Automatic Speech Recognition handout (1)

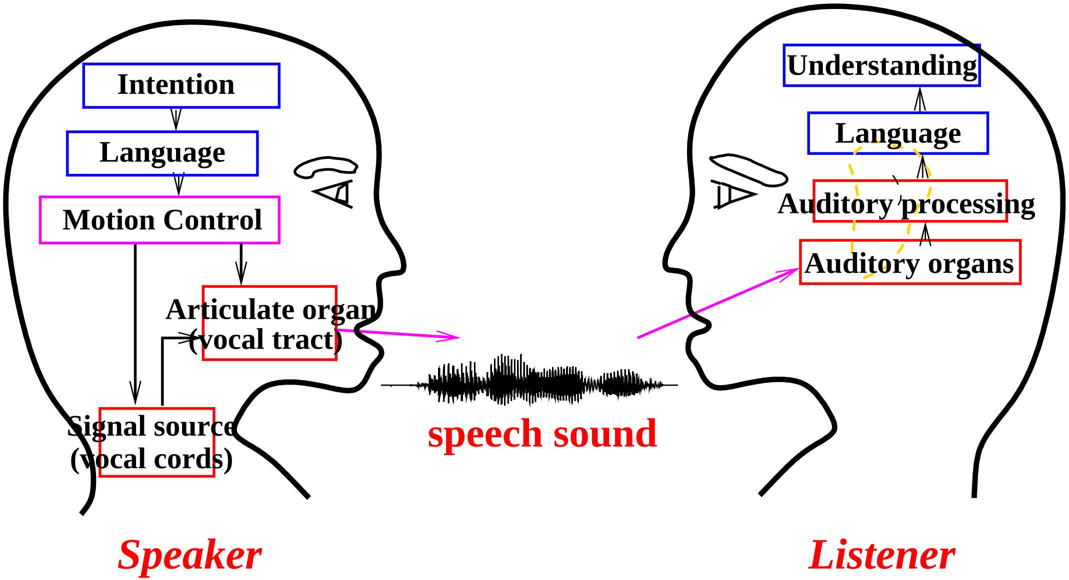
Jan - Mar 2012

$Revision$: 1.1

Speech Signal Processing and Feature Extraction

Hiroshi Shimodaira (h.shimodaira@ed.ac.uk)

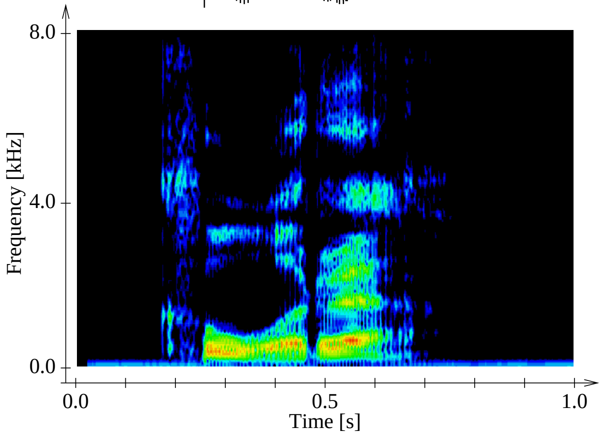
Speech Communication



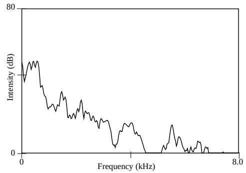
Spectrogram

Waveform

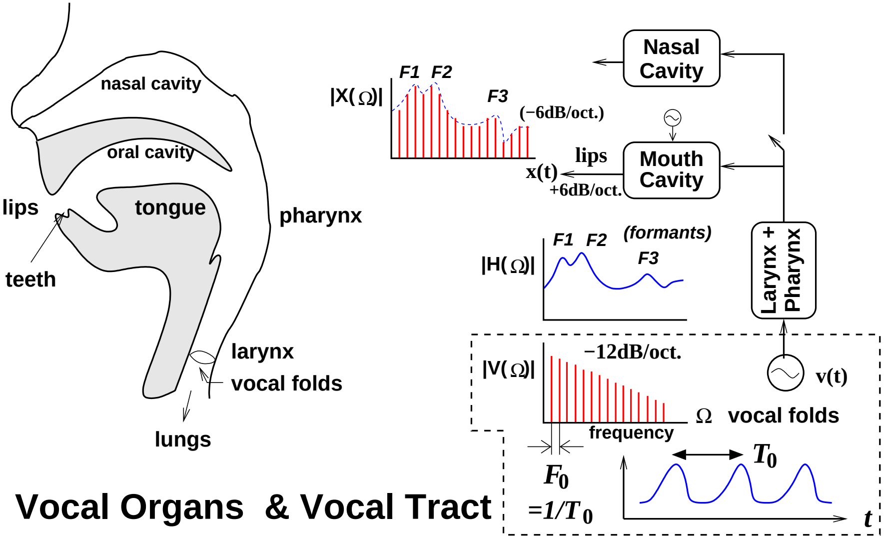
Spectrogram



Cross-section of spectrogram



Speech Production Model



Vocal Organs & Vocal Tract

Time domain: $x(t) = h(t) * v(t) = \int_0^{\infty} h(\tau)v(t-\tau)d\tau$

$\downarrow$ *Fourier transform*

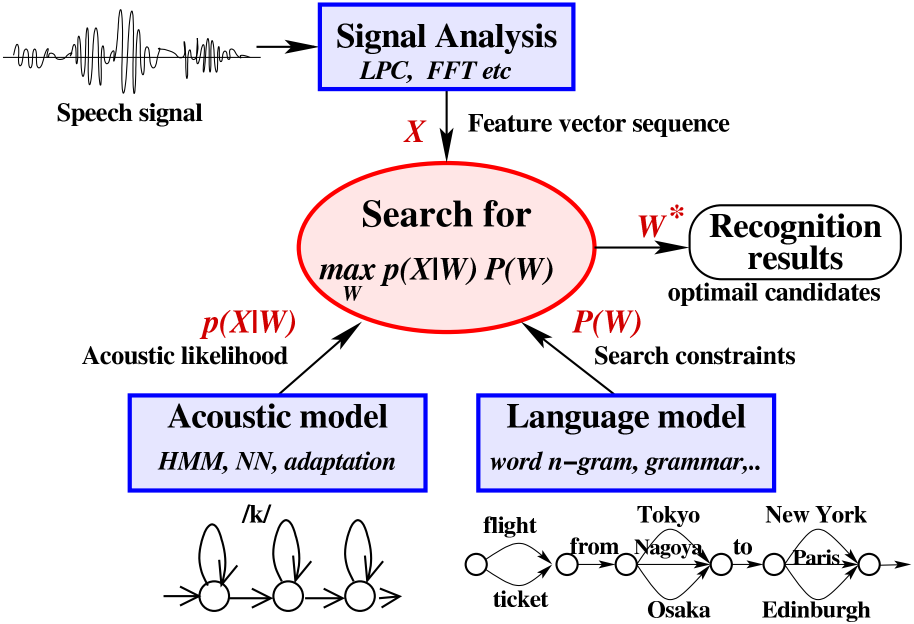
Frequency domain: $X(\Omega) = H(\Omega)V(\Omega)$

