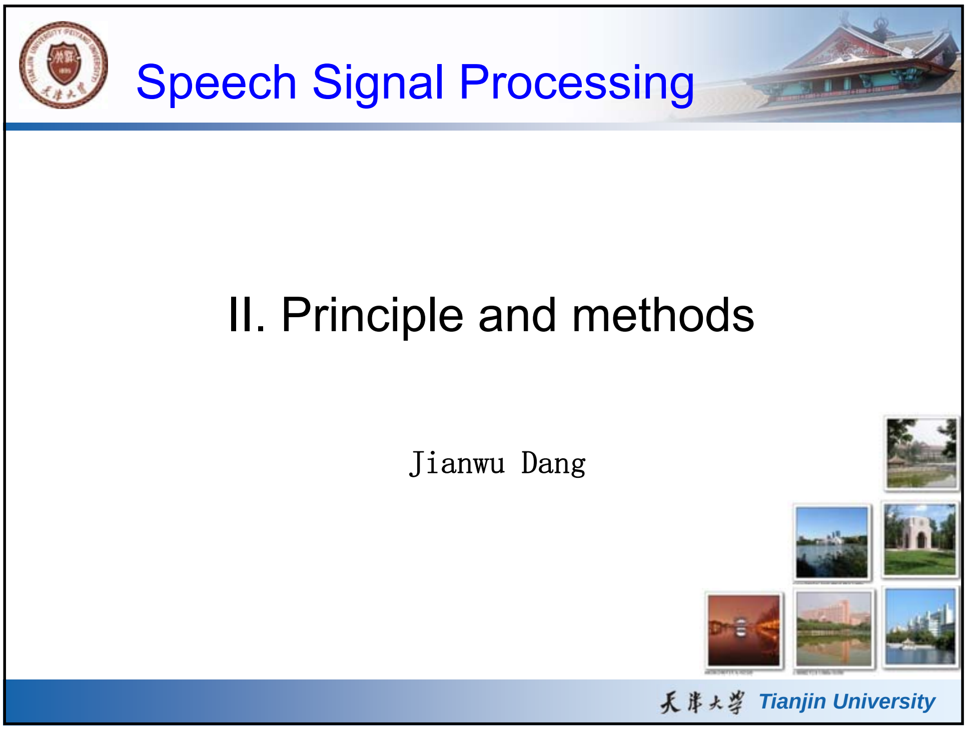
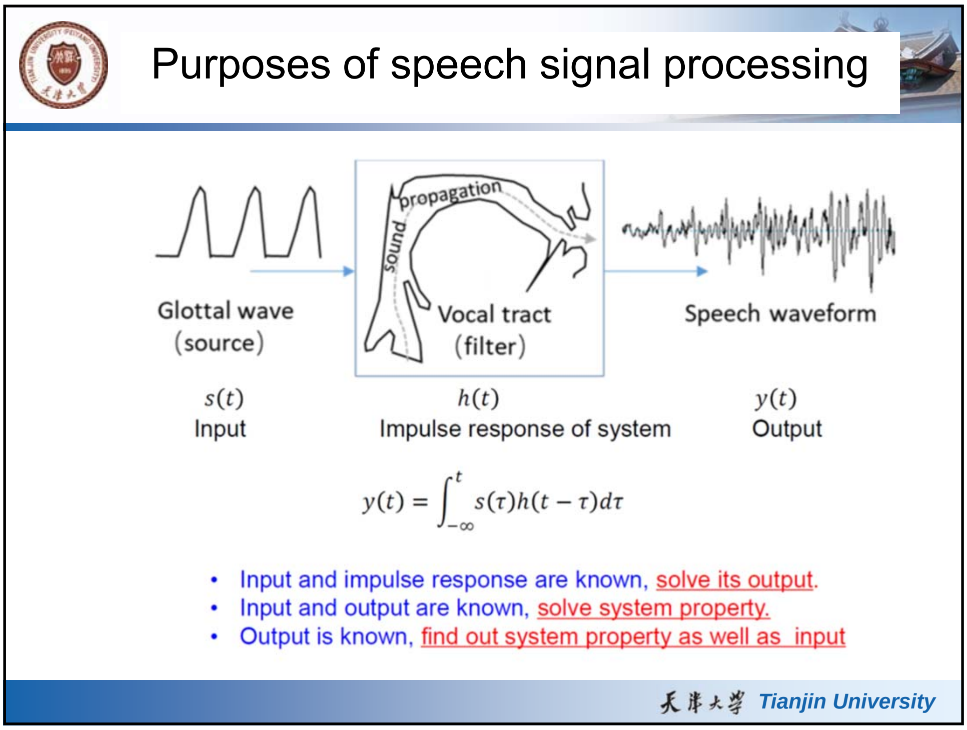


Speech Signal Processing

II. Principle and methods

Jianwu Dang



Purposes of speech signal processing

$s(t)$
Input

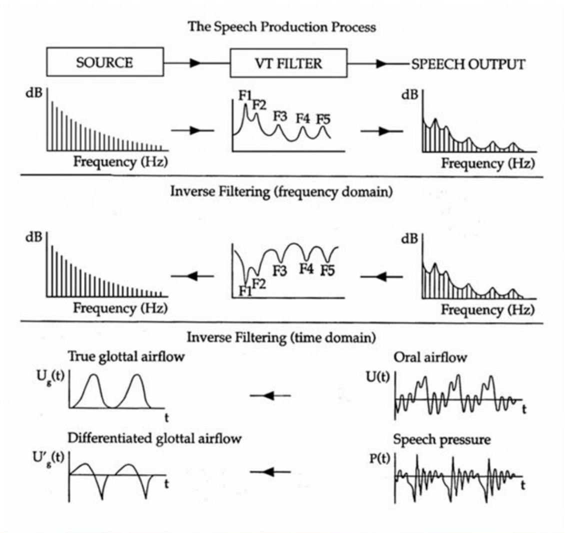
$h(t)$
Impulse response of system

$y(t)$
Output

$$y(t) = \int_{-\infty}^{t} s(\tau)h(t-\tau)d\tau$$

- Input and impulse response are known, solve its output.
- Input and output are known, solve system property.
- Output is known, find out system property as well as input

Purposes of speech signal processing



Output is known alone:

