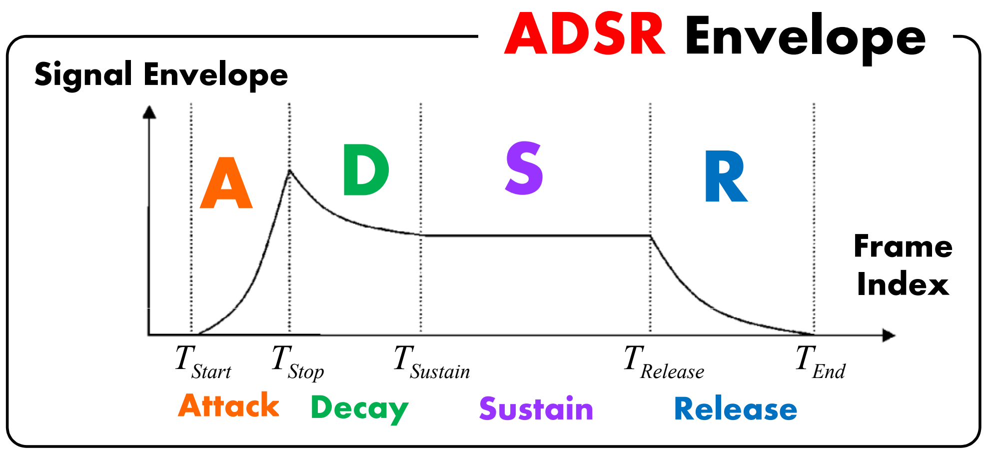


Harmonic Ratio





Log Attack Time

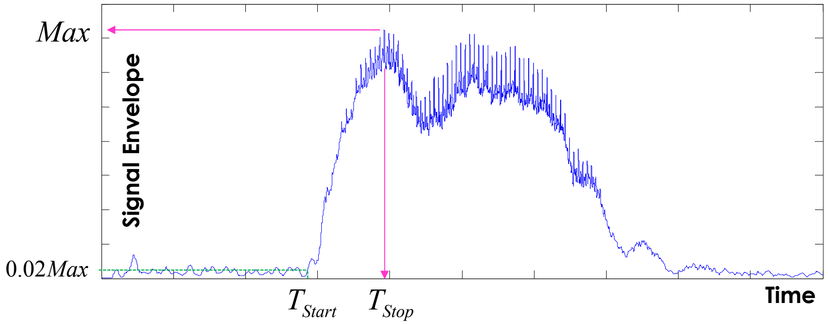


Log Attack Time

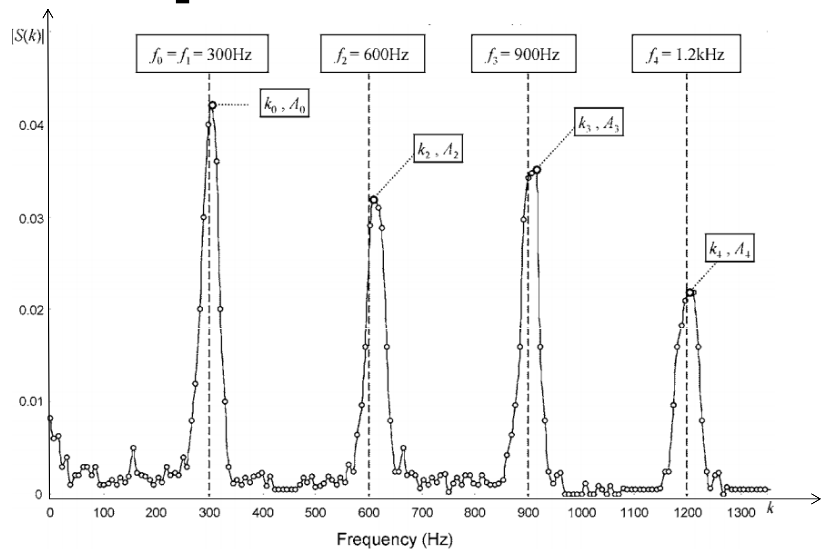
$$LAT = \log_{10}\left(T_{Stop} - T_{Start}\right)$$

T_{Start} = The time the signal envelope exceeds 2% of its maximal value

T_{Stop} = The time the signal evelope reaches its maximal value



Harmonic Spectral Centroid



MPEG-7 Audio and. Beyond. Audio Content Indexing and. Retrieval. Hyoung-Gook Kim, et. al.

Harmonic Spectral Centroid

$LHSC$ = Local Harmonic Spectral Centroid of Audio Frame

$$LHSC = \frac{\sum_{h=1}^{N_H} f_h A_h}{\sum_{h=1}^{N_H} A_h}$$

f_h = Frequency of the h^{th} harmonic

A_h = Amplitude of the h^{th} harmonic

N_H = No. of harmonic peaks

$$HSC = \frac{1}{L}\sum_{i=1}^{L} LHSC[i]$$

L = No. of time frames

Harmonic Spectral Deviation

Spectral Envelope

$$SE_h = \begin{cases} (A_h + A_{h+1})/2 & ; h = 1 \\ (A_{h-1} + A_h + A_{h+1})/2 & ; h \in [2, N_H - 1] \\ (A_{h-1} + A_h)/2 & ; h = N_H \end{cases}$$

$$LHSD = \frac{\sum_{h=1}^{N_H} \left|\log_{10} A_h - \log_{10} SE_h\right|}{\sum_{h=1}^{N_H} \log_{10} A_h} \Rightarrow HSD = \frac{1}{L}\sum_{i=1}^{L} LHSD[i]$$

Harmonic Spectral Spread

$$LHSS = \sqrt{\frac{\sum_{h=1}^{N_H} (f_h - LHSC)^2 A_h^2}{\sum_{h=1}^{N_H} A_h^2}}$$

⬇

$$HSS = \frac{1}{L}\sum_{i=1}^{L} LHSS[i]$$

Harmonic Spectral Variation

$$LHSV[i] = 1 - \frac{\sum_{h=1}^{N_H} A_h[i-1]A_h[i]}{\sqrt{\sum_{h=1}^{N_H} A_h^2[i-1]}\sqrt{\sum_{h=1}^{N_H} A_h^2[i]}}$$

⬇

$$HSS = \frac{1}{L}\sum_{i=1}^{L} LHSS[i]$$

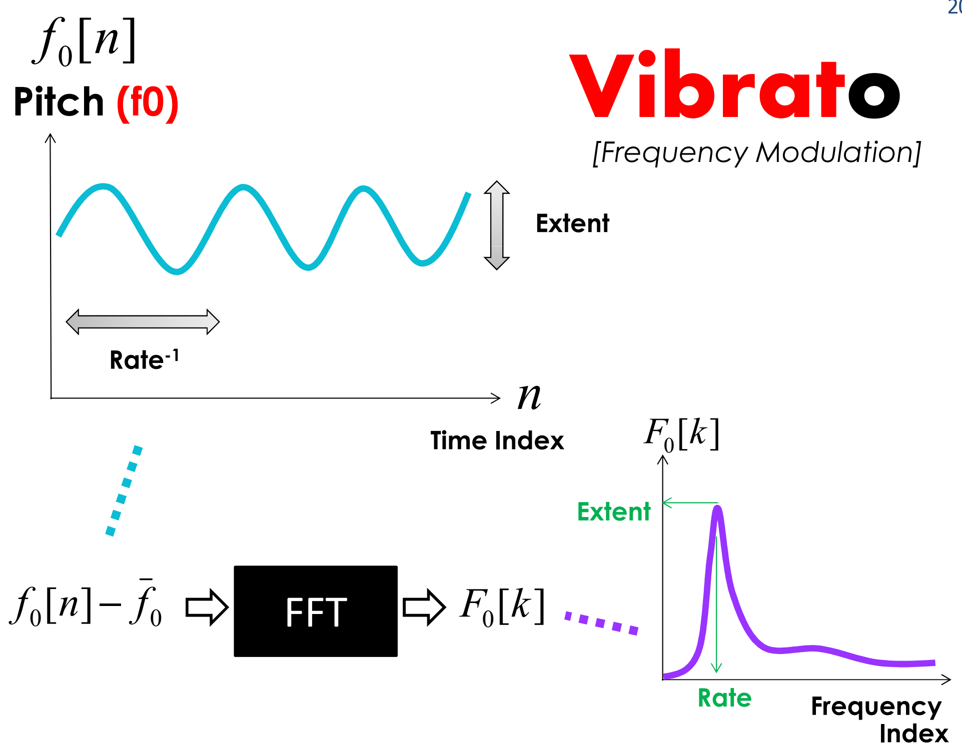
Vibrato

[Frequency Modulation]

- Variation of pitch.
- **Extent of vibrato**
  = Amount of pitch variation
- **Rate of vibrato**
  = Speed which the pitch is varied
- String instruments produce the FM dominant sounds

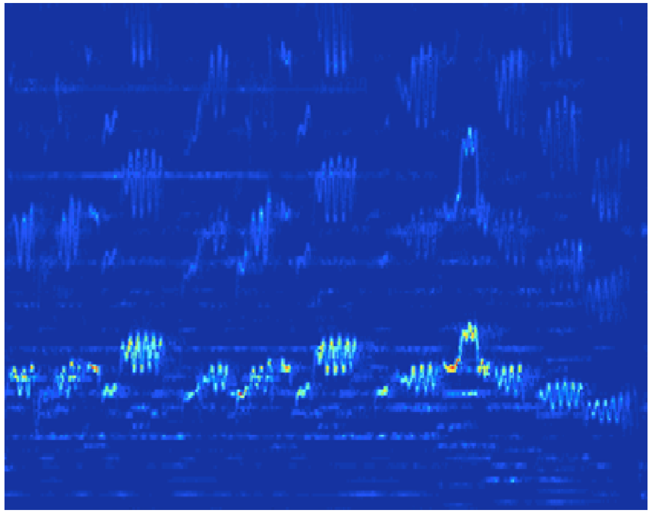
Vibrato

[Frequency Modulation]



Vibrato

[Frequency Modulation]



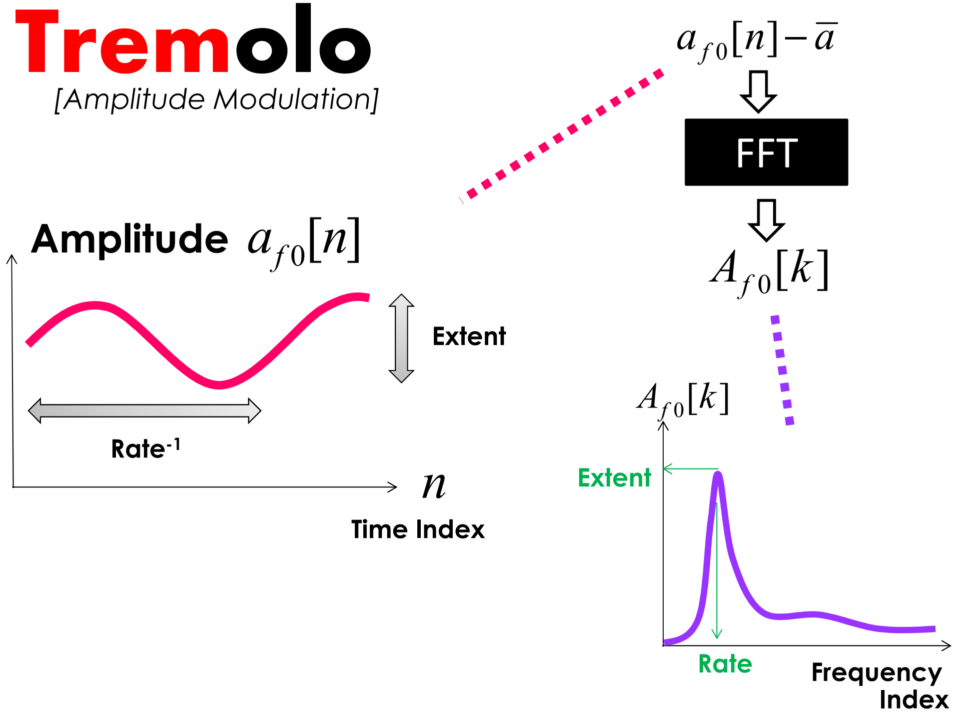
Tremolo

[Amplitude Modulation]

- Variation of sound intensity.
- **Extent of vibrato**
  = Amount of intensity variation
- **Rate of vibrato**
  = Speed which the intensity is varied
- Wind and brass instruments produce AM dominant sounds.