Ω : **angular frequency** ($= 2\pi F$)

F : **frequency**

Automatic Speech Recognition

Find the word sequence W **such that** $\max_{W} P(W|X) = \max_{W} \frac{P(X|W)P(W)}{P(X)}$



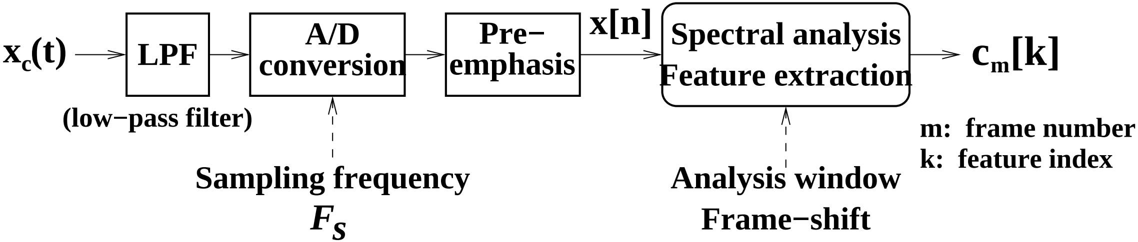
(after Sagayama, "Speech Translation Telephony",1994)

Signal Analysis for ASR

Front-end analysis

Convert acoustic signal into a sequence of feature vectors

e.g. MFCCs, PLP cepstral coefficients



Feature parameters for ASR

Features should

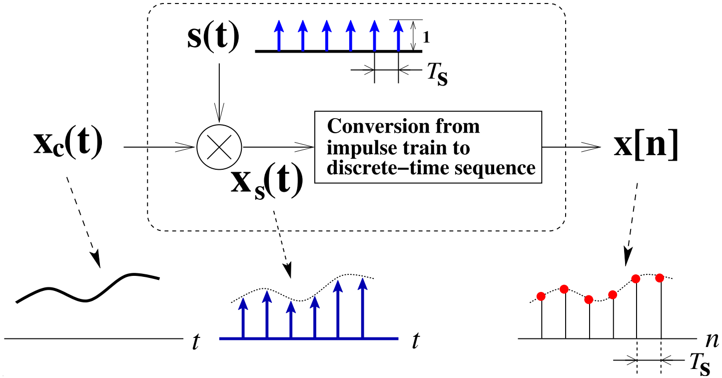
- **contain sufficient information to distinguish phonemes / phones**
  - **good time-resolutions [e.g. 10ms]**
  - **good frequency-resolutions [e.g. 20 channels/Bark-scale]**
- **not contain (or be separated from) F_0 and its harmonics**
- **be robust against speaker variation**
- **be robust against noise / channel distortions**
- **have good characteristics in terms of pattern recognition**
  - **The number of features is as few as possible**
  - **Features are independent of each other**

Converting analogue signals to machine readable form

- **Discretisation (sampling)** $x_c(t) \longrightarrow x[n]$
  - **continuous time** $\Rightarrow$ **discrete time**
  - **continuous amplitude** $\Rightarrow$**discrete amplitude**

Problem: information can be lost by sampling

Sampling of continuous-time signals



- **Continuous-time signal:** $x_c(t)$
- **Modulated signal by a periodic impulse train:**

$$x_s(t) = x_c(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s) = \sum_{n=-\infty}^{\infty} x_c(nT_s)\delta(t - nT_s)$$

- **Sampled signal:** $x[n] = x_s(nT_s)$ $\cdots$ **discrete-time signal**

T_s **:** **Sampling interval**

Sampling of continuous-time signals *(cont. 2)*

Q: Is the C/D conversion invertible ?

$$x_c(t) \xrightarrow{\text{C/D}} x[n] \xrightarrow{\text{D/C}} x_c(t)\textbf{?}$$

Sampling of continuous-time signals *(cont. 3)*

Q: Is the C/D conversion invertible ?

$$x_c(t) \xrightarrow{C/D} x[n] \xrightarrow{D/C} x_c(t)?$$

A: "No" in general, but
"Yes" under a special condition:
"Nyquist sampling theorem"

If $x_c(t)$ is band-limited (i.e. no frequency components $> F_s/2$), then $x_c(t)$ can be fully reconstructed by $x[n]$.

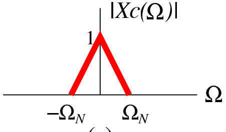
$$x_c(t) = h_{T_s}(t) * \sum_{k=-\infty}^{\infty} x[k]\delta(t - kT_s) = \sum_{k=-\infty}^{\infty} x[k]h_{T_s}(t - kT_s)$$

$$h_{T_s}(t) = \mathrm{sinc}(t/T_s) = \frac{\sin(\pi t/T_s)}{\pi t/T_s}$$

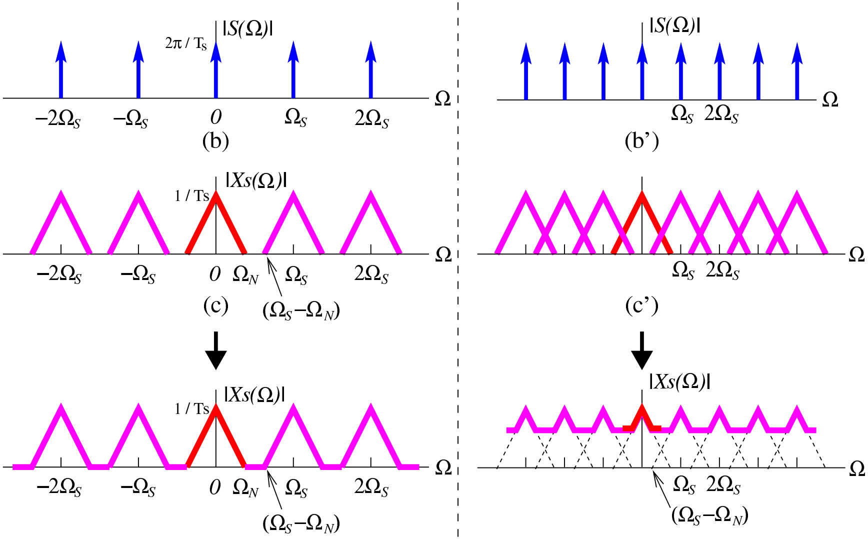
$F_s/2$: **Nyquist Frequency**, $F_s = 1/T_s$: **Sampling Frequency**