- Find system property

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

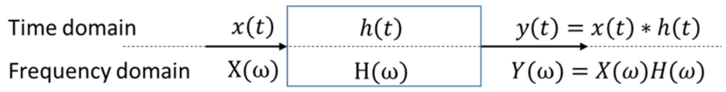
- Find source properties

$$E(z) = \left(1 - \sum_{i=1}^{p} a_i z^{-i}\right) S(z)$$

- Mechanism in time domain

$$\varepsilon(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i)$$

2.1 Modeling of speech production and perception



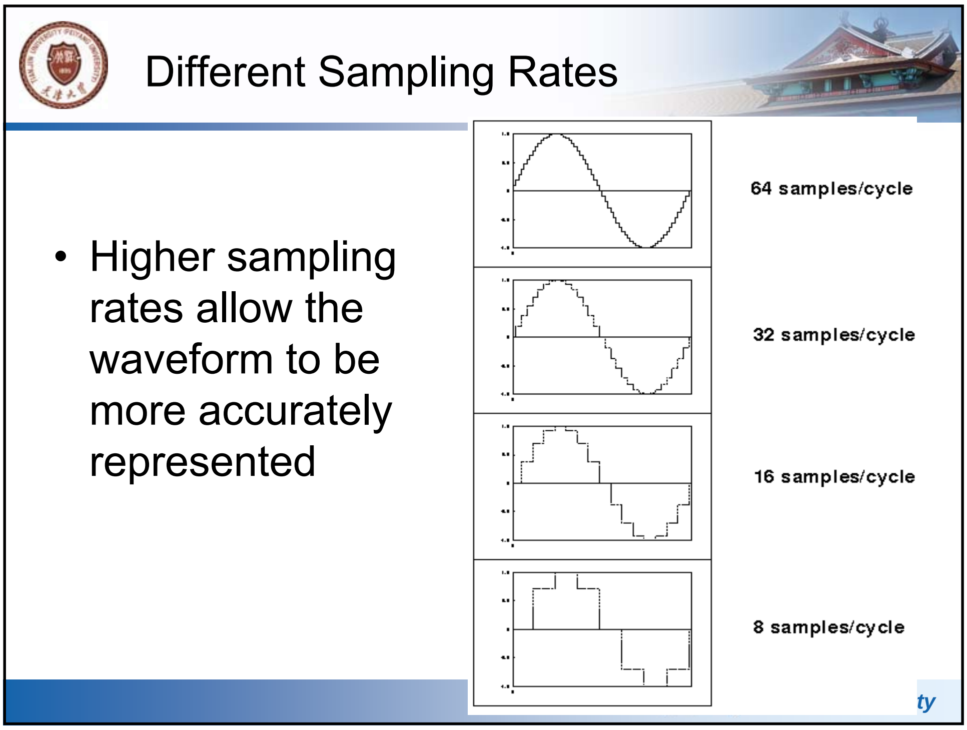
Time domain $x(t)$ $h(t)$ $y(t) = x(t) * h(t)$

Frequency domain $X(\omega)$ $H(\omega)$ $Y(\omega) = X(\omega)H(\omega)$

Time domain: $s(t) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)d\tau$

Frequency domain: $Y(\omega) = X(\omega)H(\omega)$

Sampling: $s(n) = s_a(nT)$， $-\infty < n < \infty$



Different Sampling Rates

- Higher sampling rates allow the waveform to be more accurately represented

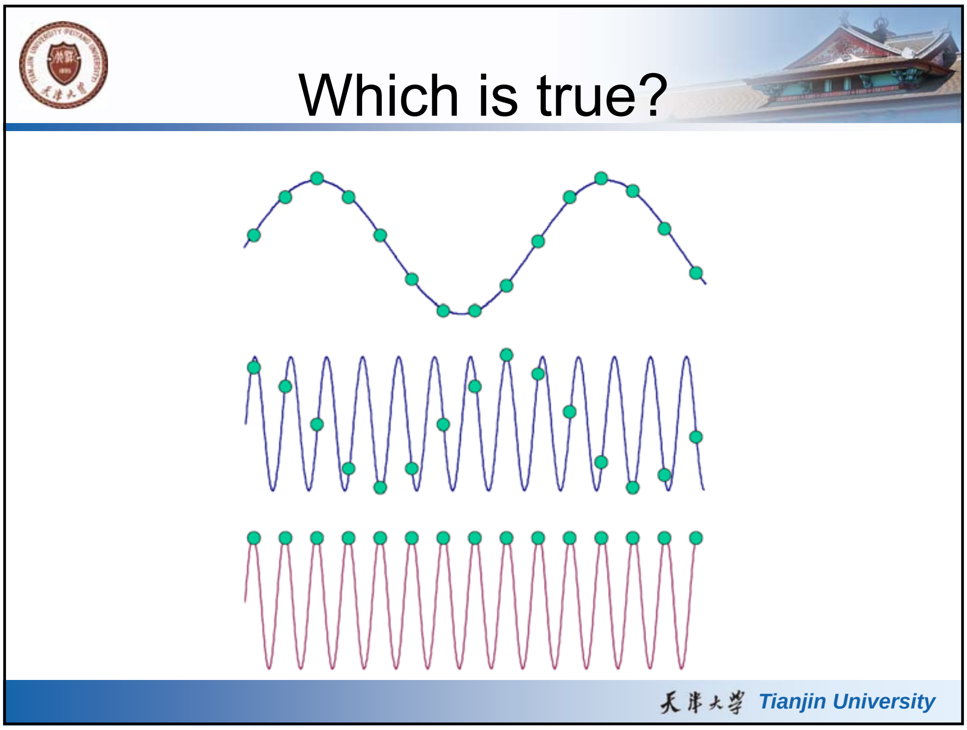
64 samples/cycle

32 samples/cycle

16 samples/cycle

8 samples/cycle

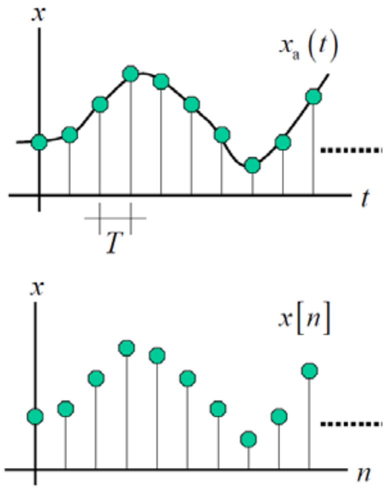
ty



Which is true?

天津大学 Tianjin University

What is sampling (traditional)



Sampling is an operation which acquires the values of a function at a certain interval.

$$x_s[n] = x_a(nT), \quad -\infty < n < \infty$$

$$= \int_{t=-\infty}^{\infty} x_a(t)\delta(t - nT)dt$$

$x_a(t)$: analog signal (continuous time signal)

T: sampling period

δ: dirac delta function

Sampling: Time Domain

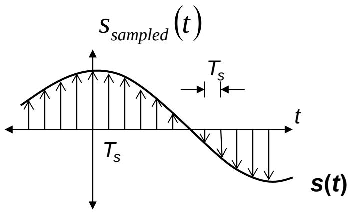
- Many signals originate as continuous-time signals, e.g. conventional music or voice
- By sampling a continuous-time signal at isolated, equally-spaced points in time, we obtain a sequence of numbers

$$s[n] = s(nT_s)$$

$n \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$

T_s is the sampling period.

$$s_{sampled}(t) = s(t)\underbrace{\sum_{n=-\infty}^{\infty}\delta(t - nT_s)}_{\text{impulse train}}$$



Sampled analog waveform

Sampling: Time Domain

Calculating the product $f_s(t)$ of a signal $f(t)$ and the function $\delta_T(t)$,

$$f_s(t) = f(t)\cdot\delta_T(t) = \sum_{n=-\infty}^{\infty} f(t)\delta(t-nT) \qquad (8.2)$$

$f_s(t)$ is a function which acquires the values of $f(t)$ with every interval T and is just a sampled function of $f(t)$. Thus, $\delta_T(t)$ is called sampling function and $f_s(t)$ is called sampled series.

Fourier transformation of sampling function

Since the sampling function $\delta_T(t)$ is

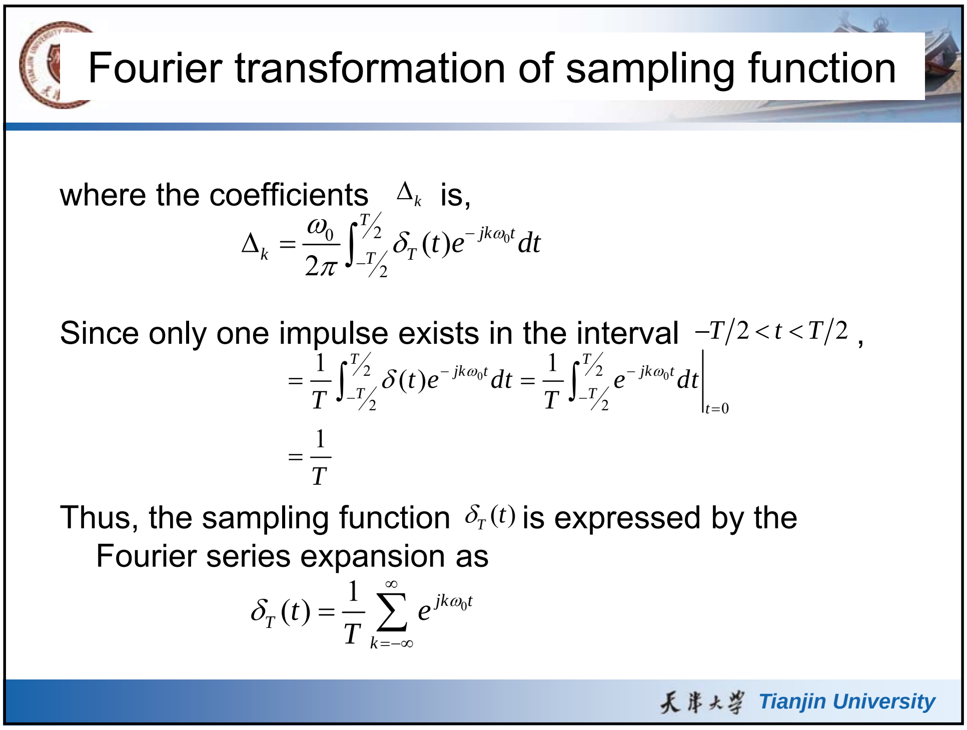
$$\delta_T(t+T) = \sum_{n=-\infty}^{\infty} \delta(t+T-nT) = \sum_{n=-\infty}^{\infty} \delta\big(t-(n-1)T\big)$$