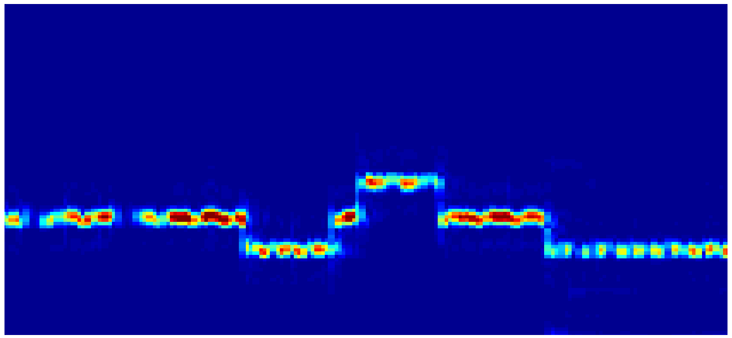
Tremolo

[Amplitude Modulation]



Tremolo

[Amplitude Modulation]



Linear Predictive Coding

- A model used for predicting a value of current sample of signal from the previous samples
- Used in Speech Analysis, Speech Synthesis, Audio Classification, Audio Compression

Current sample

Previous samples

$$s[n] = \sum_{k=1}^{P} a_k s[n-k] + \varepsilon[n]$$

LPC Coefficients

Prediction Error

Linear Predictive Coding

$s[n]$

Training Signal

$\{a_k\}$

Training LPC

- Find $\{a_k\}$
- To minimize $\sum_n \varepsilon^2[n]$

$$\varepsilon[n] = s[n] - \hat{s}[n]$$

$$\hat{s}[n] = \sum_{k=1}^{P} a_k s[n-k]$$

Linear Predictive Coding

$$E = \sum_n \varepsilon^2[n] = \sum_n \left(s[n] - \sum_{k=1}^{P} a_k s[n-k] \right)^2$$

$$\frac{\partial E}{\partial a_i} = \sum_n (-2s[n-i]) \left(s[n] - \sum_{k=1}^{P} a_k s[n-k] \right)$$

$$0 = \sum_n s[n]s[n-i] - \sum_n s[n-i] \sum_{k=1}^{P} a_k s[n-k]$$

$$0 = \sum_n s[n]s[n-i] - \sum_{k=1}^{P} a_k \sum_n s[n-k]s[n-i]$$

Linear Predictive Coding

$$0 = \sum_n s[n]s[n-i] - \sum_{k=1}^{P} a_k \sum_n s[n-k]s[n-i]$$

$$0 = R[i] - \sum_{k=1}^{P} a_k R[k-i]$$

$$\sum_{k=1}^{P} a_k R[k-i] = R[i]$$

Autocorrelation Function $R[i-j] = R[j-i] = \dfrac{\sum_n s[n-i]s[n-j]}{\sum_n s^2[n]}$

Linear Predictive Coding

$$\frac{\partial E}{\partial a_1}=0 \Rightarrow \sum_{k=1}^{P} a_k R[k-1]=R[1]$$

$$\frac{\partial E}{\partial a_2}=0 \Rightarrow \sum_{k=1}^{P} a_k R[k-2]=R[2]$$

$$\vdots$$

$$\frac{\partial E}{\partial a_P}=0 \Rightarrow \sum_{k=1}^{P} a_k R[k-P]=R[P]$$

Linear Predictive Coding

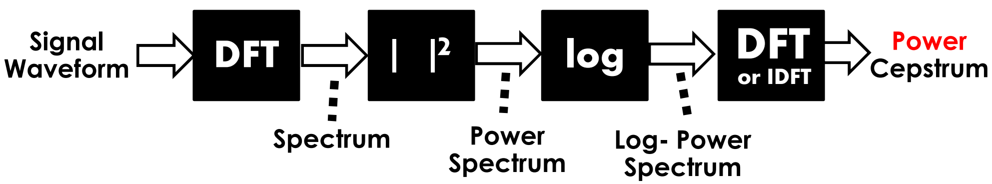
$$\begin{bmatrix} R[0] & R[1] & R[2] & \cdots & R[P-1] \\ R[1] & R[0] & R[1] & \cdots & R[P-2] \\ R[2] & R[1] & R[0] & \cdots & R[P-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R[P-1] & R[P-2] & R[P-3] & \cdots & R[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R[1] \\ R[2] \\ R[3] \\ \vdots \\ R[P] \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R[0] & R[1] & R[2] & \cdots & R[P-1] \\ R[1] & R[0] & R[1] & \cdots & R[P-2] \\ R[2] & R[1] & R[0] & \cdots & R[P-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R[P-1] & R[P-2] & R[P-3] & \cdots & R[0] \end{bmatrix}^{-1} \begin{bmatrix} R[1] \\ R[2] \\ R[3] \\ \vdots \\ R[P] \end{bmatrix}$$