Sampling of continuous-time signals *(cont. 4)*

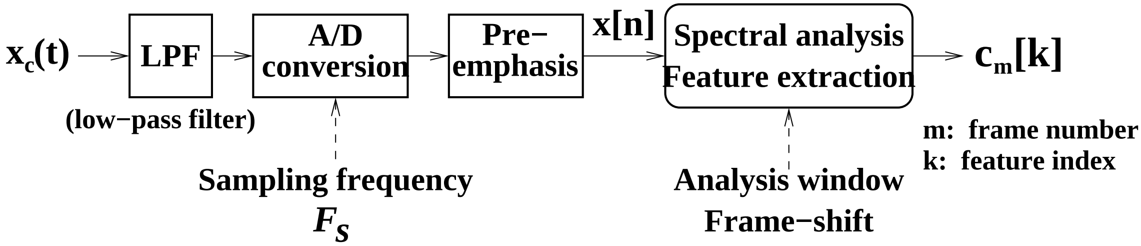
Interpretation in frequency domain:



$$\underbrace{X_s(\Omega)}_{Spectrum\ of\ x_s(t)} = \frac{1}{T_s}\sum_{k=-\infty}^{\infty} \underbrace{Xc(\Omega - k\Omega_s)}_{Spectrum\ of\ x_c(t)}$$



Sampling of continuous-time signals *(cont. 5)*



Questions

1. **What sampling frequencies (F_s) are used for ASR ?**
   - **microphone voice:** $12kHz \sim 20kHz$
   - **telephone voice:** $\sim 8kHz$
2. **What are the advantages / disadvantages of using higher F_s ?**
3. **Why is pre-emphasis (+6dB/oct.) employed?**

$$x[n] = x_0[n] - a x_0[n-1], \quad a = 0.95 \sim 0.97$$

Spectral analysis: Fourier Transform

- **FT for continuous-time signals (& continuous-frequency)**

$$X_c(\Omega) = \int_{-\infty}^{\infty} x_c(t) e^{-j\Omega t} dt$$ **(time domain → freq. domain)**

$$x_c(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_c(\Omega) e^{j\Omega t} d\Omega$$ **(freq. domain → time domain)**

- **FT for discrete-time signals (& continuous-frequency)**

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n}$$

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega$$

$|X(e^{j\omega})|^2 \quad \cdots$ **Power spectrum**

$\log |X(e^{j\omega})|^2 \quad \cdots$ **Log power spectrum**

where $\omega = T_s \Omega = 2\pi f$,

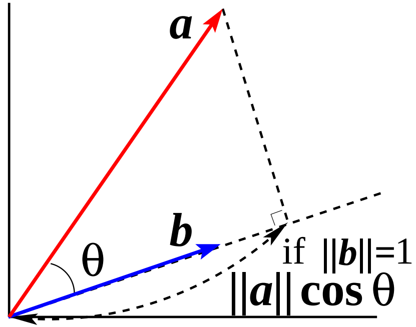
$e^{-j\omega n} = \cos(\omega n) + j \sin(\omega n)$, $\quad j :$ the imaginary unit

An interpretation of FT

Inner product between two vectors (Linear Algebra)

- **2-dimensional case**

$$\boldsymbol{a} = (a_1, a_2)^t$$
$$\boldsymbol{b} = (b_1, b_2)^t$$
$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^t \boldsymbol{b} = a_1 b_1 + a_2 b_2$$
$$= \| \boldsymbol{a} \| \, \| \boldsymbol{b} \| \cos\theta$$



- **Infinite-dimensional case**

$$\boldsymbol{x} \triangleq \{x[n]\}_{-\infty}^{\infty}$$
$$\boldsymbol{e}_\omega \triangleq \left\{\boldsymbol{e}^{j\omega n}\right\}_{-\infty}^{\infty} = \{\cos(\omega n) + j\sin(\omega n)\}_{-\infty}^{\infty}$$
$$\triangleq \mathbf{cos}_\omega + j\mathbf{sin}_\omega$$
$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n} = \boldsymbol{x} \cdot \boldsymbol{e}^{j\omega n} = \boldsymbol{x} \cdot \mathbf{cos}_\omega + j\boldsymbol{x} \cdot \mathbf{sin}_\omega$$

$\boldsymbol{x} \cdot \mathbf{cos}_\omega$ **: proportion of how much $\cos_\omega$ component is contained in x**

Short-time Spectrum Analysis

Problem with FT

- **Assuming signals are stationary:**
  signal properties do not change over time
- **If signals are non-stationary**
  $\Rightarrow$ loses information on time varying features

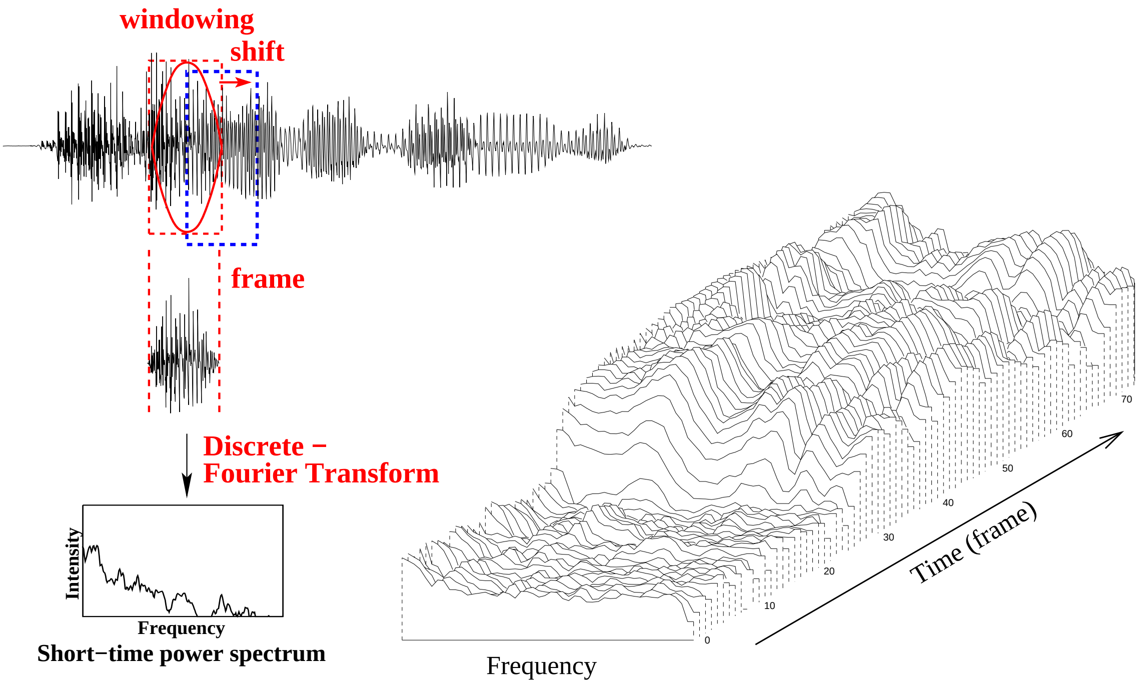
$\Rightarrow$ **Short-time Fourier transform (STFT)**
(Time-dependent Fourier transform)

$$\Downarrow$$

Divide the signal $x[n]$ into short-time segments (frames) $x_k[m]$ and apply FT to each segment.