$$= \sum_{m=-\infty}^{\infty} \delta(t-mT) = \delta_T(t)$$

then it is a periodic function with period T .

A periodic function is expanded by the Fourier series and $\delta_T(t)$is expressed as

$$\delta_T(t) = \sum_{k=-\infty}^{\infty} \Delta_k e^{-jk\omega_0 t} \qquad \omega_0 = \frac{2\pi}{T}$$



Fourier transformation of sampling function

where the coefficients Δ_k is,

$$\Delta_k = \frac{\omega_0}{2\pi}\int_{-T/2}^{T/2}\delta_T(t)e^{-jk\omega_0 t}dt$$

Since only one impulse exists in the interval $-T/2 < t < T/2$,

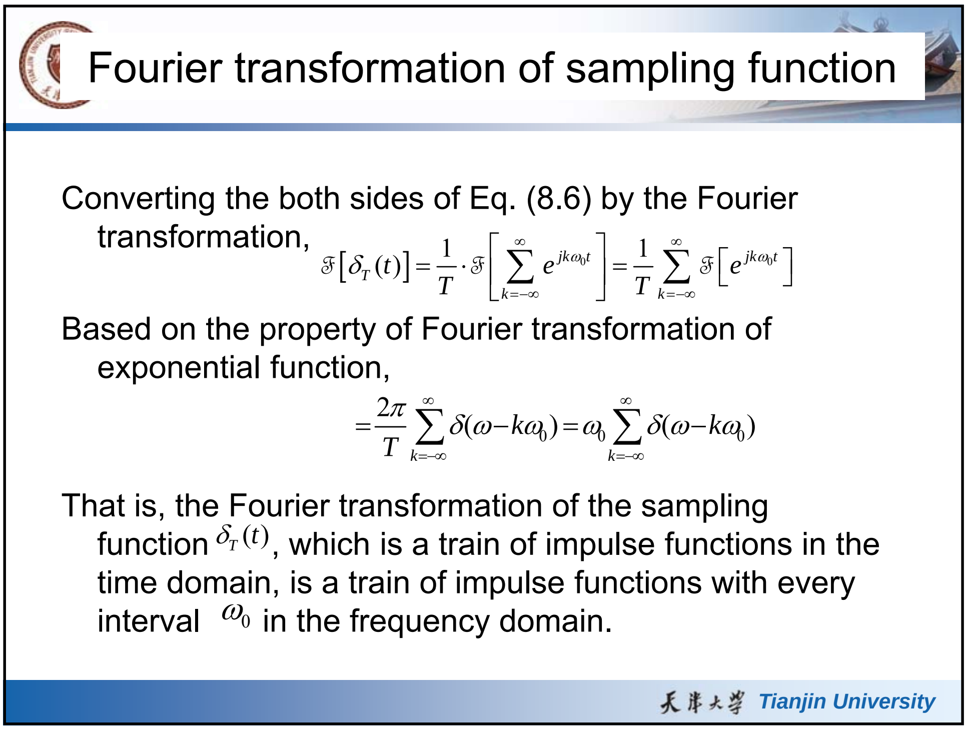
$$= \frac{1}{T}\int_{-T/2}^{T/2}\delta(t)e^{-jk\omega_0 t}dt = \frac{1}{T}\int_{-T/2}^{T/2}e^{-jk\omega_0 t}dt\bigg|_{t=0}$$

$$= \frac{1}{T}$$

Thus, the sampling function $\delta_T(t)$ is expressed by the Fourier series expansion as

$$\delta_T(t) = \frac{1}{T}\sum_{k=-\infty}^{\infty}e^{jk\omega_0 t}$$

天津大学 Tianjin University



Fourier transformation of sampling function

Converting the both sides of Eq. (8.6) by the Fourier transformation,

$$\mathfrak{F}[\delta_T(t)] = \frac{1}{T}\cdot\mathfrak{F}\left[\sum_{k=-\infty}^{\infty}e^{jk\omega_0 t}\right] = \frac{1}{T}\sum_{k=-\infty}^{\infty}\mathfrak{F}\left[e^{jk\omega_0 t}\right]$$

Based on the property of Fourier transformation of exponential function,

$$= \frac{2\pi}{T}\sum_{k=-\infty}^{\infty}\delta(\omega - k\omega_0) = \omega_0\sum_{k=-\infty}^{\infty}\delta(\omega - k\omega_0)$$

That is, the Fourier transformation of the sampling function $\delta_T(t)$, which is a train of impulse functions in the time domain, is a train of impulse functions with every interval ω_0 in the frequency domain.

天津大学 Tianjin University

Fourier transformation of sampled series

Since a sampled series $f_s(t)$ is the product of a signal $f(t)$ and a sampling function $\delta_T(t)$, if the Fourier transformations of these exist, these are connected by convolution as

$$F_s(\omega) = \mathfrak{F}[f_s(t)] = \frac{1}{2\pi}\mathfrak{F}[f(t)] * \mathfrak{F}[\delta_T(t)]$$

Thus,

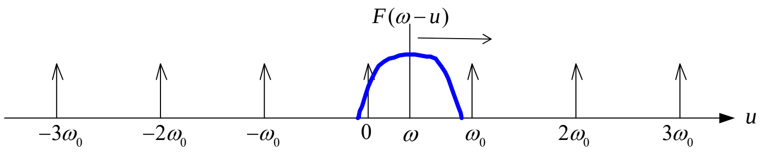
$$F_s(\omega) = \frac{1}{2\pi}\mathfrak{F}[f(t)] * \mathfrak{F}[\delta_T(t)] = \frac{1}{2\pi}F(\omega) * \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0)$$