Cepstrum

Fourier transform of the logarithm of the spectrum of signal

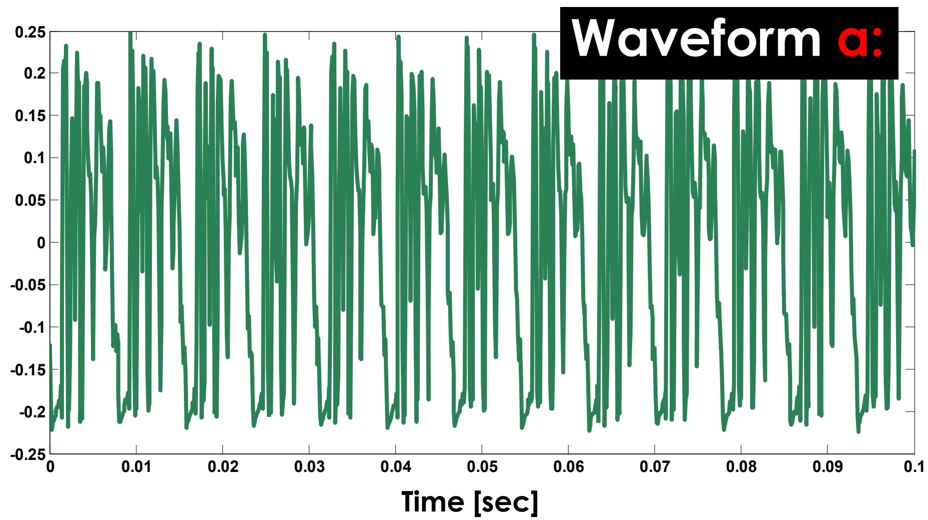


Spectrum ⇒ Cepstrum

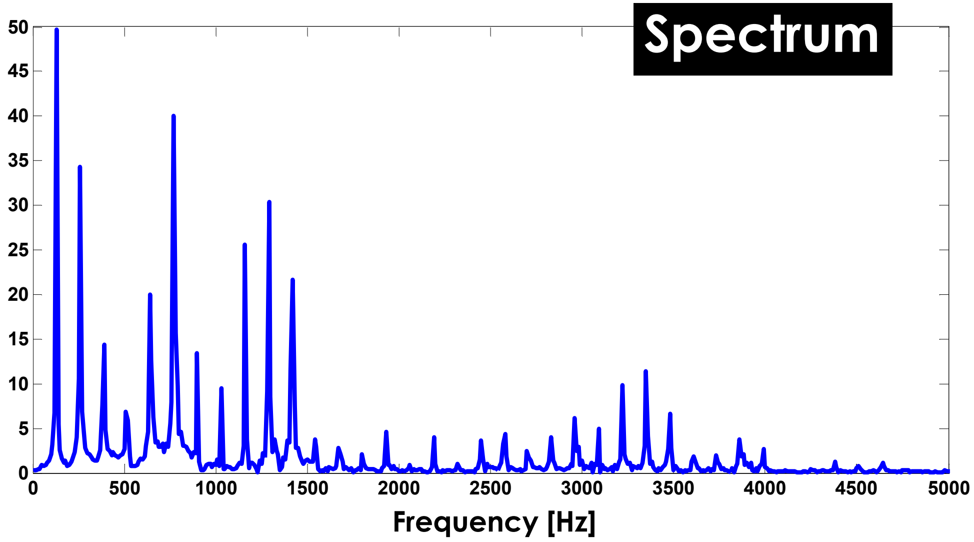
Frequency ⇒ Quefrency

Filter ⇒ Lifter

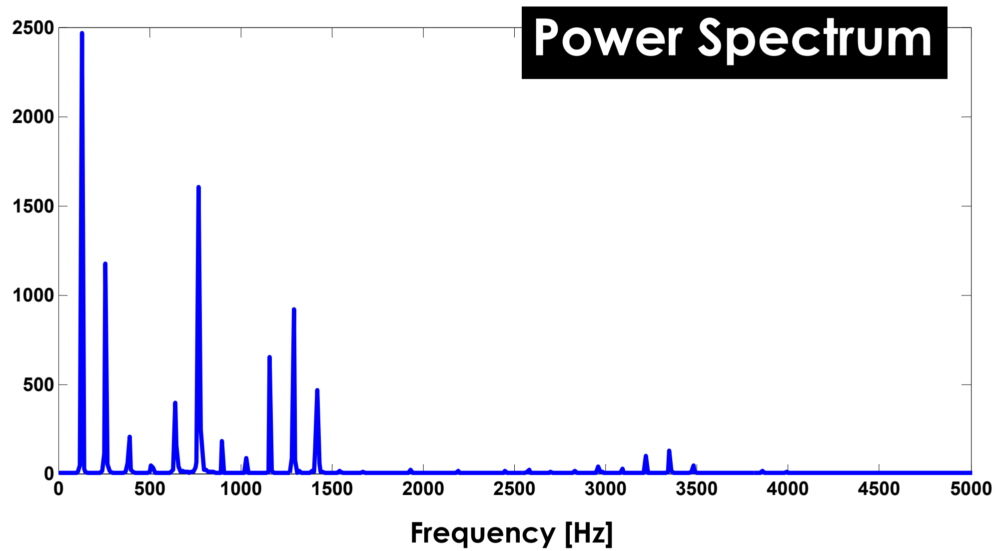
Cepstrum



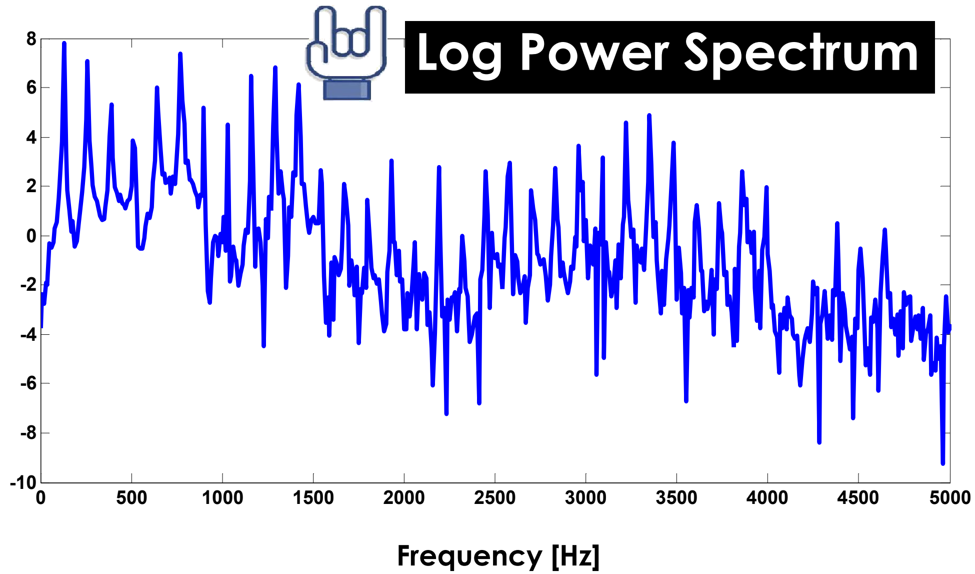
Cepstrum



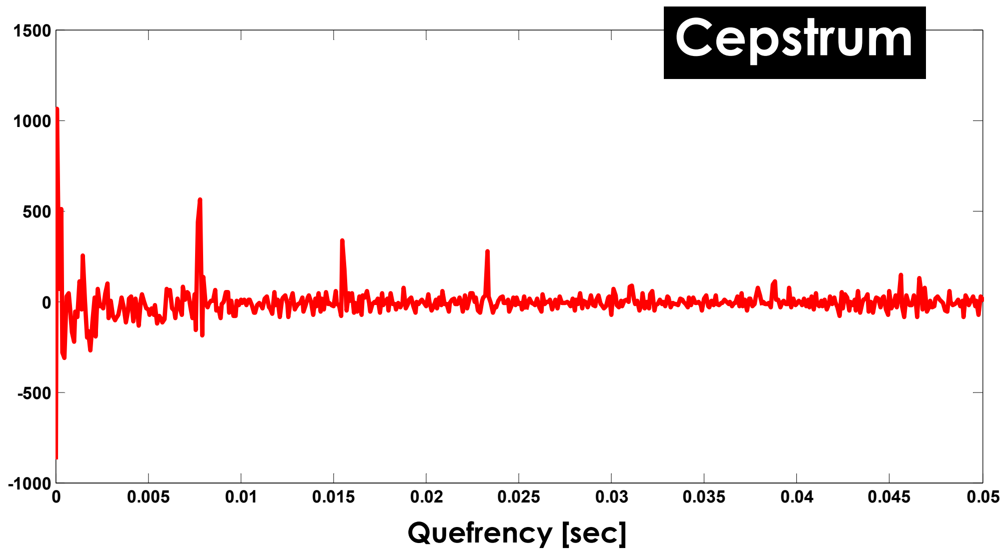
Cepstrum



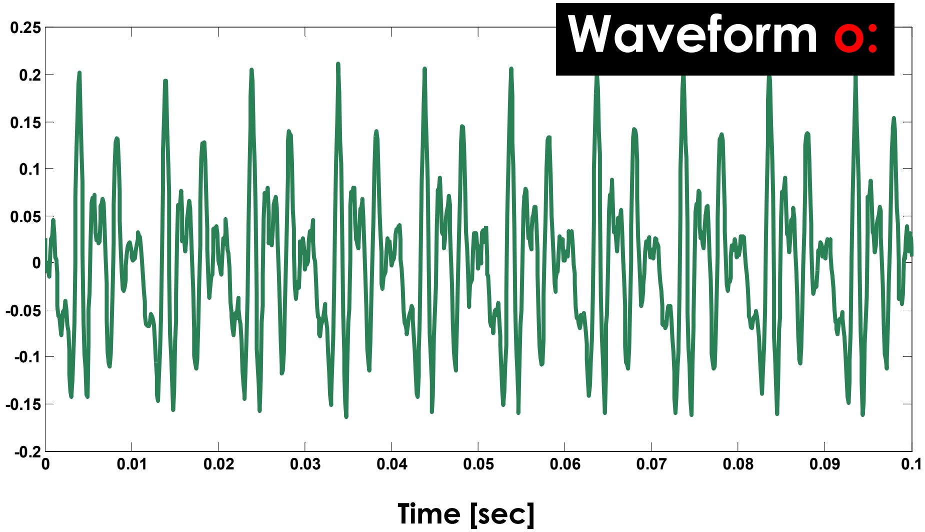
Cepstrum



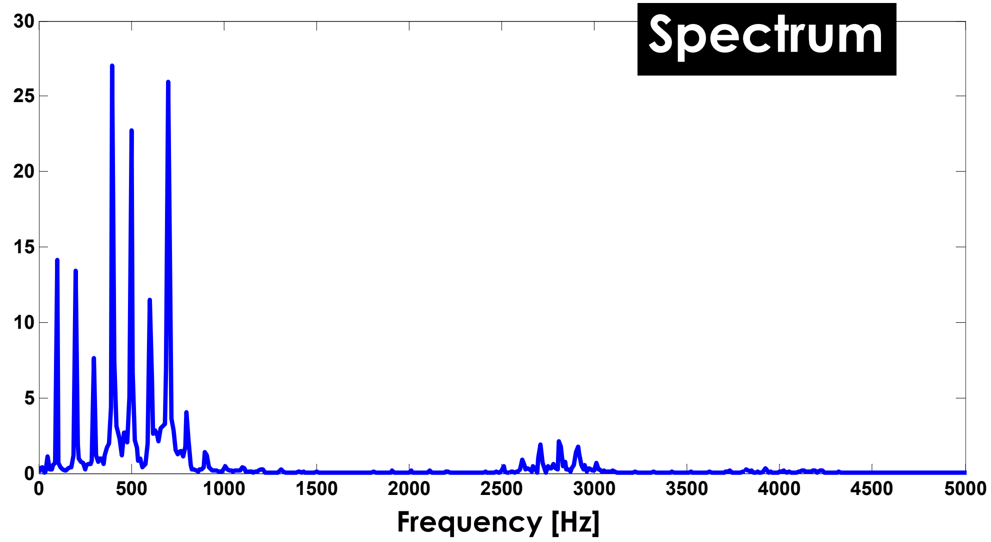
Cepstrum



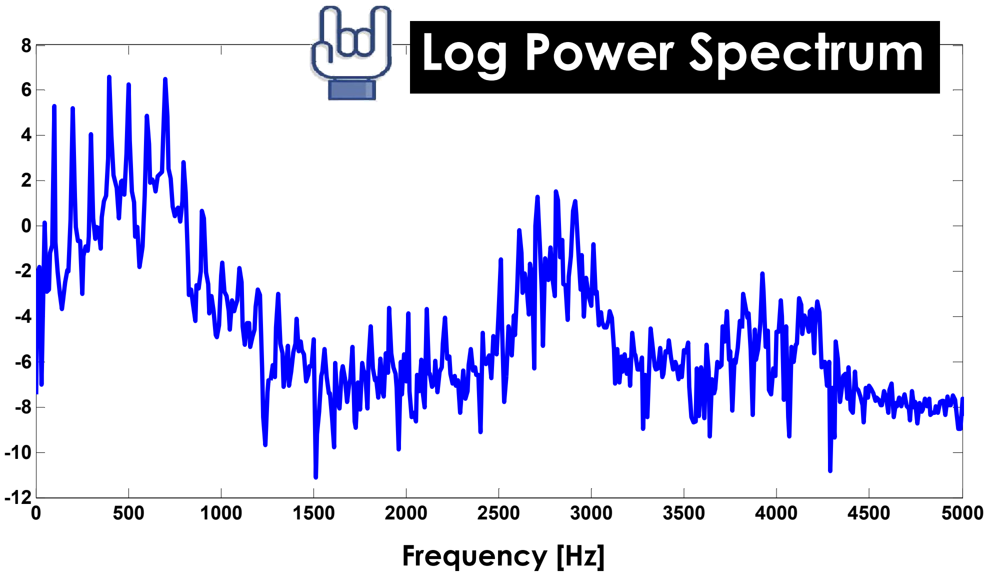
Cepstrum



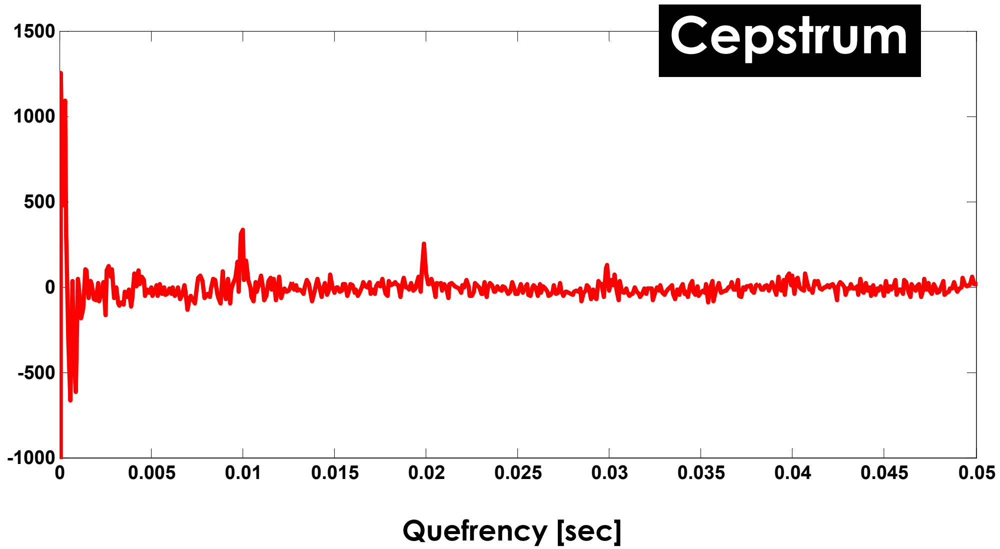
Cepstrum



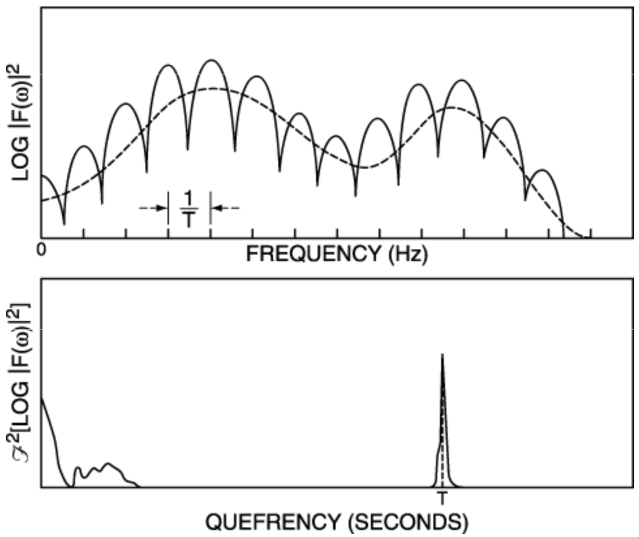
Cepstrum



Cepstrum



Cepstrum



http://postgavindisorder.blogspot.com/2010_09_01_archive.html

Cepstrum

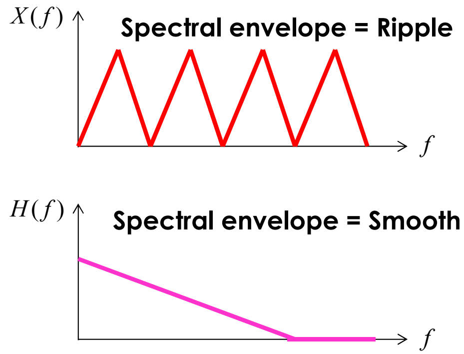
$x(t)$ ⇨ **System** ⇨ $s(t)$

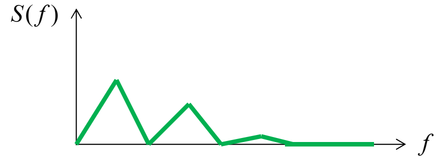