$$
\begin{array}{c|ccccc}
\hline
x[n] & x_1[m], & x_2[m], & \ldots, & x_k[m], & \ldots \\
\downarrow & \downarrow & \downarrow & & \downarrow & \\
X(\omega) & X_1(\omega), & X_2(\omega), & \ldots, & X_k(\omega), & \ldots \\
\hline
\end{array}
$$

Short-time Spectrum Analysis *(cont. 2)*



Short-time Spectrum Analysis *(cont. 3)*

- **Trade-off problem of short time spectrum analysis**

	window width
	short $\longrightarrow$ long
frequency resolution	$\nearrow$
time resolution	$\searrow$

$\Rightarrow$ **a compromise for ASR:**

window width (frame width): $20 \sim 30$ **ms**
window shift (frame shift): $5 \sim 15$ **ms**

The Effect of Windowing in STFT

Time domain:

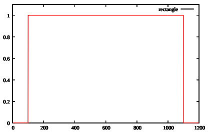
$$y_k[n] = w_k[n]x[n], \quad w_k[n] \ : \ \text{time-window for } k\text{-th frame}$$

Simply cutting out a short segment (frame) from $x[n]$ implies applying a rectangular window on to $x[n]$.

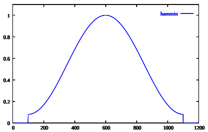
$\Rightarrow$ causes discontinuities at the edges of the segment.

Instead, a tapered window is usually used.. e.g. Hamming ($\alpha = 0.46164$) or Hanning ($\alpha = 0.5$) window)

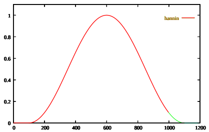
$$w[\ell] = (1-\alpha) - \alpha \cos\left(\frac{2\pi\ell}{N-1}\right) \qquad N : \text{window width}$$



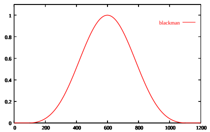
rectangle



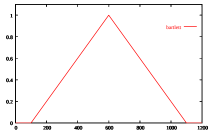
Hamming



Hanning



Blackman



Bartlett

The Effect of Windowing in STFT *(cont. 2)*

Frequency domain:

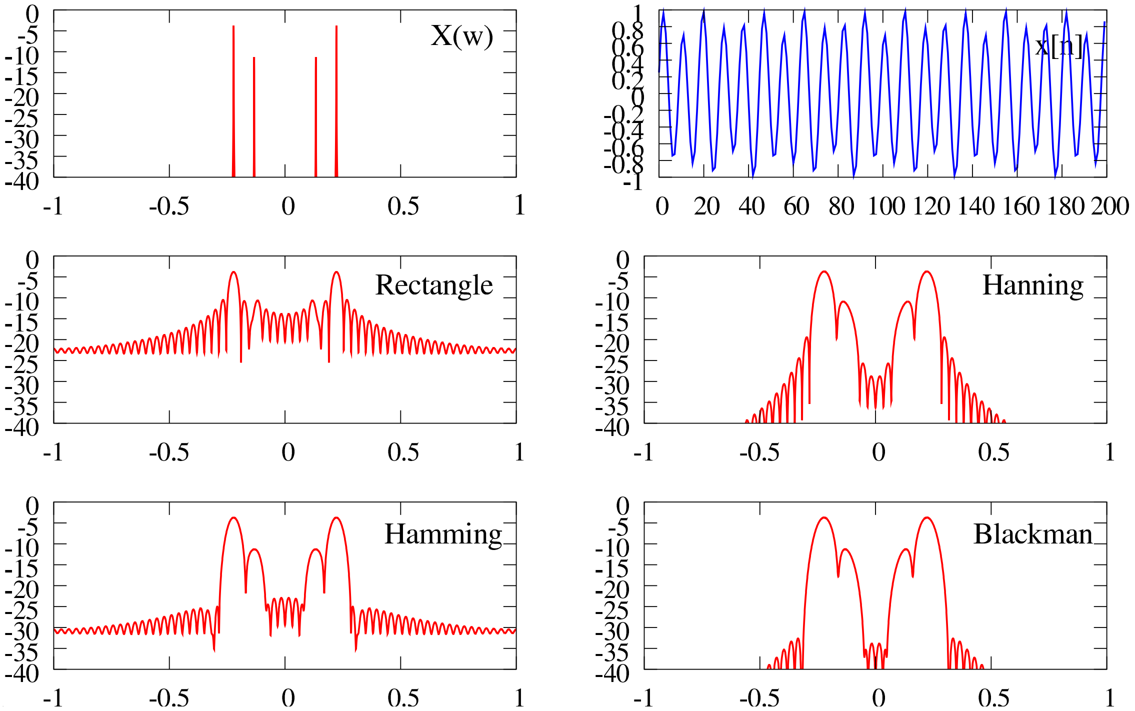
$$Y_k(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W_k(e^{j\theta}) X(e^{j(\omega-\theta)}) d\theta \quad \cdots \; \textbf{Periodic convolution}$$

- **Power spectrum of the frame is given as a periodic convolution between the power spectra of $x[n]$ and $w_k[n]$.**
- **If we want $Y_k(e^{j\omega}) = X(e^{j\omega})$, the necessary and sufficient condition for this is $W_k(e^{j\omega}) = \delta(\omega)$,**
  i.e. $w_k[n] = \mathcal{F}^{-1}\delta(\omega) = 1$, which means the length of $w_k[n]$ is infinite.
  $\Rightarrow$ there is no window function of finite length that causes no distortion.

NB: hereafter $x[n]$ will be also used to denote a segmented signal for simplicity.