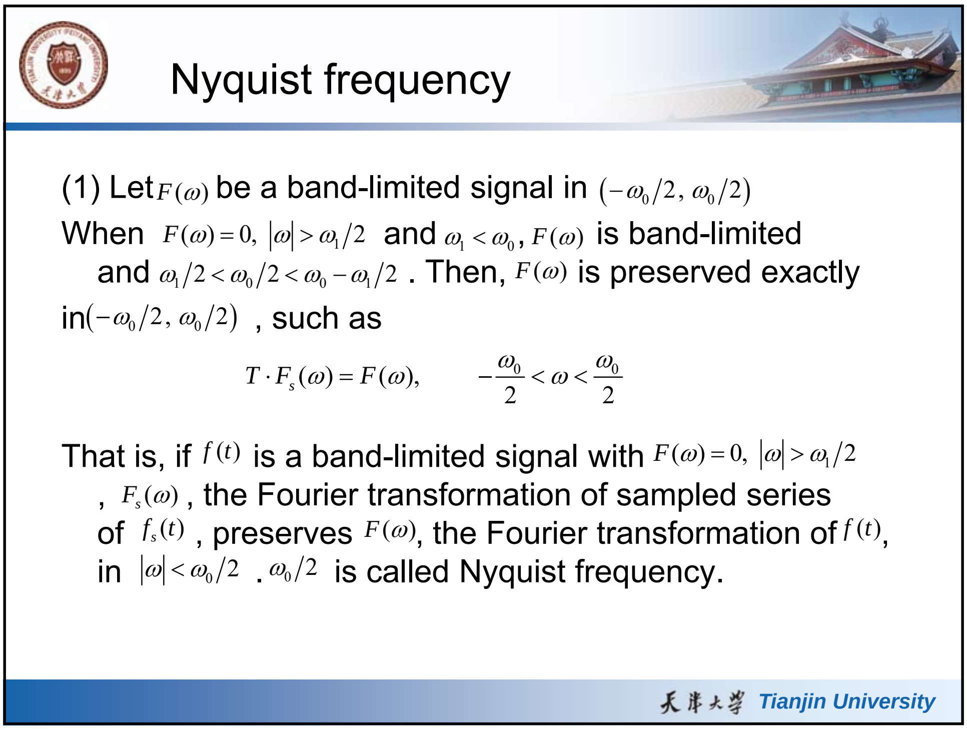
$$= \frac{\omega_0}{2\pi}\int_{-\infty}^{\infty} F(\omega - u) \sum_{n=-\infty}^{\infty} \delta(u - n\omega_0) du$$

$$= \frac{\omega_0}{2\pi}\sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} F(\omega - u)\delta(u - n\omega_0) du$$

$$= \frac{\omega_0}{2\pi}\sum_{n=-\infty}^{\infty} F(\omega - n\omega_0) = \frac{1}{T}\sum_{n=-\infty}^{\infty} F(\omega - n\omega_0)$$

Convolution of signal and sampling function in Frequency domain





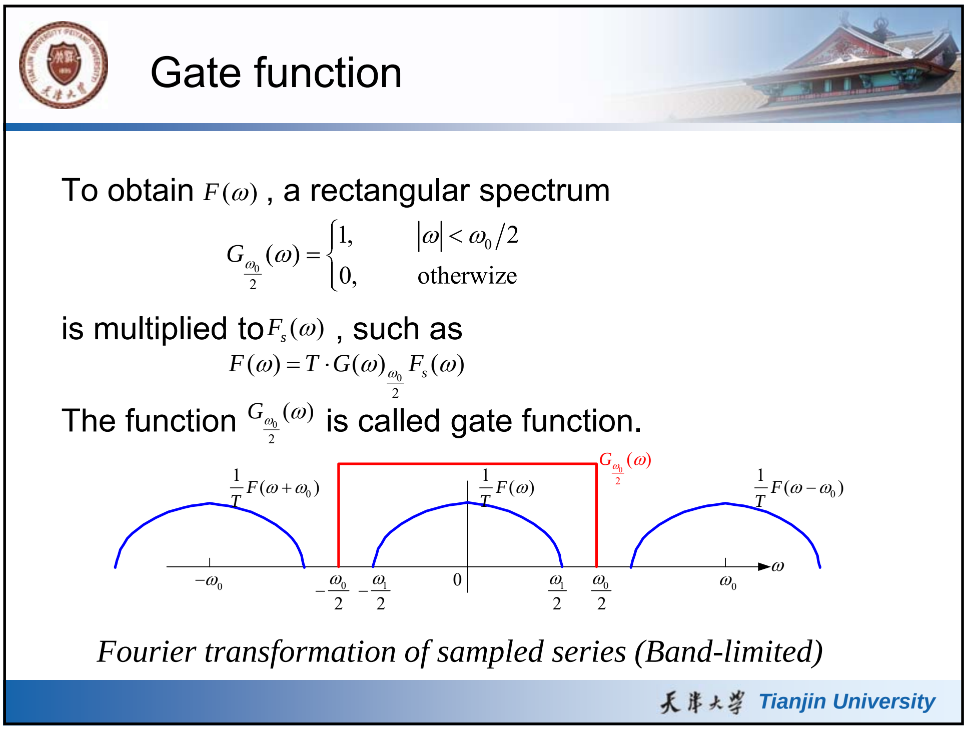
Nyquist frequency

(1) Let $F(\omega)$ be a band-limited signal in $(-\omega_0/2,\ \omega_0/2)$

When $F(\omega)=0,\ |\omega|>\omega_1/2$ and $\omega_1<\omega_0$, $F(\omega)$ is band-limited and $\omega_1/2<\omega_0/2<\omega_0-\omega_1/2$. Then, $F(\omega)$ is preserved exactly in$(-\omega_0/2,\ \omega_0/2)$, such as

$$T\cdot F_s(\omega)=F(\omega),\qquad -\frac{\omega_0}{2}<\omega<\frac{\omega_0}{2}$$

That is, if $f(t)$ is a band-limited signal with $F(\omega)=0,\ |\omega|>\omega_1/2$, $F_s(\omega)$, the Fourier transformation of sampled series of $f_s(t)$, preserves $F(\omega)$, the Fourier transformation of $f(t)$, in $|\omega|<\omega_0/2$. $\omega_0/2$ is called Nyquist frequency.



Gate function

To obtain $F(\omega)$, a rectangular spectrum

$$G_{\frac{\omega_0}{2}}(\omega)=\begin{cases}1, & |\omega|<\omega_0/2\\ 0, & \text{otherwize}\end{cases}$$

is multiplied to$F_s(\omega)$, such as

$$F(\omega)=T\cdot G(\omega)_{\frac{\omega_0}{2}}F_s(\omega)$$

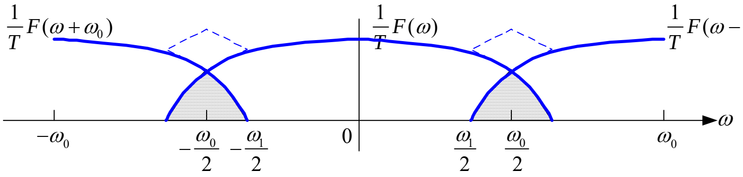
The function $G_{\frac{\omega_0}{2}}(\omega)$ is called gate function.

Fourier transformation of sampled series (Band-limited)

Aliasing

(2) Let $F(\omega)$ be NOT band-limited signal in $(-\omega_0/2,\ \omega_0/2)$

When $F(\omega)=0,\ |\omega|>\omega_1/2$ and $\omega_1>\omega_0$, as shown in Fig. 8.4, there are overlaps around $(2n+1)\omega_0/2,\ n=\pm1,\pm2,\ldots$. Thus, even if the rectangular spectrum $G_{\frac{\omega_0}{2}}(\omega)$ is multiplied to $F_s(\omega)$, $F(\omega)$, the Fourier transformation of $f(t)$, can not be extracted. This phenomenon is called aliasing.