Impulse Response $h(t)$

$$s(t)=h(t)*x(t)$$

Fourier Transform

$$S(f)=H(f)X(f)$$





Cepstrum

Spectrum $$S(f)=H(f)X(f)$$

Power spectrum $$|S(f)|^2=|H(f)|^2|X(f)|^2$$

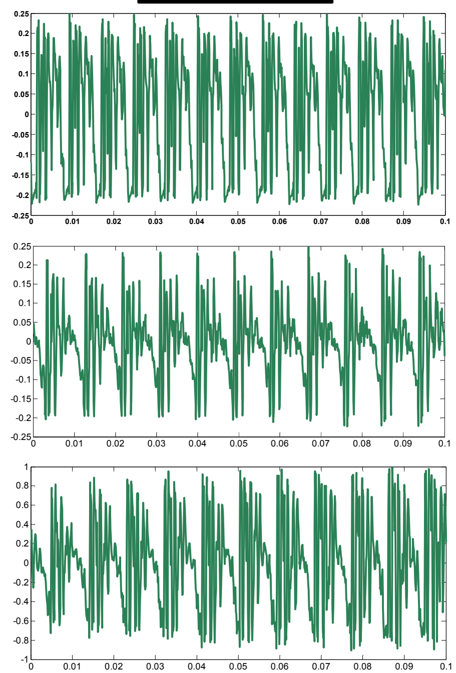
Log Power spectrum $$\log|S(f)|=\underbrace{\log|H(f)|}_{\text{Smooth SE}}+\underbrace{\log|X(f)|}_{\text{Ripple SE}}$$

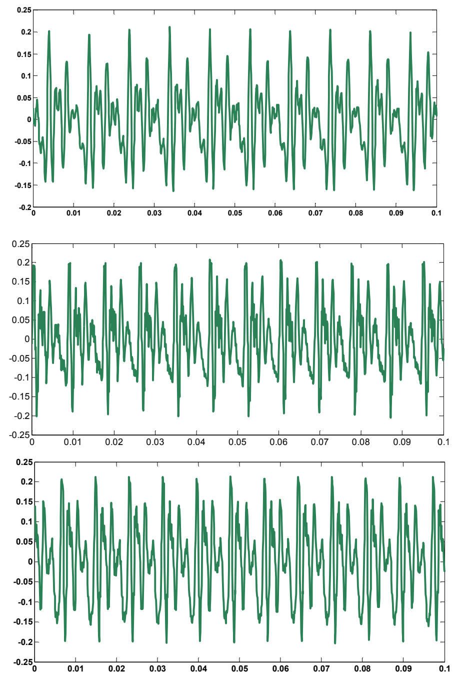
Fourier Transform

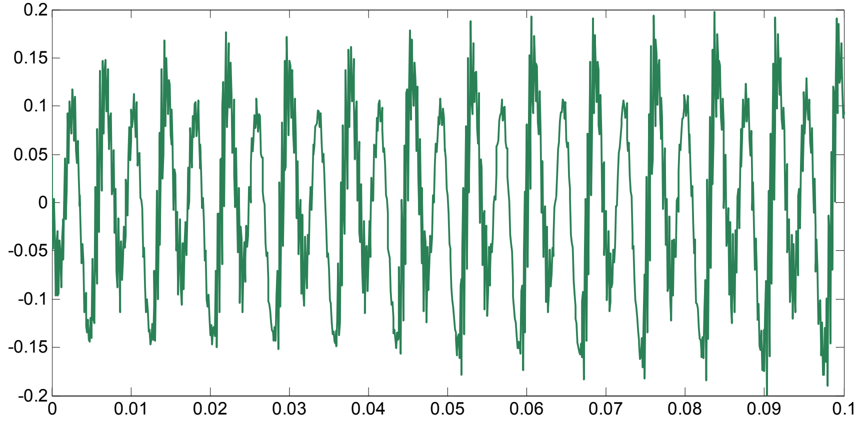
Cepstrum $$F\{\log|S(f)|\}=\underbrace{F\{\log|H(f)|\}}_{\text{Low Quefrency Component}}+\underbrace{F\{\log|X(f)|\}}_{\text{High Quefrency Component}}$$

อาาาาหห์

โอวววว์

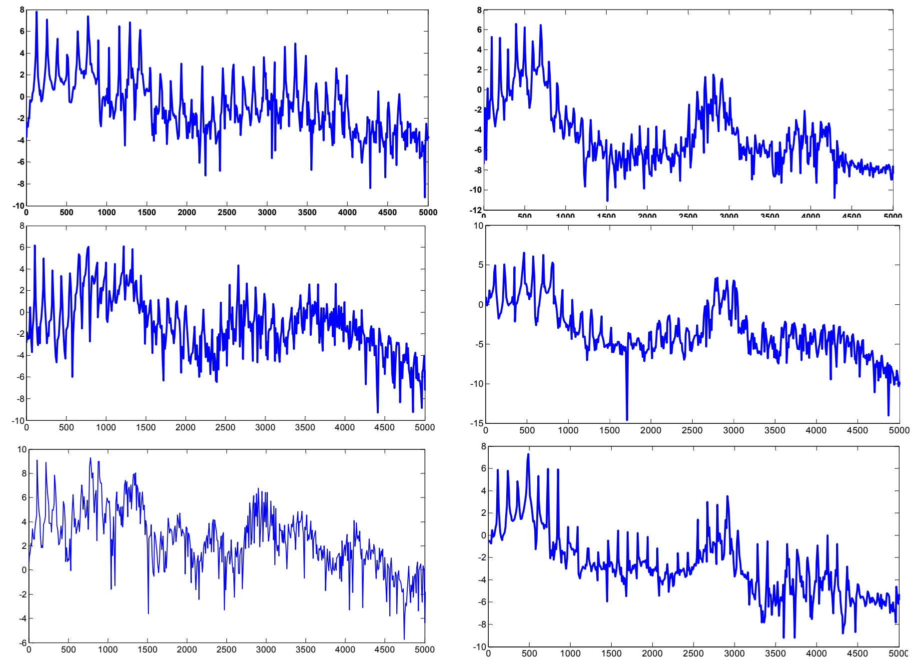


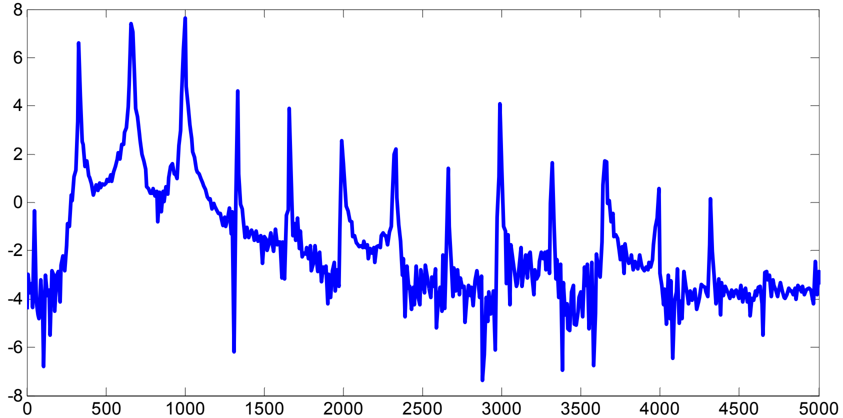




ให้ทายว่าเสียงอะไร?

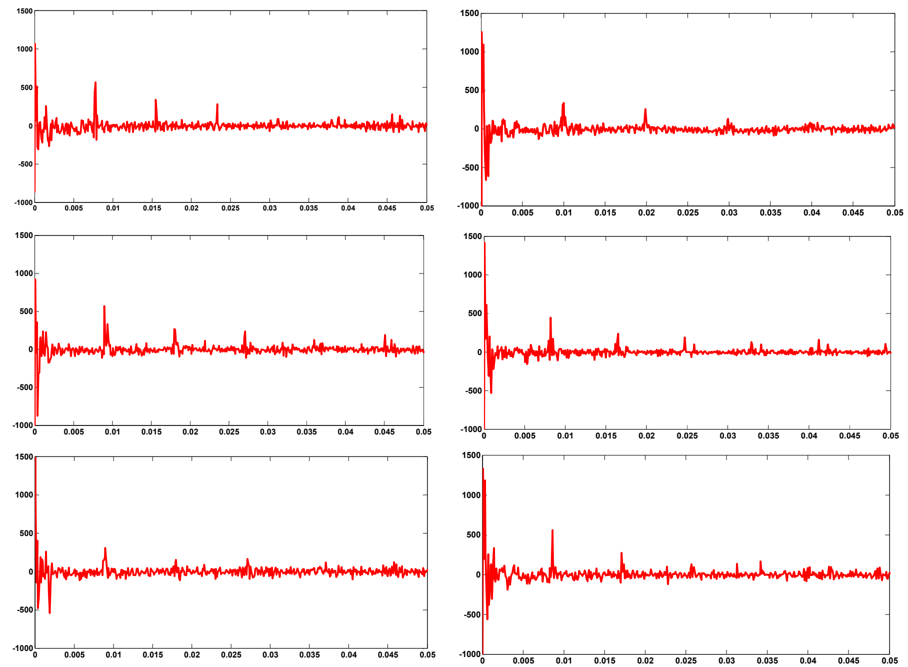
อาาาาาหห์ **โอวววว์**

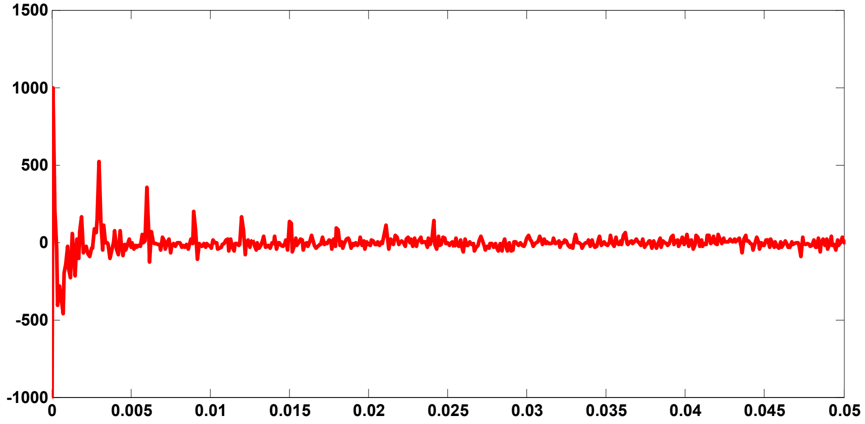




ให้ทายว่าเสียงอะไร?