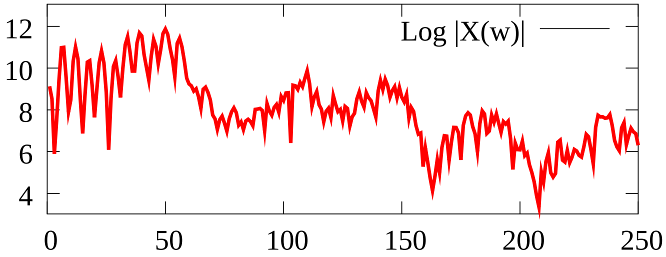
The Effect of Windowing in STFT *(cont. 3)*

Spectral analysis of two sine signals of close frequencies



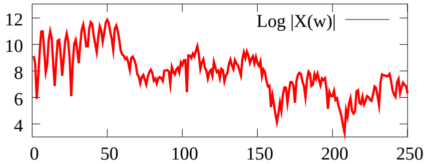
Problems with STFT

- **The estimated power spectrum contains harmonics of F_0, which makes it difficult to estimate the envelope of the spectrum.**
- **Frequency bins of STFT are highly correlated each other, i.e. power spectrum representation is highly redundant.**



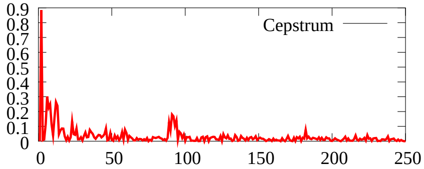
Cepstrum Analysis

Idea: split(deconvolve) the power spectrum into spectrum envelope and F_0 harmonics.



Log-spectrum [freq. domain]

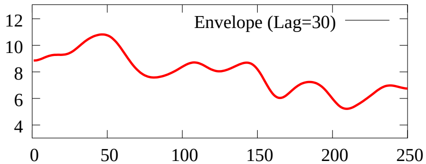
⇓ **Inverse Fourier Transform**



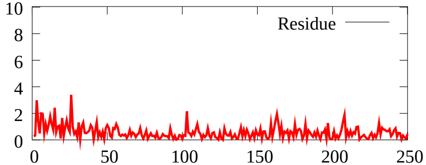
Cepstrum [time domain] (quefrency)

⇓ **Liftering to get low/high part**
(lifter: filter used in cepstral domain)

⇓ **Fourier Transform**



Smoothed-spectrum [freq. domain]
(low-part of cepstrum)



Log-spectrum of high-part of cepstrum

Cepstrum Analysis (cont. 2)

$$x[n] = h[n] * v[n]$$

$h[n]$: vocal tract
$v[n]$: glottal sounds

$$\downarrow \mathcal{F} \quad \text{(Fourier transform)}$$

$$X(e^{j\omega}) = H(e^{j\omega})V(e^{j\omega})$$

Log spectrum

$$\downarrow \log$$

$$\log|X(e^{j\omega})| = \underbrace{\log|H(e^{j\omega})|}_{\text{(spectral envelope)}} + \underbrace{\log|V(e^{j\omega})|}_{\text{(spectral fine structure)}}$$

Cepstrum

$$\downarrow \mathcal{F}^{-1}$$

$$c(\tau) = \mathcal{F}^{-1}\left\{\log|X(e^{j\omega})|\right\}$$
$$= \mathcal{F}^{-1}\left\{\log|H(e^{j\omega})|\right\} + \mathcal{F}^{-1}\left\{\log|V(e^{j\omega})|\right\}$$

LPC Analysis

Linear Predictive Coding (LPC):

a model-based / parametric spectrum estimation

Assume a "linear system" for human speech production

sound source $v[n]$ $\Rightarrow$ **vocal tract** $\Rightarrow$ **speech** $x[n]$

$$v[n] \longrightarrow \boxed{h[n]} \longrightarrow x[n] \qquad h[n] : \text{ impulse response}$$

$$x[n] = h[n] * v[n] = \sum_{k=0}^{\infty} h[k]\, v[n-k]$$

Using a model enables us to

- **estimate a spectrum of vocal tract from small amount of observations**
- **represent the spectrum with a small number of parameters**
- **synthesise speech with the parameters**

LPC Analysis (cont. 2)

Predict $x[n]$ **from** $x[n-1], x[n-2], \cdots$

$$\hat{x}[n] = \sum_{k=1}^{N} a_k x[n-k]$$

$$e[n] = x[n] - \hat{x}[n] = x[n] - \sum_{k=1}^{N} a_k x[n-k] \quad \cdots \textbf{ prediction error}$$

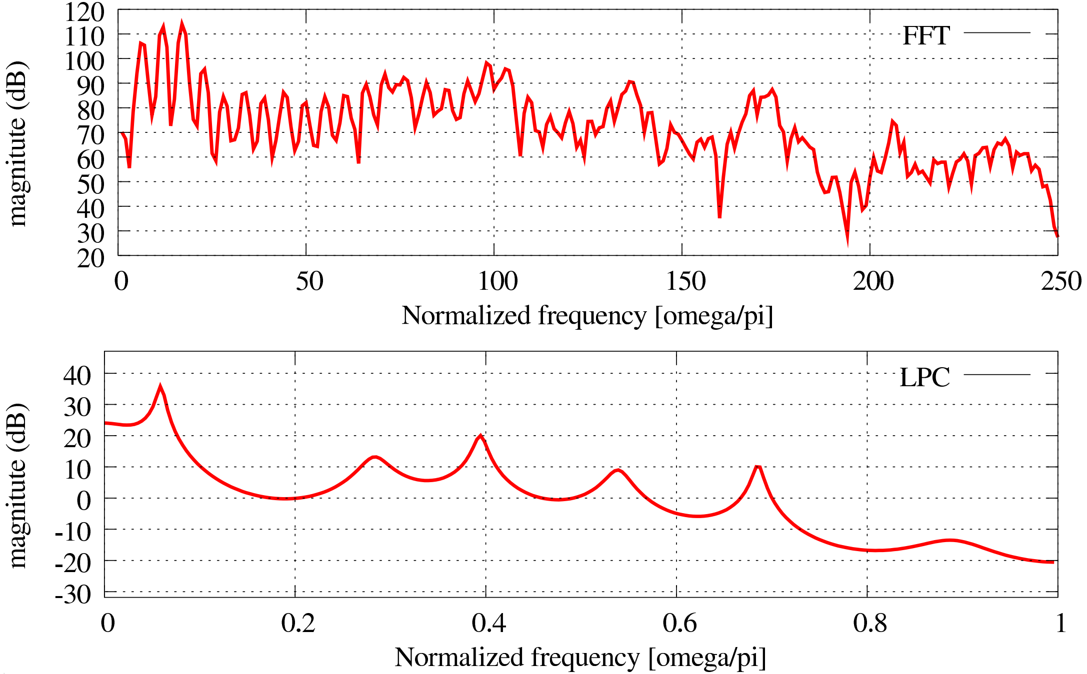
Optimisation problem

Find $\{a_k\}$ **that minimises the mean square (MS) error:**

$$P_e = E\left\{e^2[n]\right\} = E\left\{\left(x[n] - \sum_{k=1}^{N} a_k x[n-k]\right)^2\right\}$$