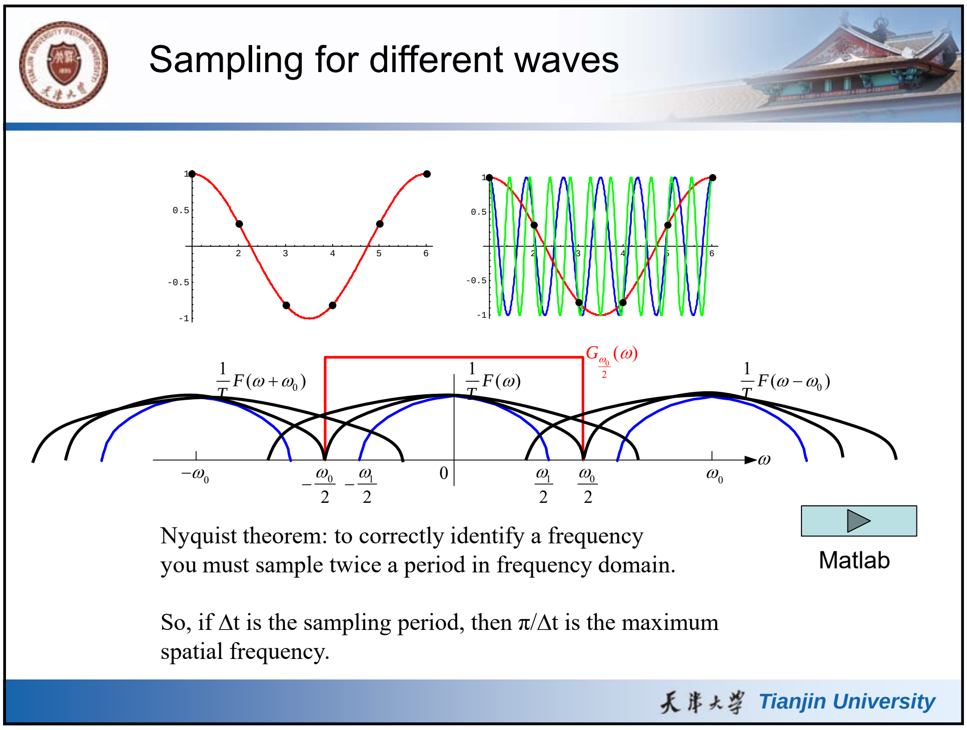


Fourier transformation of sampled series (NOT band-limited)

Low-pass filtered and cut-off frequency

That is, when sampling a signal $f(t)$ with period T and obtaining the sampled series $f_s(t)$, if $f(t)$ is band-limited with $\omega_0/2=\pi/T$, i.e., $F(\omega)=0,\ |\omega|>\omega_0/2$, then $F(\omega)$, the Fourier transformation of $f(t)$, is preserved exactly in$F_s(\omega)$, the Fourier transformation of sampled series of $f_s(t)$ with $|\omega|<\omega_0/2$.

Actually, when sampling a signal $f(t)$ with period T , i.e., sampling frequency $f=1/T$ [Hz], $f(t)$ is firstly low-pass filtered whose cut-off frequency is lower than $f/2=1/2T$ and then sampled.

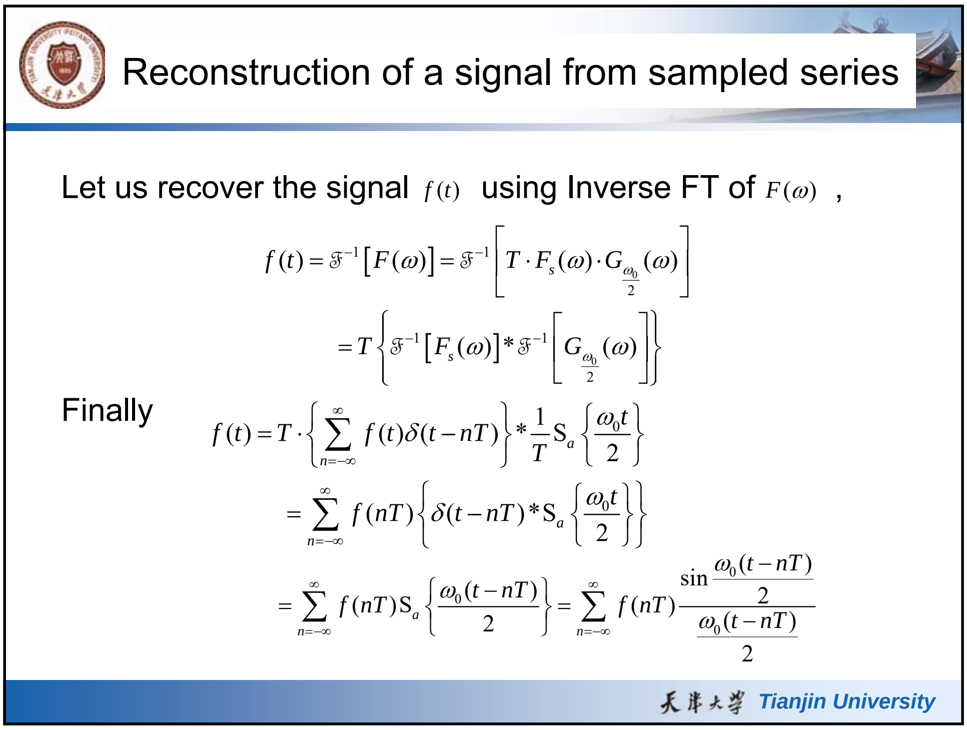


Sampling for different waves

Nyquist theorem: to correctly identify a frequency you must sample twice a period in frequency domain.

So, if Δt is the sampling period, then π/Δt is the maximum spatial frequency.

天津大学 Tianjin University



Reconstruction of a signal from sampled series

Let us recover the signal $f(t)$ using Inverse FT of $F(\omega)$,

$$f(t)=\mathfrak{F}^{-1}\left[F(\omega)\right]=\mathfrak{F}^{-1}\left[T\cdot F_s(\omega)\cdot G_{\frac{\omega_0}{2}}(\omega)\right]$$

$$=T\left\{\mathfrak{F}^{-1}\left[F_s(\omega)\right]*\mathfrak{F}^{-1}\left[G_{\frac{\omega_0}{2}}(\omega)\right]\right\}$$

Finally

$$f(t)=T\cdot\left\{\sum_{n=-\infty}^{\infty}f(t)\delta(t-nT)\right\}*\frac{1}{T}\mathrm{S}_a\left\{\frac{\omega_0 t}{2}\right\}$$

$$=\sum_{n=-\infty}^{\infty}f(nT)\left\{\delta(t-nT)*\mathrm{S}_a\left\{\frac{\omega_0 t}{2}\right\}\right\}$$

$$=\sum_{n=-\infty}^{\infty}f(nT)\mathrm{S}_a\left\{\frac{\omega_0(t-nT)}{2}\right\}=\sum_{n=-\infty}^{\infty}f(nT)\frac{\sin\frac{\omega_0(t-nT)}{2}}{\frac{\omega_0(t-nT)}{2}}$$

天津大学 Tianjin University

Reconstruction of signal

(1) Case: $t = nT$

$$f(nT) = \sum_{k=-\infty}^{\infty} f(kT)\frac{\sin \omega_0(nT-kT)\Big/2}{\omega_0(nT-kT)\Big/2}, \qquad \omega_0 = \frac{2\pi}{T}$$

$$= \begin{cases} f(nT), & k = n \\ 0, & k \neq n \end{cases}$$

Reconstructed signal at the sampled points is the original.

(2) Case: $t = t_0 \neq nT$

$$f(t_0) = \sum_{k=-\infty}^{\infty} f(kT)\frac{\sin \omega_0(t_0-kT)\Big/2}{\omega_0(t_0-kT)\Big/2}, \qquad \omega_0 = \frac{2\pi}{T}$$

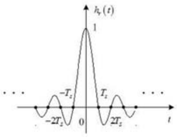
$f(t_0)$ can be obtained by interpolating during sample points using a weighted sum of sampling functions.

That is, in order to reconstruct the signal $f(t)$ from the sampled series $f_s(t)$, it is needed to convolute $f_s(t)$ with the sampling function which is the impulse response of the ideal low-pass filter.