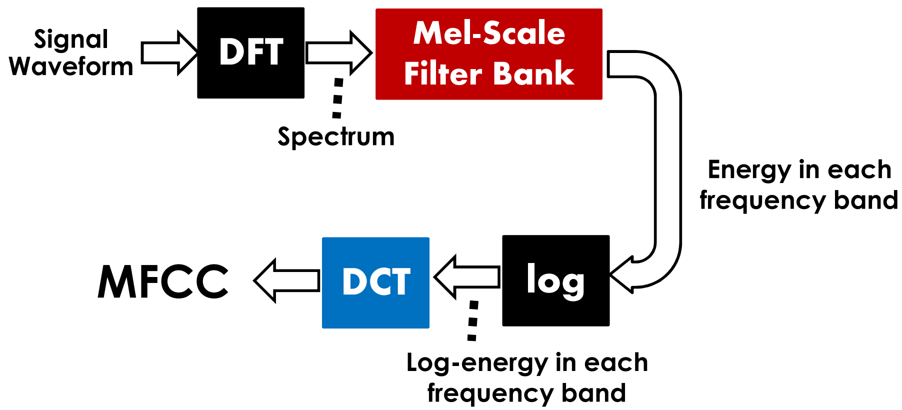
อาาาาาหห์ **โอวววว์**





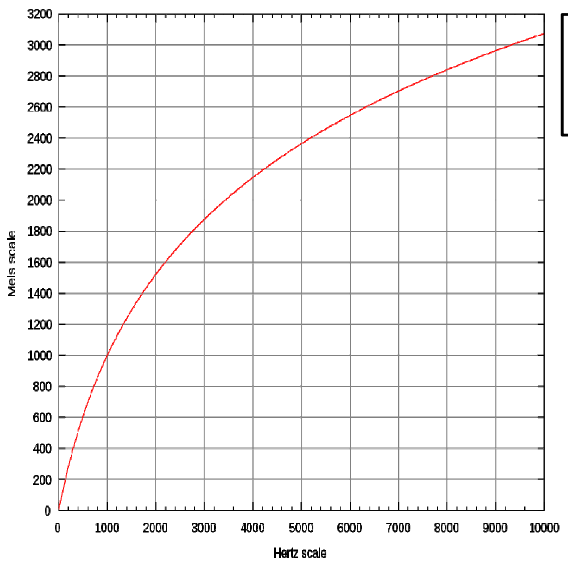
ให้ทายว่าเสียงอะไร?

Mel-Frequency Cepstrum Coefficient [MFCC]



DCT = Discrete Cosine Transform

Mel-Scale



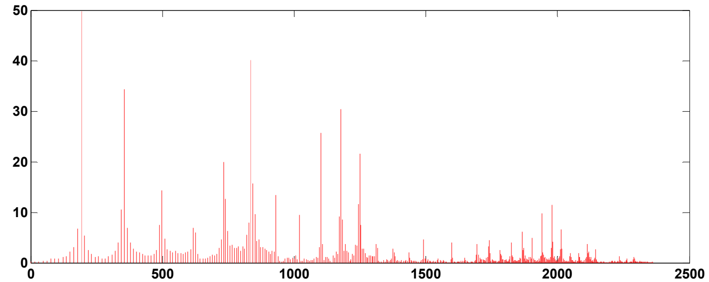
$$f_{[mel]} = 1127.0148 \ln\left(1 + \frac{f_{[Hz]}}{700}\right)$$

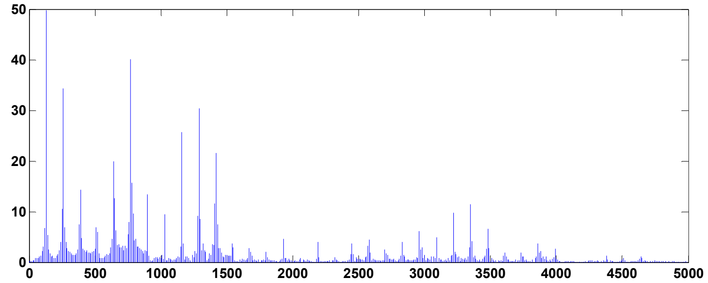
- Perceptual scale of pitches

Hertz	Mel
174	250
391	500
662	750
1000	1000
1949	1500
3429	2000
5734	2500

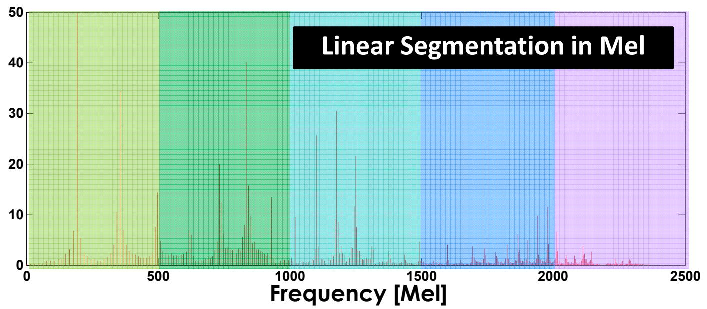
http://en.wikipedia.org/wiki/File:Mel-Hz_plot.svg

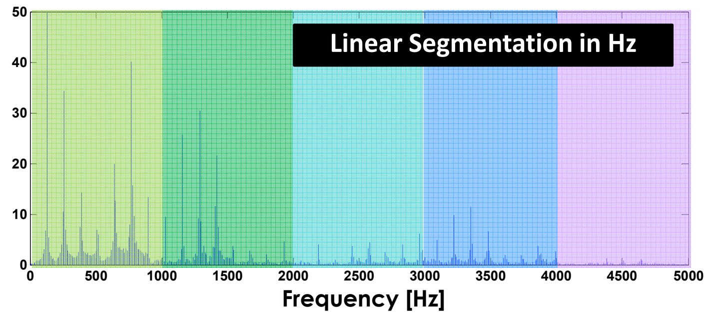
Mel-Scale



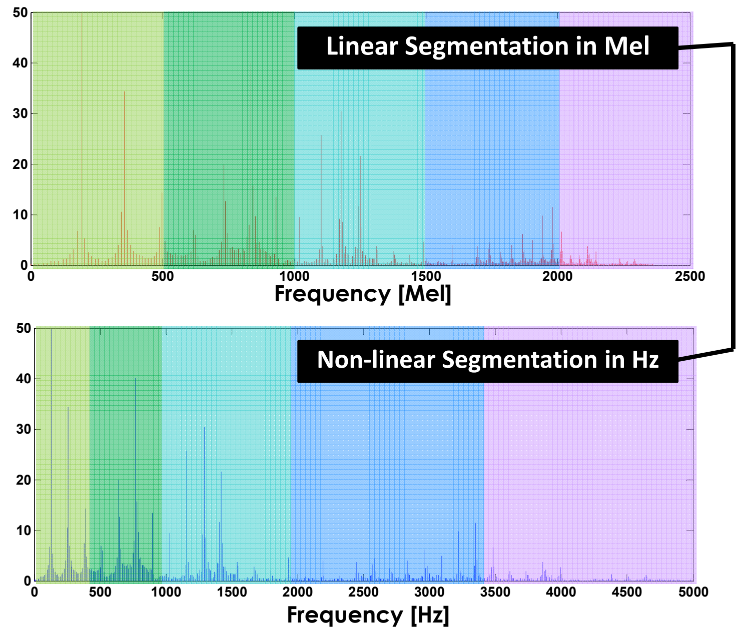


Mel-Scale





Mel-Scale

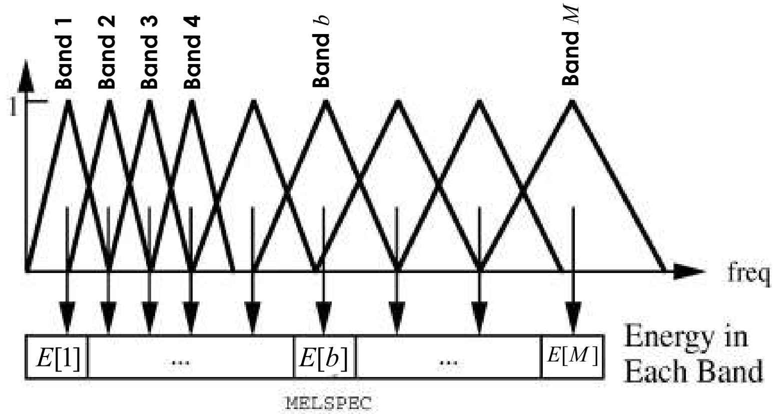


Mel-Scale Filter Bank

$$E[b] = \sum_{f \in Band[b]} |S(f)| H_b(f)$$

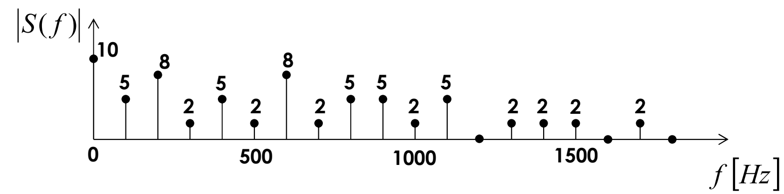
$E[b]$: Energy in Band b

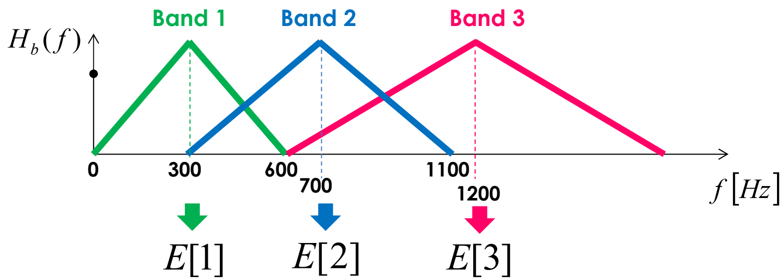
$H_b(f)$: Freq. Response of Filter for Band b



Unsupervised speaker segmentation with residual phase and MFCC features
S. Jothilakshmi, , V. Ramalingam , S. Palanivel

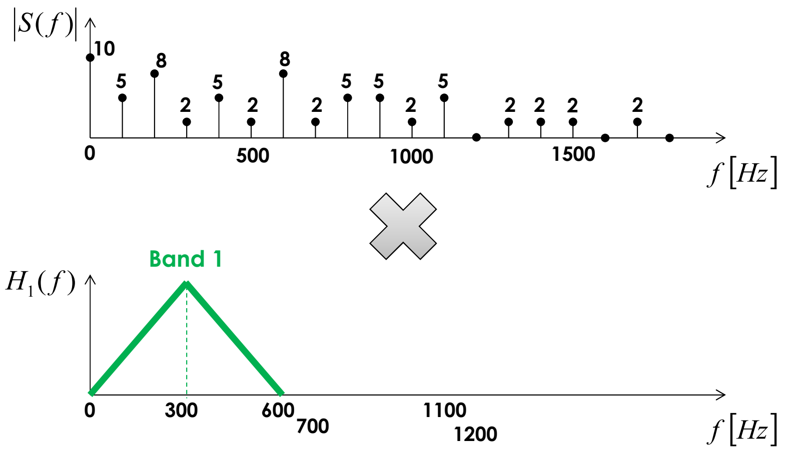
Mel-Scale Filter Bank





Mel-Scale Filter Bank

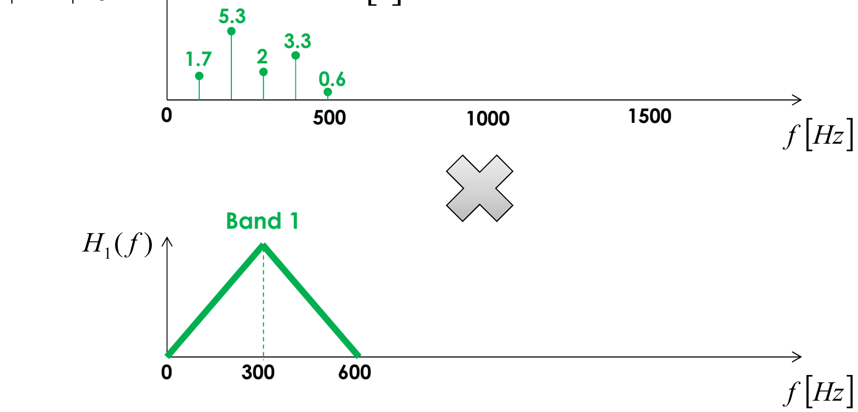
$$E[1] = \sum_{f \in Band[1]} |S(f)| H_1(f)$$