$\{a_k\}$ **: LPC coefficients**

Spectrums estimated by FT & LPC



LPC summary

- **Spectrum can be modelled/coded with around $14 LPCs$.**
- **LPC family**
  - **PARCOR (Partial Auto-Correlation Coefficient)**
  - **LSP (Line Spectral Pairs) / LSF (Line Spectrum Frequencies)**
  - **CSM (Composite Sinusoidal Model)**
- **LPC can be used to predict log-area ratio coefficients lossless tube model**
- **LPC-(Mel)Cepstrum: LPC based cepstrum.**
- **Drawback:**
  - **LPC assumes AR model which does not suit to model nasal sounds that have zeros in spectrum.**
  - **Difficult to determine the prediction order N.**

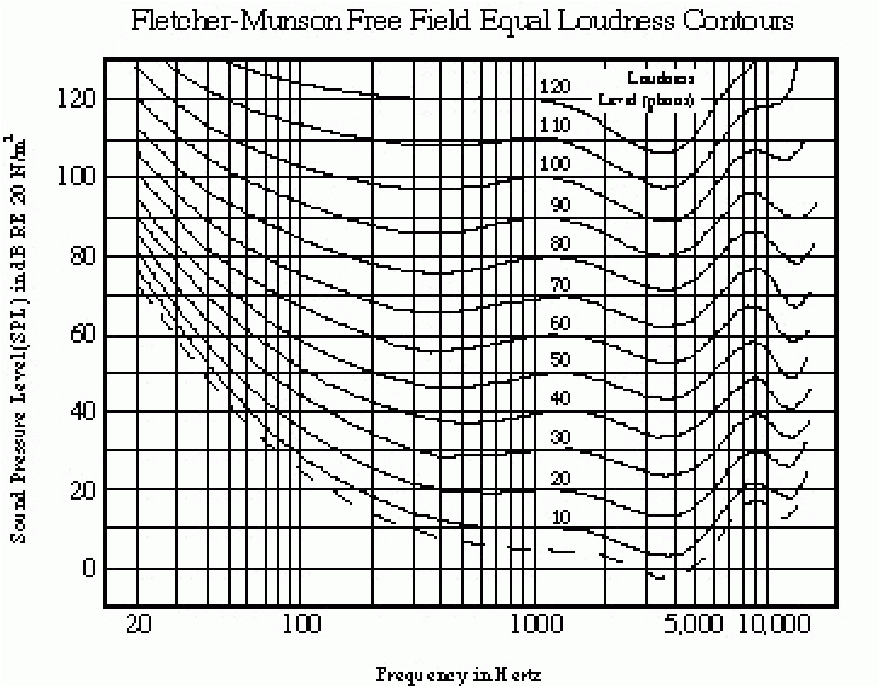
Taking into Perceptual Attributes

Physical quality	Perceptual quality
Intensity	Loudness
Fundamental frequency	Pitch
Spectral shape	Timbre
Onset/offset time	Timing
Phase difference in binaural hearing	Location

Technical terms

- **equal-loudness contours**
- **masking**
- **auditory filters (critical-band filters)**
- **critical bandwidth**