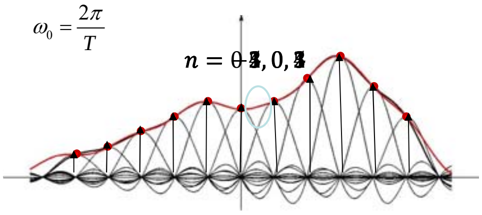
Examples of restoration

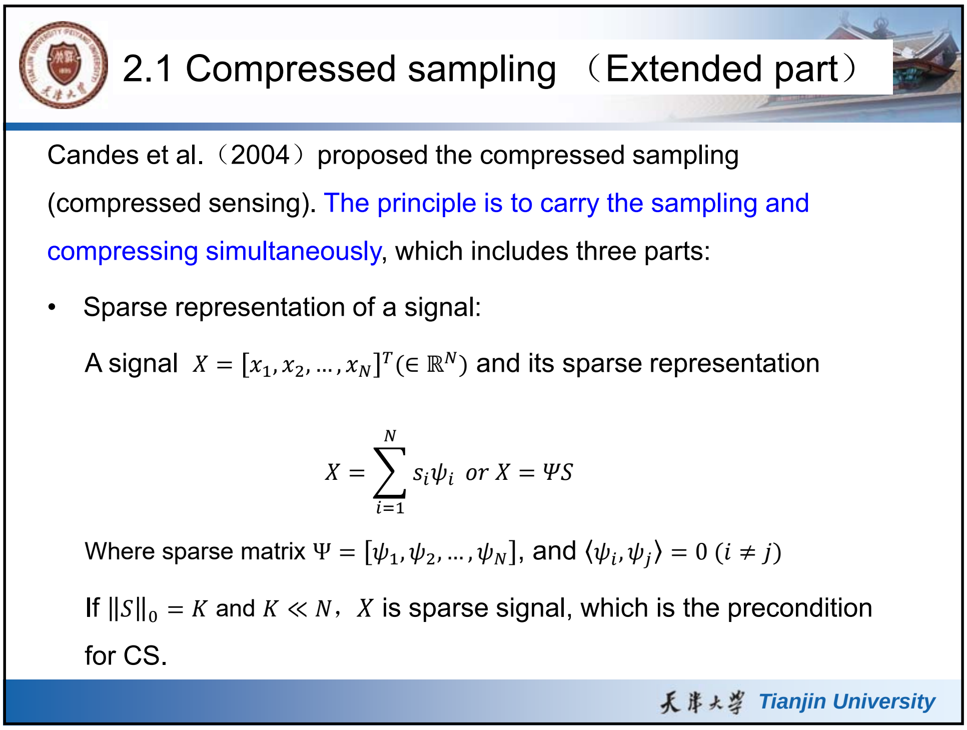
$$f(t) = T\cdot\left\{\sum_{n=-\infty}^{\infty} f(t)\delta(t-nT)\right\} * \frac{1}{T}\mathrm{S}_a\left\{\frac{\omega_0 t}{2}\right\}$$

$$= \sum_{n=-\infty}^{\infty} f(nT)\left\{\delta(t-nT) * \mathrm{S}_a\left\{\frac{\omega_0 t}{2}\right\}\right\}$$



$$f(t_0) = \sum_{k=-\infty}^{\infty} f(kT)\frac{\sin \omega_0(t_0-kT)\Big/2}{\omega_0(t_0-kT)\Big/2}, \qquad \omega_0 = \frac{2\pi}{T}$$





2.1 Compressed sampling （Extended part）

Candes et al.（2004）proposed the compressed sampling (compressed sensing). The principle is to carry the sampling and compressing simultaneously, which includes three parts:

- Sparse representation of a signal:

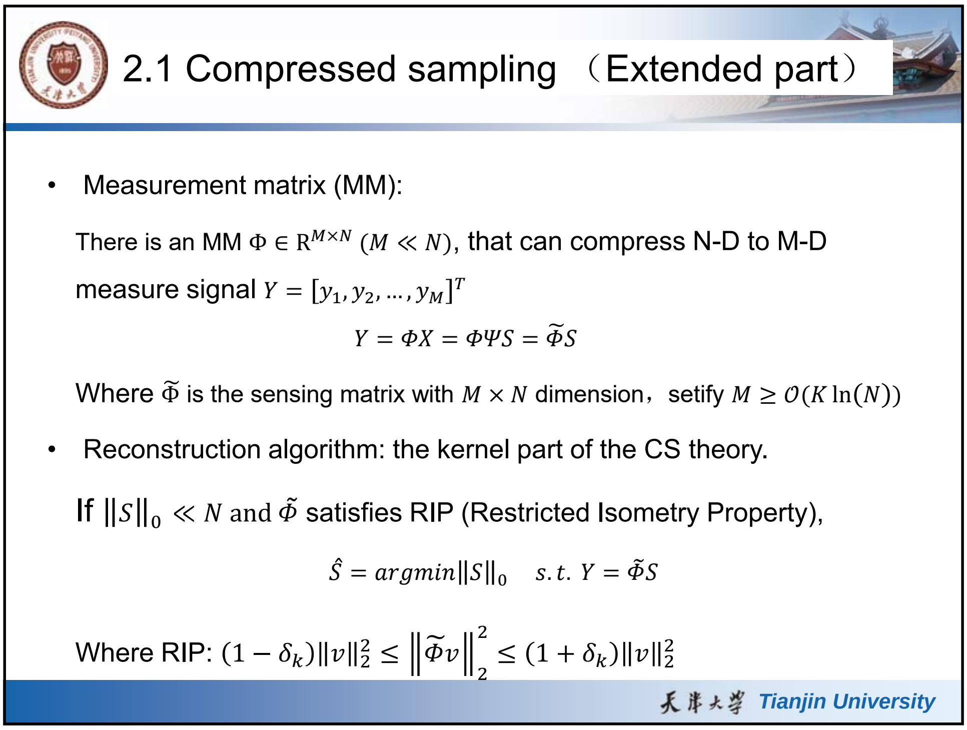
  A signal $X = [x_1, x_2, \dots, x_N]^T (\in \mathbb{R}^N)$ and its sparse representation

$$X = \sum_{i=1}^{N} s_i \psi_i \;\; or \; X = \Psi S$$

  Where sparse matrix $\Psi = [\psi_1, \psi_2, \dots, \psi_N]$, and $\langle \psi_i, \psi_j \rangle = 0 \; (i \neq j)$

  If $\|S\|_0 = K$ and $K \ll N$， X is sparse signal, which is the precondition for CS.

天津大学 Tianjin University



2.1 Compressed sampling （Extended part）

- Measurement matrix (MM):

  There is an MM $\Phi \in \mathrm{R}^{M \times N}$ $(M \ll N)$, that can compress N-D to M-D measure signal $Y = [y_1, y_2, \dots, y_M]^T$

$$Y = \Phi X = \Phi \Psi S = \widetilde{\Phi} S$$

  Where $\widetilde{\Phi}$ is the sensing matrix with $M \times N$ dimension，setify $M \geq O(K \ln(N))$

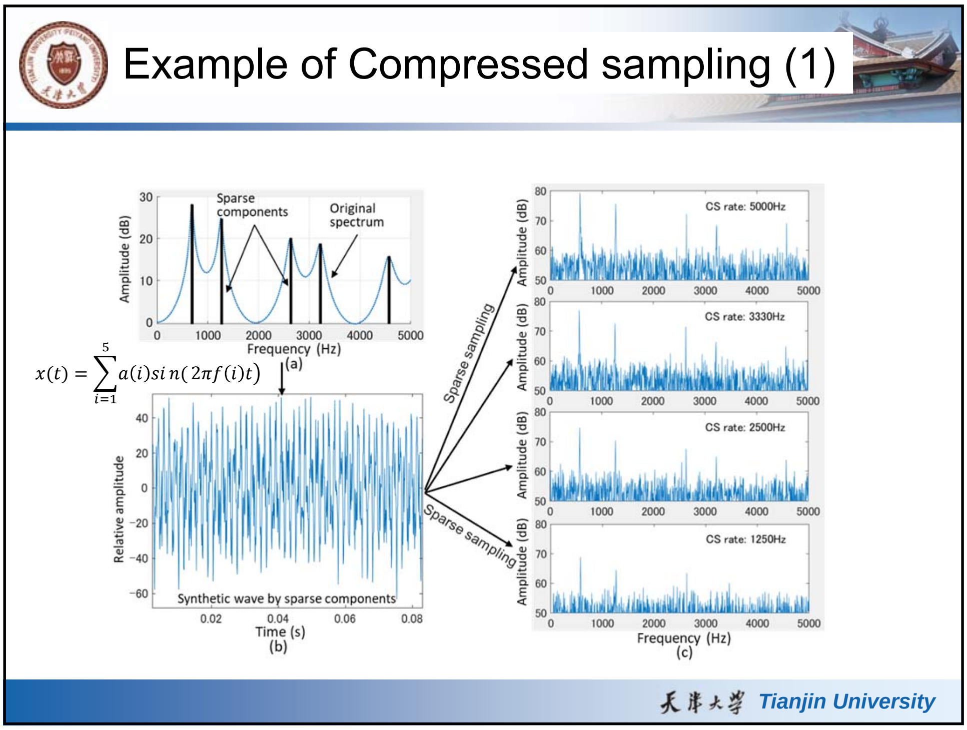
- Reconstruction algorithm: the kernel part of the CS theory.