Mel-Scale Filter Bank

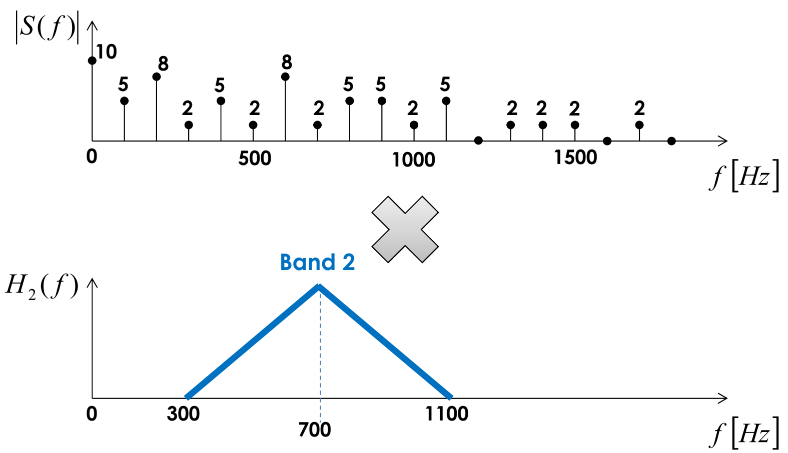
$$E[1] = \sum_{f \in Band[1]} |S(f)| H_1(f)$$

$$E[1] = 1.7 + 5.3 + 2 + 3.3 + 0.6 = 12.9$$



Mel-Scale Filter Bank

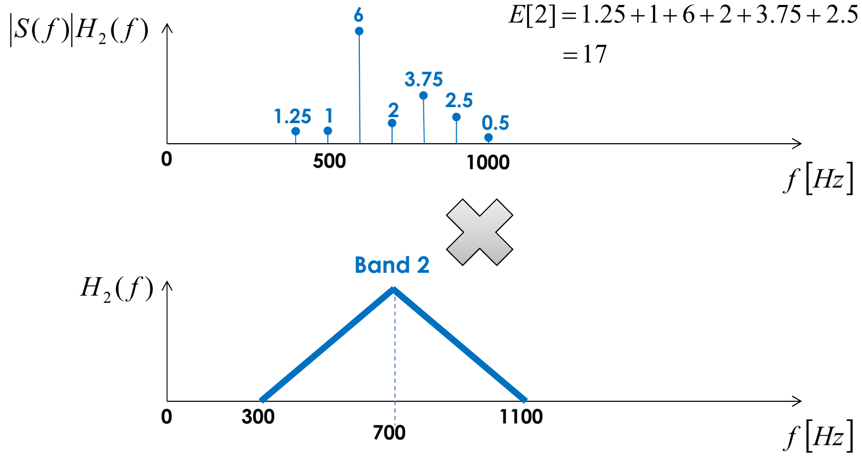
$$E[2] = \sum_{f \in Band[2]} |S(f)| H_2(f)$$



Mel-Scale Filter Bank

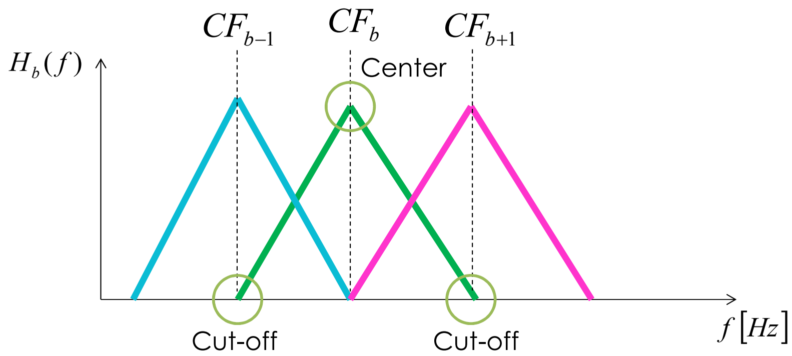
$$E[2] = \sum_{f \in Band[2]} |S(f)| H_2(f)$$

$$E[2] = 1.25 + 1 + 6 + 2 + 3.75 + 2.5 + 0.5$$
$$= 17$$



Mel-Scale Filter Bank

- Cut-off frequency of filter in the current band determined by the center frequencies of the two adjacent filters.



Mel-Scale Filter Bank

- Triangular Filter

$$H_b(f) = \begin{cases} \dfrac{f - CF_{b-1}}{CF_b - CF_{b-1}}; & CF_{b-1} < f < CF_b \\ \dfrac{f - CF_{b+1}}{CF_b - CF_{b+1}}; & CF_b \le f < CF_{b+1} \\ 0; & ;otherwise \end{cases}$$

CF_b = Center Frequency of Band $b\,[Hz]$

Mel-Scale Filter Bank

- Center Frequency

$$CF_b = F_{\min} + b \cdot \frac{F_{\max} - F_{\min}}{M+1} \quad in\ [mel]$$

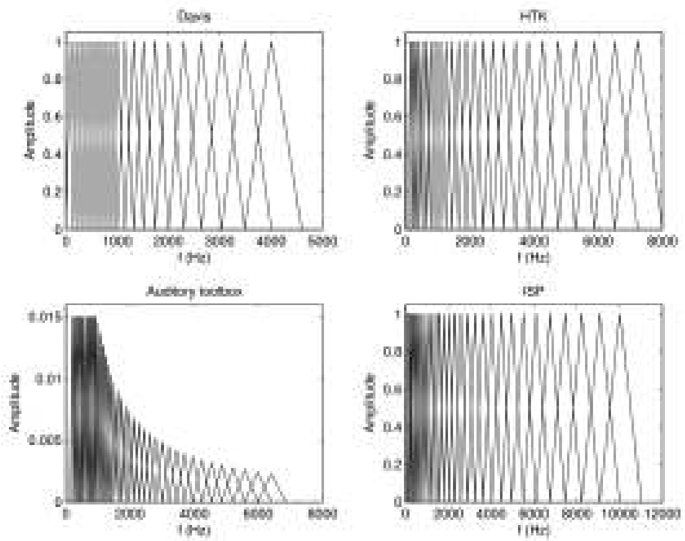
$$\Downarrow$$

$$CF_b \quad in\ [Hz]$$

M = Number of Bands

$F_{\min} \sim F_{\max}$ = Frequency Range

Mel-Scale Filter Bank



Discrete Cosine Transform [DCT]

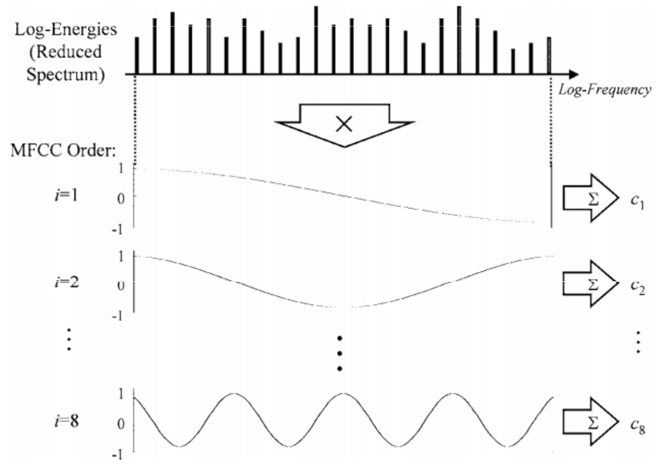
$$\boxed{c[i] = \sum_{b=1}^{M} \ln(E[b]) \cos\left(i\left(b - \frac{1}{2}\right)\frac{\pi}{M}\right)}$$

$c[i]$ = Mel Frequency Cepstrum Coefficient

i = Order of MFCC

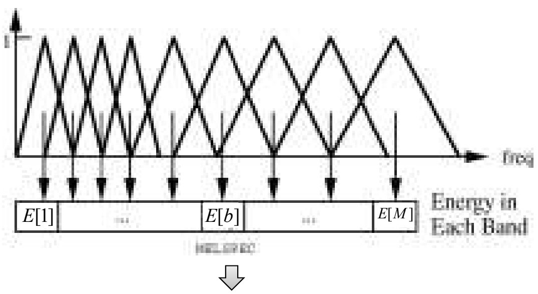
M = Number of Frequency Bands

Discrete Cosine Transform [DCT]



MPEG-7 Audio and. Beyond. Audio Content Indexing and. Retrieval. Hyoung-Gook Kim, et. al.

Discrete Cosine Transform [DCT]



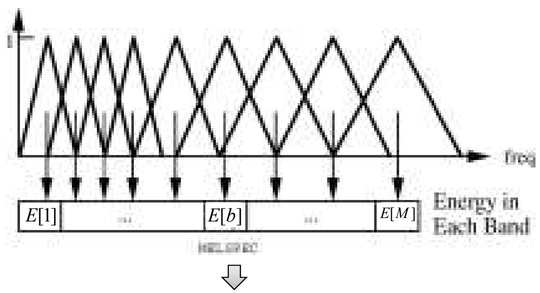
100	50	0.5	20	0.1	2	5	1

$$c[1] = \sum_{b=1}^{8} \ln(E[b]) \cos\left(\left(b - \frac{1}{2}\right)\frac{\pi}{8}\right)$$

$$= \ln(100)\cos\left(\frac{\pi}{16}\right) + \ln(50)\cos\left(\frac{3\pi}{16}\right) + \ln(0.5)\cos\left(\frac{5\pi}{16}\right) + \ldots + \ln(1)\cos\left(\frac{15\pi}{16}\right)$$

$$= 6.6947$$

Discrete Cosine Transform [DCT]

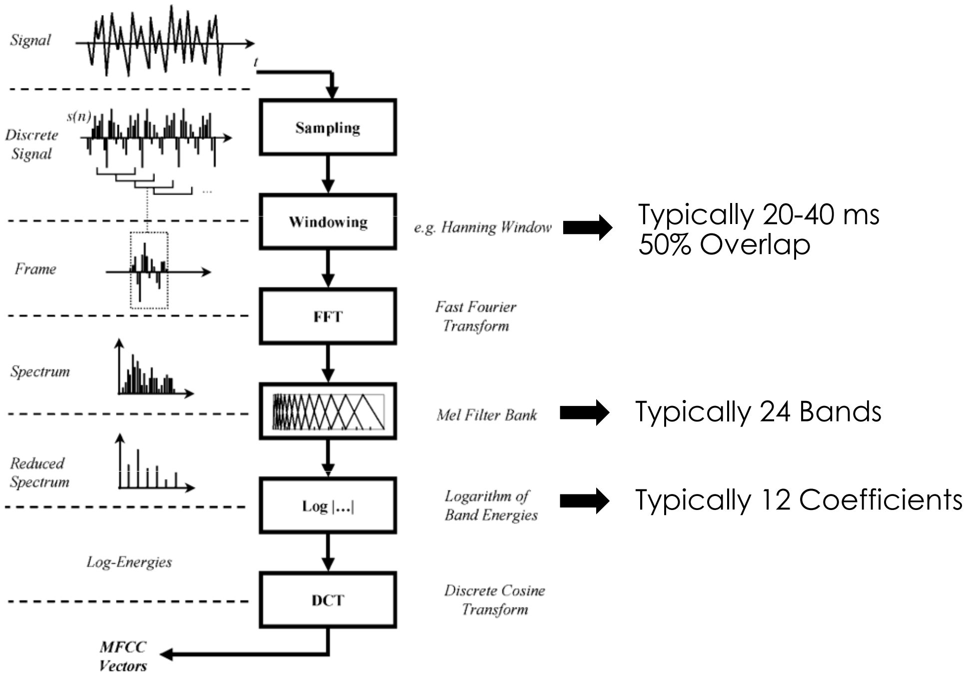


100	50	0.5	20	0.1	2	5	1

$$c[2] = \sum_{b=1}^{8} \ln(E[b]) \cos\left(2\left(b - \frac{1}{2}\right)\frac{\pi}{8}\right)$$