Taking into Perceptual Attributes *(cont. 2)*



Taking into Perceptual Attributes *(cont. 3)*

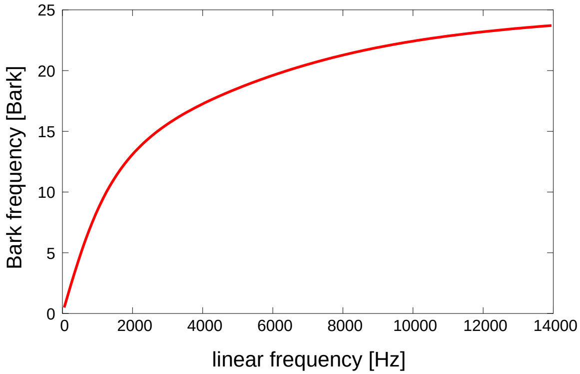
Non-linear frequency scale

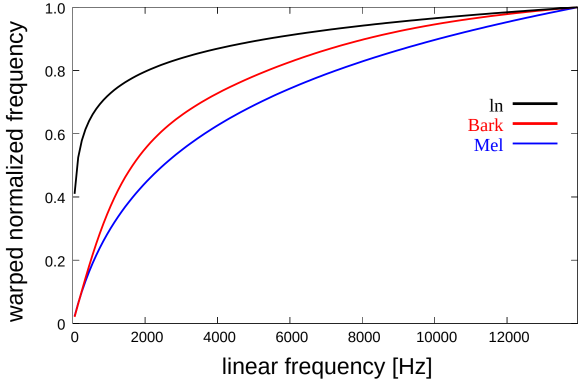
- **Bark scale**

$$b(f) = 13\arctan(0.00076f) + 3.5\arctan((f/7500)^2) \quad [\text{Bark}]$$

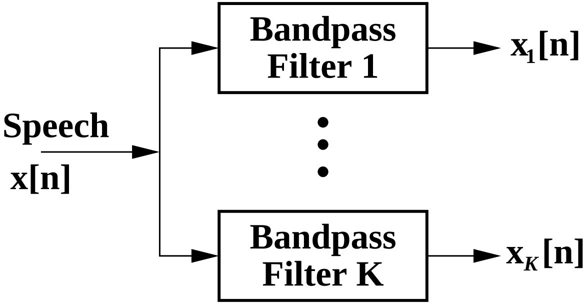
- **Mel scale**

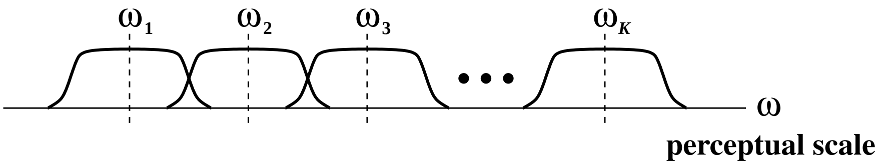
$$B(f) = 1127\ln(1 + f/700)$$





Filter Bank Analysis

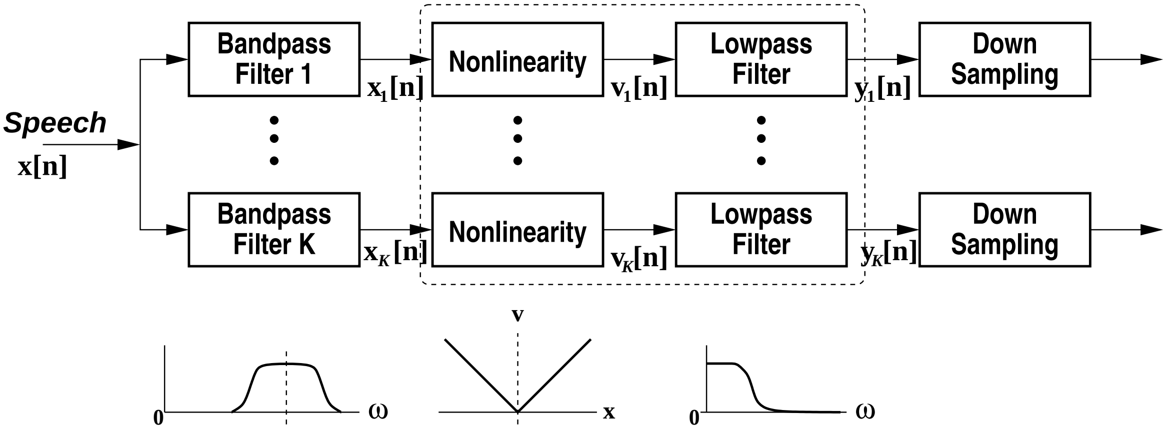




$$x_i[n] = h_i[n] * x[n] = \sum_{k=0}^{M_i-1} h_i[k]x[n-k]$$

$h_i[n]$: **Impulse response of Bandpass filter** i

Filter Bank Analysis *(cont. 2)*



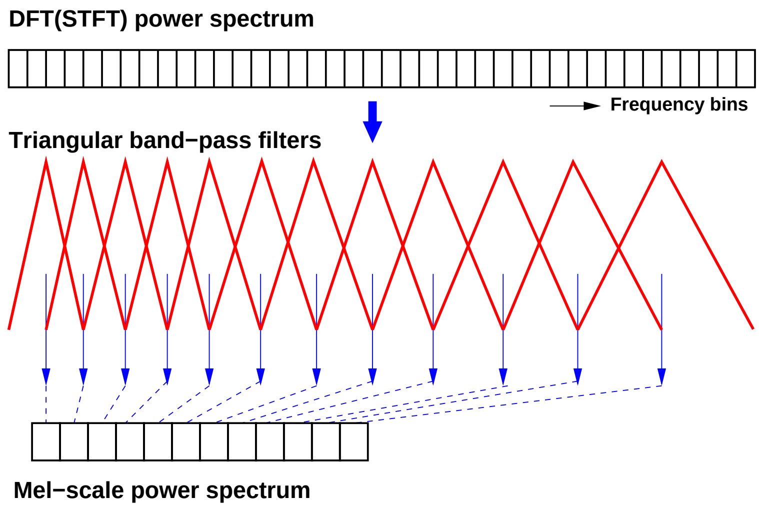
Trade-off problem

Freq. resolution	# of filters	length of filter	Time resolution
↗	↗	↗	↘
↘	↘	↘	↗

Filter Bank Analysis *(cont. 3)*

Another implementation of filter banks:

apply a mel-scale filter bank to STFT power spectrum to obtain mel-scale power spectrum



MFCC

MFCC: Mel-frequency Cepstral Coefficients $c[n]$

$$x[n] \xrightarrow{\textbf{DFT}} X[k] \rightarrow |X[k]|^2 \xrightarrow{\substack{\textbf{Mel-frequency} \\ \textbf{filterbank}}} \log|S[m]| \xrightarrow{\textbf{DCT}} c[n]$$

$$\text{DCT:} \quad c[n] = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} s[i] \cos\left(\frac{\pi n(i-0.5)}{N}\right), \quad \text{where } s[i] = \log|S[i]|$$

DFT: discrete Fourier transform, DCT: discrete cosine transform

- **MFCCs are widely used in HMM-based ASR systems.**
- **The first 12 MFCCs ($c[1] \sim c[12]$) are generally used.**

MFCC(cont. 2)

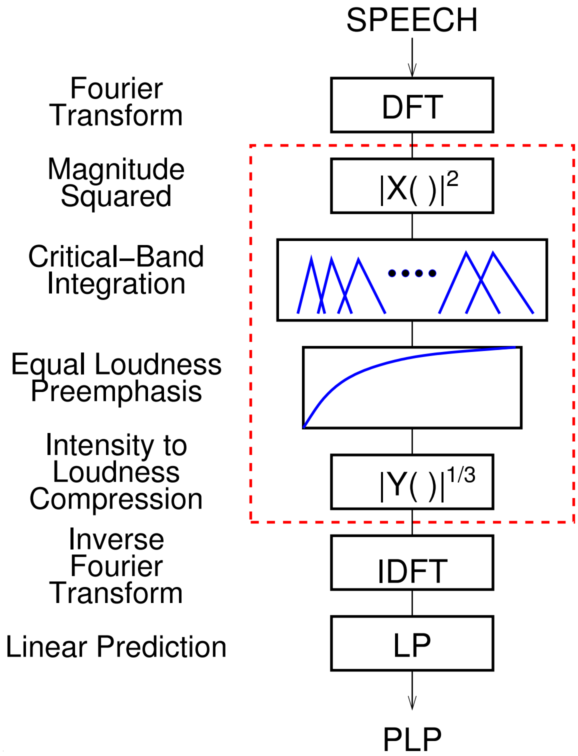
- **MFCCs are less correlated each other than DCT/Filter-bank based spectrum.**
- **Good compression rate.**

Feature	dimensionality / frame
Speech wave	400
DCT Spectrum	$64 \sim 256$
Filter-bank	$10 \sim 20$
MFCC	12

where $F_s = 16kHz$**, frame-width** $= 25ms$**, frame-shift** $= 10ms$ **are assumed.**

- **MFCCs show better ASR performance than filter-bank features, but MFCCs are not robust against noises.**

Perceptually-based Linear Prediction (PLP)



[Hermansky, 1985,1990]

PLP had been shown experimentally to be

- **more noise robust**
- **more speaker independent**

than MFCCs

Other features with low dimensionality

- **Formants ($F_1, F_2, F_3, \cdots$)**

  They are not used in modern ASR systems, but why ?