  If $\|S\|_0 \ll N$ and $\tilde{\Phi}$ satisfies RIP (Restricted Isometry Property),

$$\hat{S} = argmin \|S\|_0 \quad s.t. \; Y = \tilde{\Phi} S$$

  Where RIP: $(1 - \delta_k)\|v\|_2^2 \leq \left\|\widetilde{\Phi} v\right\|_2^2 \leq (1 + \delta_k)\|v\|_2^2$

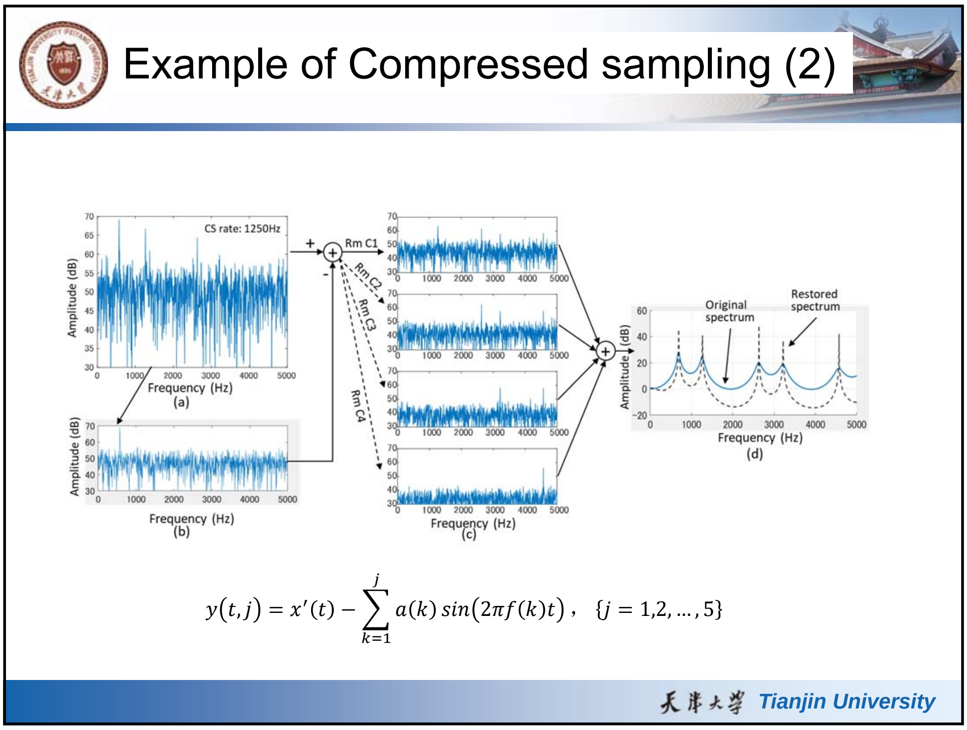
天津大学 Tianjin University



Example of Compressed sampling (1)

$$x(t) = \sum_{i=1}^{5} a(i) sin(2\pi f(i)t)$$

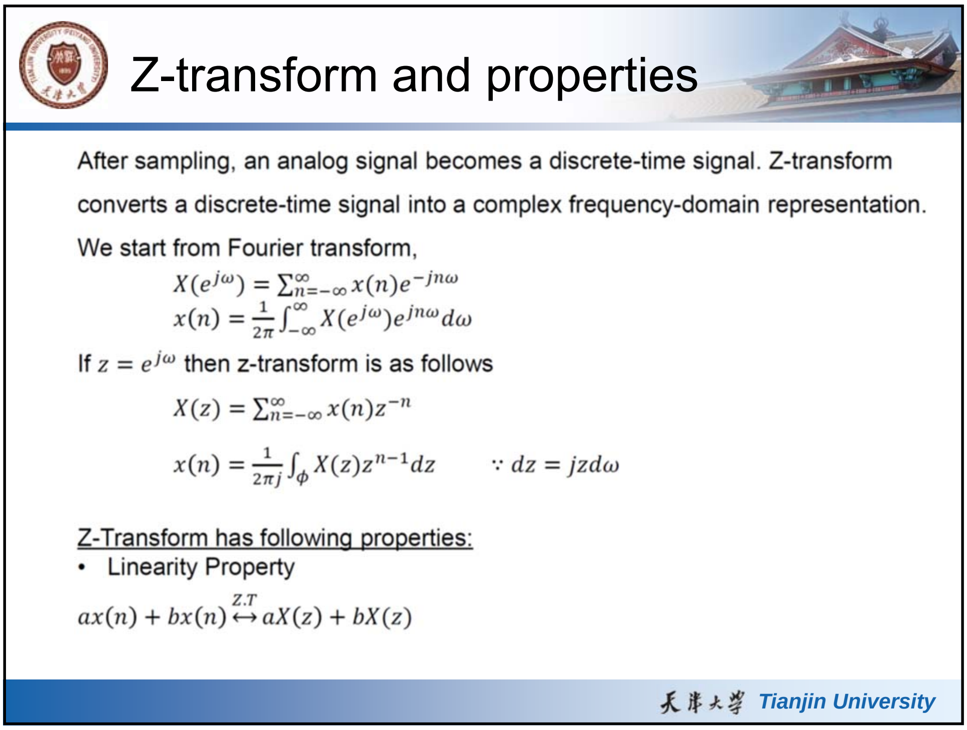
天津大学 *Tianjin University*



Example of Compressed sampling (2)

$$y(t,j) = x'(t) - \sum_{k=1}^{j} a(k)\, sin\big(2\pi f(k)t\big)\,, \quad \{j = 1,2,\dots,5\}$$

天津大学 *Tianjin University*



Z-transform and properties

After sampling, an analog signal becomes a discrete-time signal. Z-transform converts a discrete-time signal into a complex frequency-domain representation.