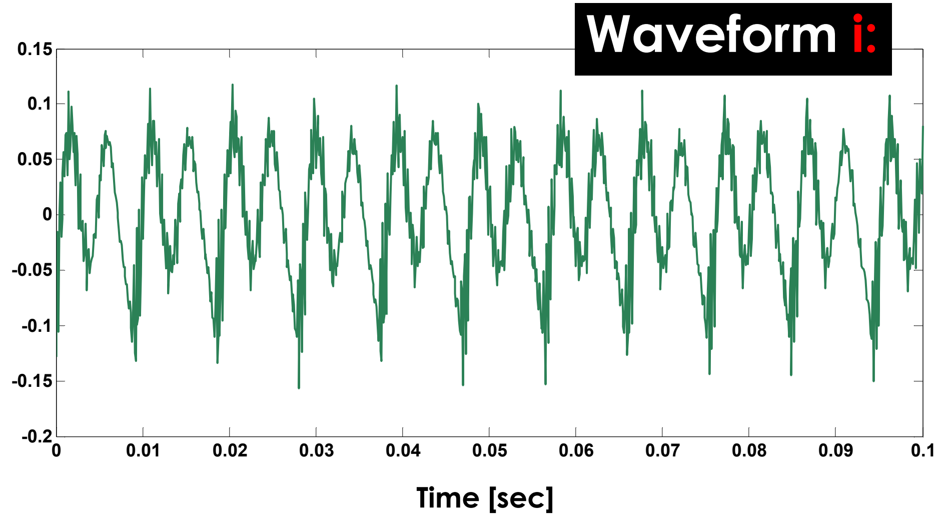
$$= \ln(100)\cos\left(\frac{\pi}{8}\right) + \ln(50)\cos\left(\frac{3\pi}{8}\right) + \ln(0.5)\cos\left(\frac{5\pi}{8}\right) + \ldots + \ln(1)\cos\left(\frac{15\pi}{8}\right)$$

$$= 5.7272$$

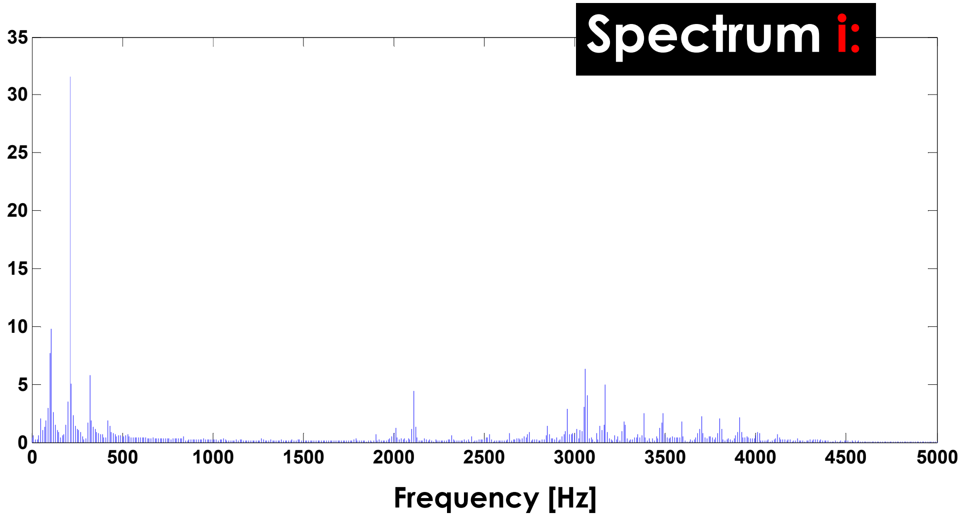


MPEG-7 Audio and. Beyond. Audio Content Indexing and. Retrieval. Hyoung-Gook Kim, et. al.