Using temporal features: dynamic features

In SP lab-sessions on speech recognition using HTK,

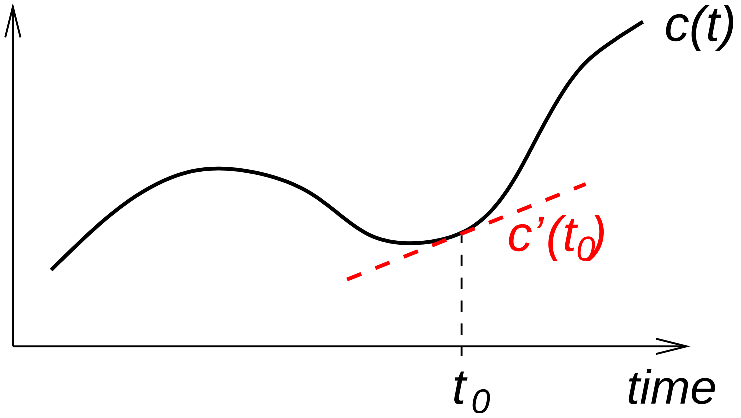
- **MFCCs, and energy**
- **Δ MFCCs, Δ energy**
- **Δ^2 MFCCs, Δ^2 energy**

$\Rightarrow \Delta *, \Delta^2 *$: **delta features**

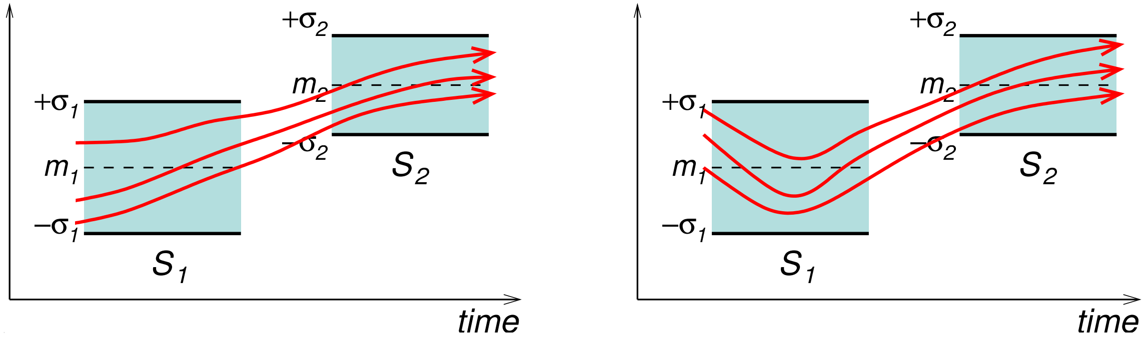
(dynamic features / time derivatives) [Furui, 1986]

continuous time	discrete time	
$c(t)$	$c[n]$	
$c'(t) = \dfrac{dc(t)}{dt}$	$\Delta c[n]$	$\sum_{i=-M}^{M} w_i \, c[n+i]$ **e.g.** $\Delta c[n] = \dfrac{c[n+1] - c[n-1]}{2}$
$c''(t) = \dfrac{d^2 c(t)}{dt^2}$	$\Delta^2 c[n]$	$\sum_{i=-M}^{M} w_i \, \Delta c[n+i]$

Using temporal features: dynamic features (cont. 2)



Using temporal features: dynamic features *(cont. 3)*



- **An acoustic feature vector, eg MFCCs, representing part of a speech signal is highly correlated with its neighbours.**
- **HMM based acoustic models assume there is no dependency between the observations.**
- **Those correlations can be captured to some extent by augmenting the original set of static acoustic features, eg. MFCCs, with dynamic features.**

General Feature Transformation

- Orthogonal transformation (orthogonal bases)
  - DCT (discrete cosine transform)
  - PCA (principal component analysis)
- Transformation based on the bases that maximises the separability between classes.
  - LDA (linear discriminant analysis) / Fisher's linear discrminant
  - HLDA (heteroscedastic linear discriminant analysis)

A comparison of speech features

I. Mporas, *et al.*, "Comparison of Speech Features on the Speech Recognition Task", Journal of Computer Science, Vol.3, pp.608–616, 2007.

Feature	WER(%)	SER(%)
SBC (16)	6.2	21.3
WPSR125 (16)	6.3	21.8
OWPF (16)	6.4	22.1
LFCC-FB40	6.9	23.5
HFCC-FB23	8.2	27.3
HFCC-FB40	8.7	28.2
PLP-FB19	9.0	29.4
MFCC-FB40	9.0	29.9

SBC	Subband-based Cepstral Coefficients
WPSR	Wavelet packet features
OWPF	Overlapping wavelet packet features
WPSR	Wavelet packet-based speech features
LFCC-FB	Linear-spaced filter-bank based cepstral coefficients
HFCC-FB	Human factor cepstral coefficients

NB The above result was obtained for TIMIT speech corpus. Results might change a lot under different conditions (e.g. noise, tasks, ASR systems)

Further topics on feature extraction

- Feature normalisation/enhancement in terms of
  - noise / environments
  - speakers / speaking styles
  - speech recognition
- Pitch (F_0) adapted feature extraction

SUMMARY

- **Nyquist Sampling theory**
- **Short-time Spectrum Analysis**
  - **Non-parametric method**
    - **Short-time Fourier Transform**
    - **Cepstrum, MFCC**
    - **Filter bank**
  - **Parametric methods**
    - **LPC, PLP**
  - **Windowing effect: trade-off between time and frequency resolutions**
- **Dynamic features (delta features)**
- **There is no best feature that can be used for any purposes, but MFCC is widely used for ASR and TTS.**

SUMMARY *(cont. 2)*

- **Front-end analysis has a great influence on ASR performance.**
- **For robust ASR in real environments, various techniques for front-end processing have been proposed. e.g. spectral subtraction (SS), cepstral mean normalisation (CMN)**
- **Spectrum analysis and feature extraction involve information loss and non-linear distortions. There is always a trade-off between accuracy and efficiency. (e.g. spatial resolution vs. temporal resolution)**

References

- **John N. Holmes, Wendy J. Holmes, ”Speech Synthesis and Recognition”, Taylor and Francis (2001), 2nd edition (chapter 2, 4, 10)**
- **http://mi.eng.cam.ac.uk/comp.speech/**
- **http://mi.eng.cam.ac.uk/˜ajr/SpeechAnalysis/**
- **http://cslu.cse.ogi.edu/HLTsurvey/**
- **B. Gold, N. Morgan, ”Speech and Audio Signal Processing: Processing and Perception of Speech and Music”, John Wiley and Sons (1999).**
- **“Spoken language processing: a guide to theory, algorithm, and system development”, Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, Prentice Hall (2001). isbn: 0130226165**

References*(cont. 2)*

- **"Robusness in Automatic Speech Recognition", J-C Junqua and J-P Hanton, , Kluwer Academic Publications (1996). isbn: 0-7923-9646-4**
- **"A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress", Sahar Bou-Ghazale and John H.L. Hansen, IEEE Trans SAP, vol. 8, no. 4, pp.429–442, July 2000.**