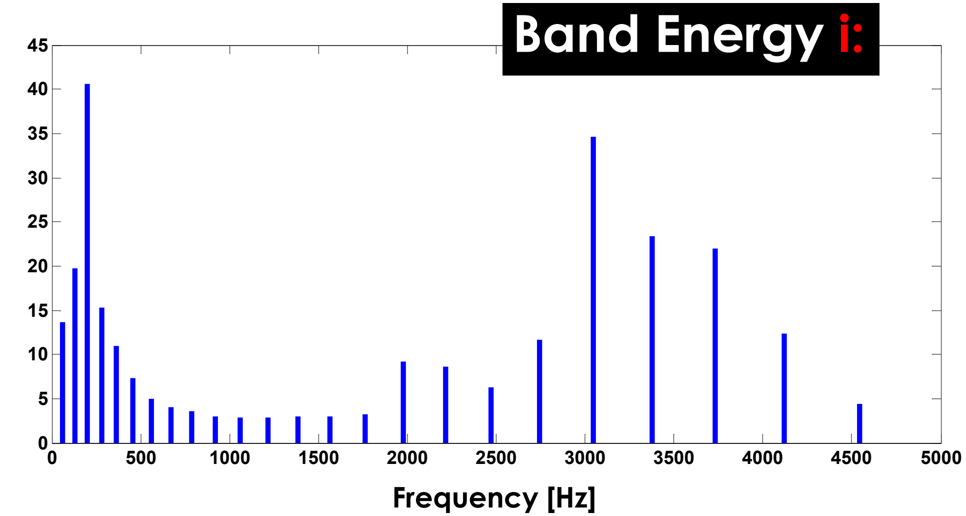
Mel-Frequency Cepstrum Coefficient



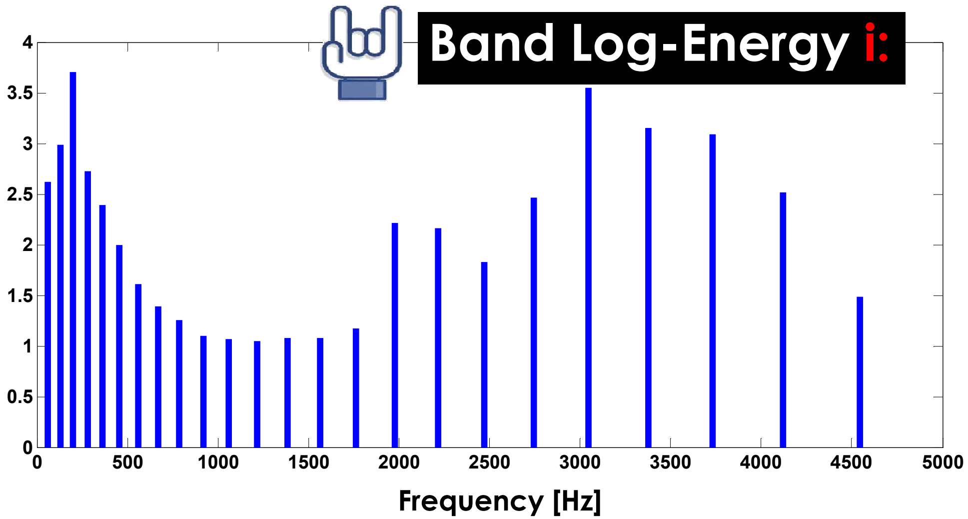
Mel-Frequency Cepstrum Coefficient



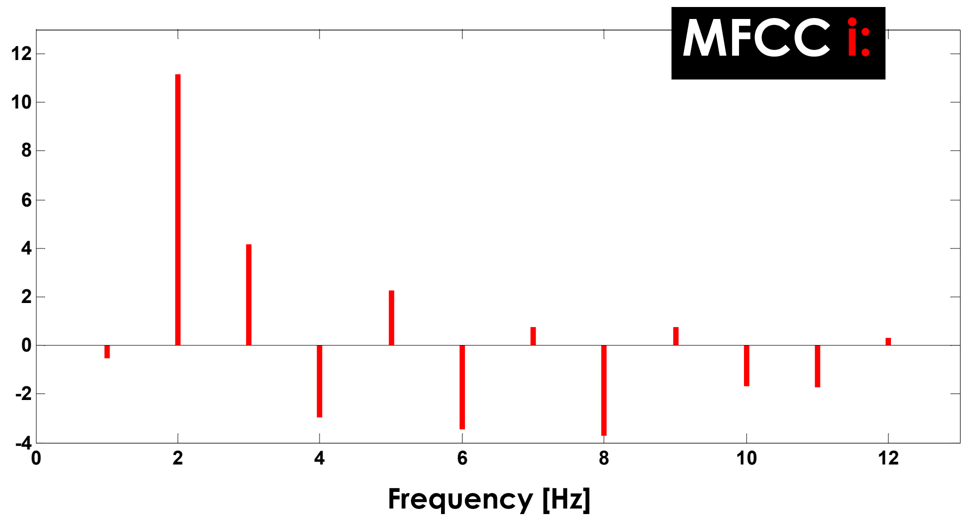
Mel-Frequency Cepstrum Coefficient

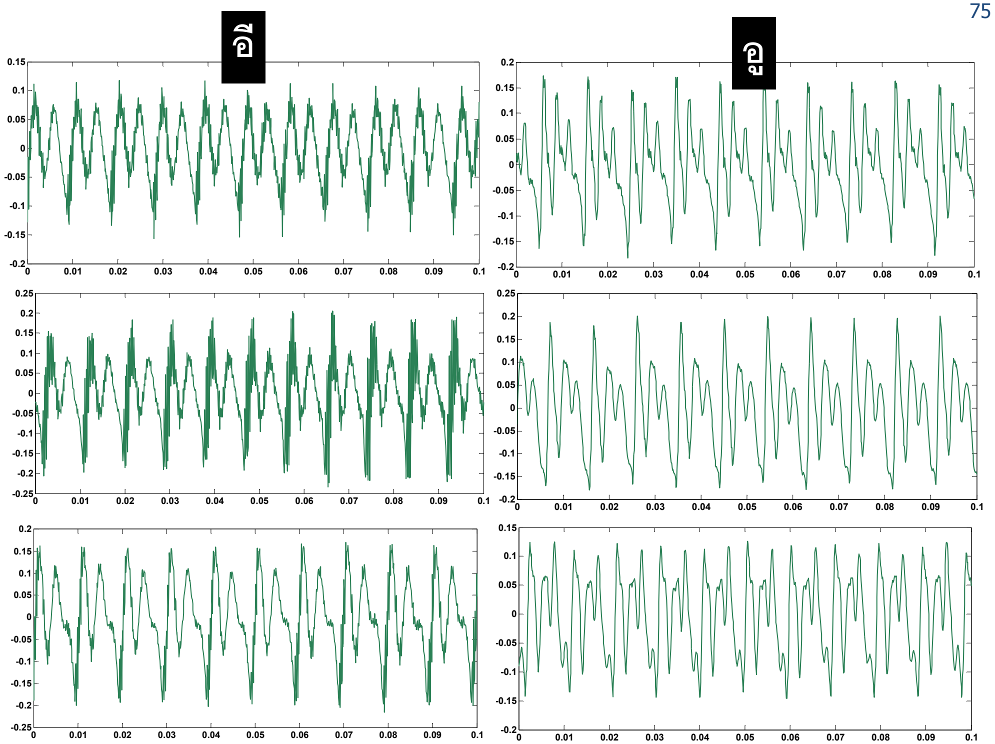


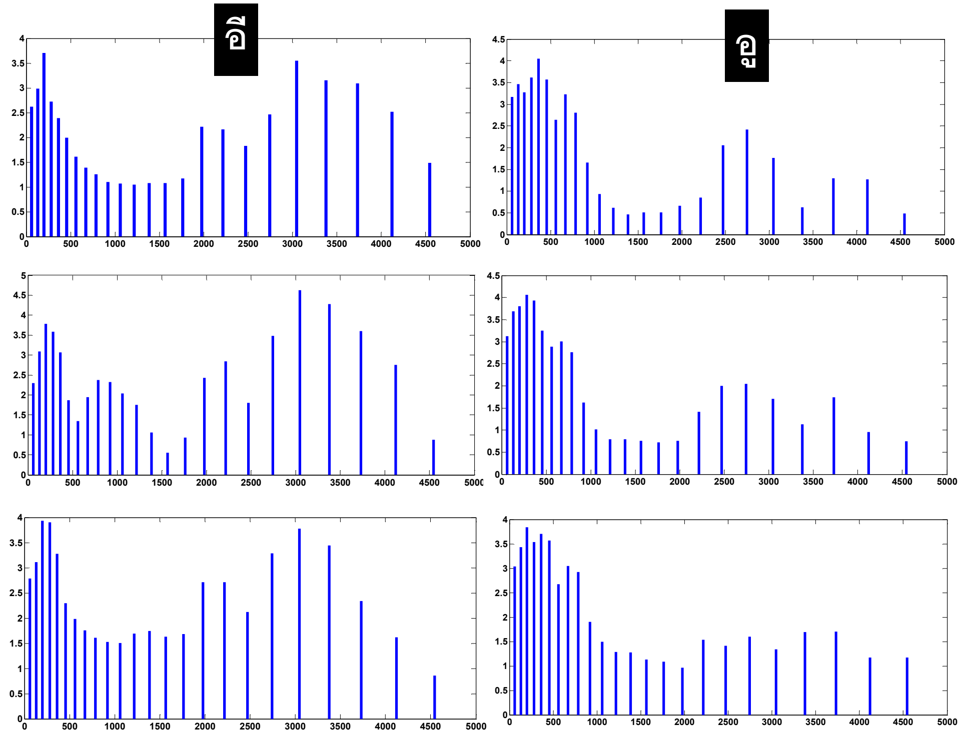
Mel-Frequency Cepstrum Coefficient

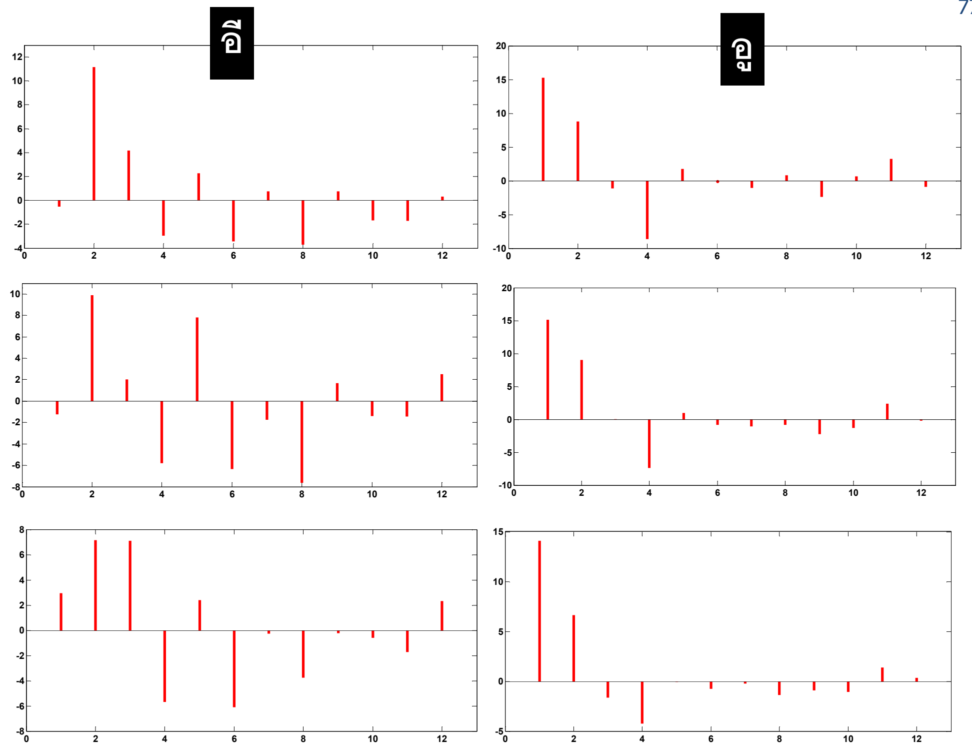


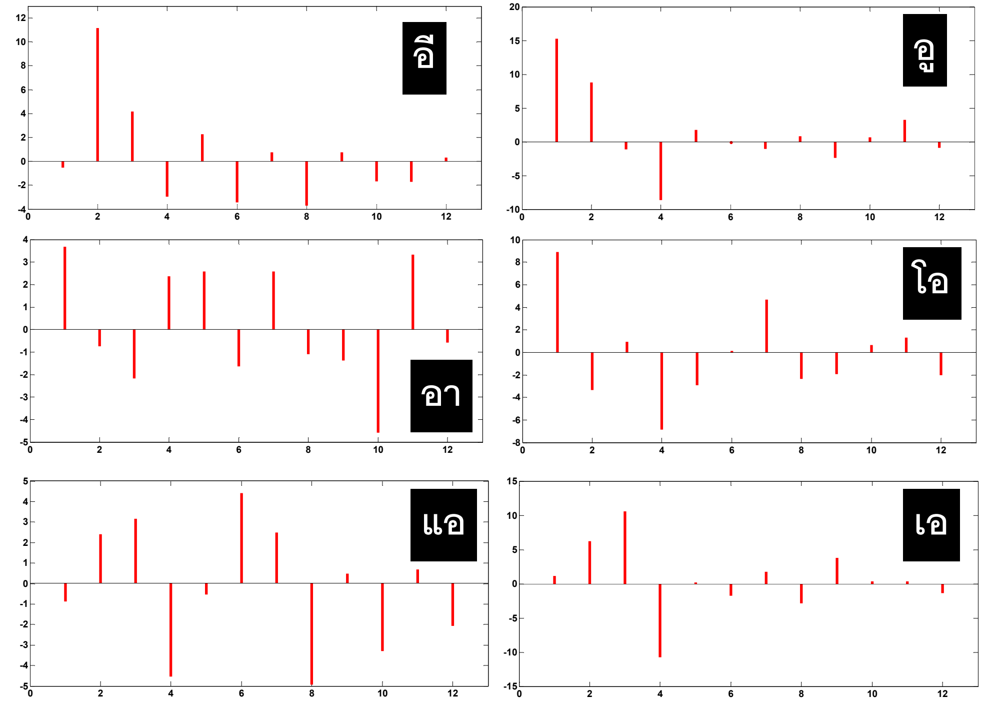
Mel-Frequency Cepstrum Coefficient

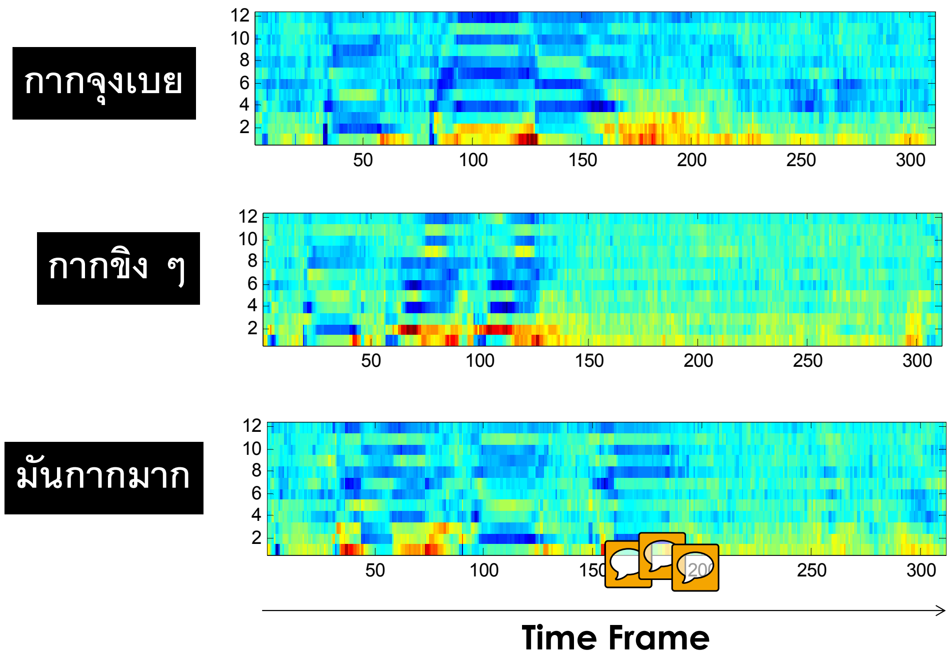


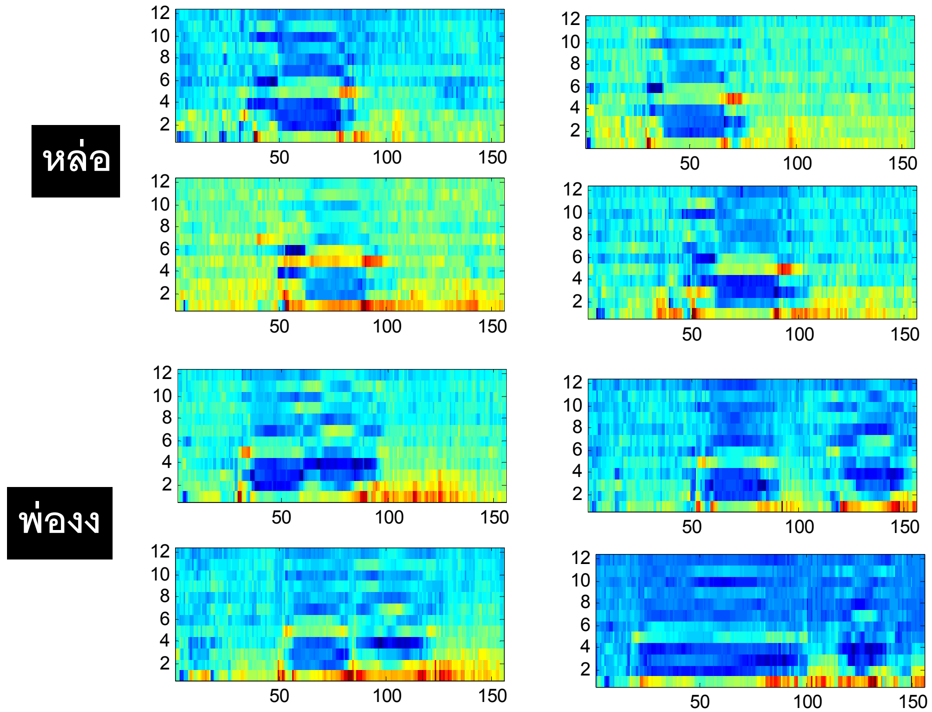





อี
อู


อี
อู
อา
โอ
แอ
เอ


กากจุงเบย
กากขิง ๆ
มันกากมาก
Time Frame


หล่อ
พ่องง