We start from Fourier transform,

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

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

If $z = e^{j\omega}$ then z-transform is as follows

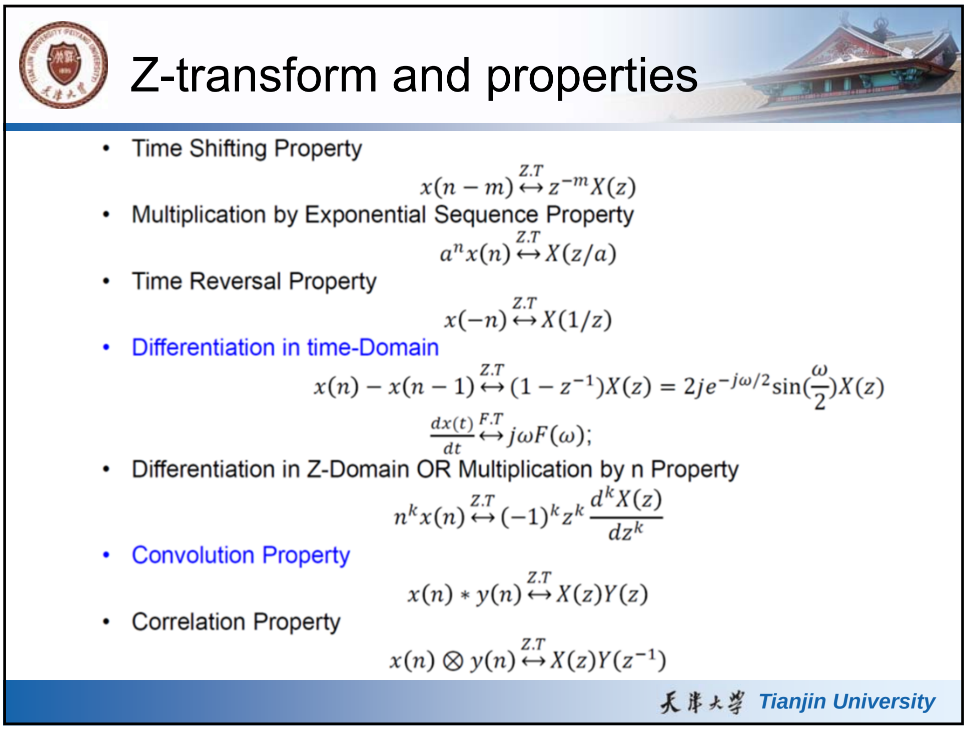
$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$$

$$x(n) = \frac{1}{2\pi j}\int_{\phi} X(z)z^{n-1}dz \qquad \because dz = jzd\omega$$

Z-Transform has following properties:

- Linearity Property

$$ax(n) + bx(n) \overset{Z.T}{\leftrightarrow} aX(z) + bX(z)$$



Z-transform and properties

- Time Shifting Property

$$x(n-m) \overset{Z.T}{\leftrightarrow} z^{-m}X(z)$$

- Multiplication by Exponential Sequence Property

$$a^n x(n) \overset{Z.T}{\leftrightarrow} X(z/a)$$

- Time Reversal Property

$$x(-n) \overset{Z.T}{\leftrightarrow} X(1/z)$$

- Differentiation in time-Domain

$$x(n) - x(n-1) \overset{Z.T}{\leftrightarrow} (1 - z^{-1})X(z) = 2je^{-j\omega/2}\sin(\frac{\omega}{2})X(z)$$

$$\frac{dx(t)}{dt} \overset{F.T}{\leftrightarrow} j\omega F(\omega);$$

- Differentiation in Z-Domain OR Multiplication by n Property

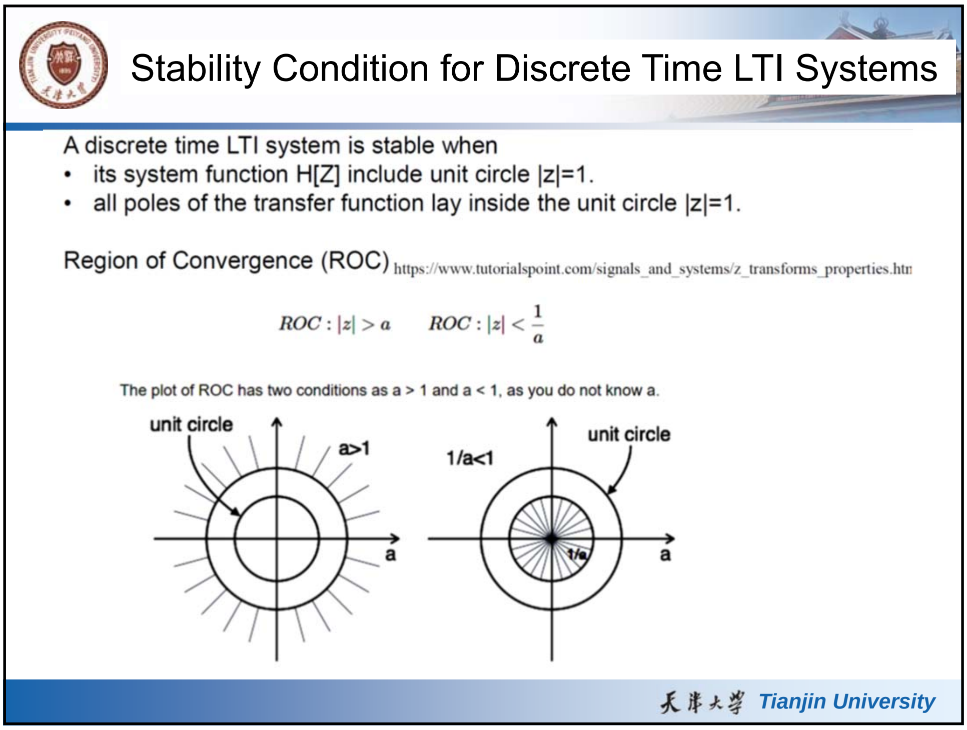
$$n^k x(n) \overset{Z.T}{\leftrightarrow} (-1)^k z^k \frac{d^k X(z)}{dz^k}$$

- Convolution Property

$$x(n) * y(n) \overset{Z.T}{\leftrightarrow} X(z)Y(z)$$

- Correlation Property

$$x(n) \otimes y(n) \overset{Z.T}{\leftrightarrow} X(z)Y(z^{-1})$$



Stability Condition for Discrete Time LTI Systems

A discrete time LTI system is stable when

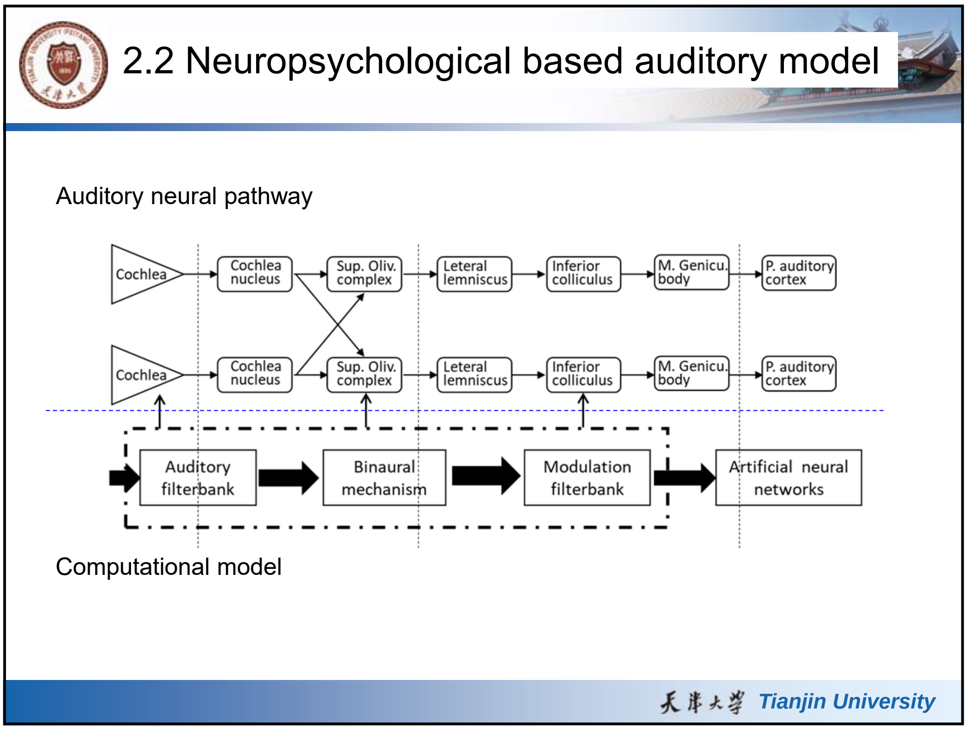
- its system function H[Z] include unit circle |z|=1.
- all poles of the transfer function lay inside the unit circle |z|=1.

Region of Convergence (ROC) https://www.tutorialspoint.com/signals_and_systems/z_transforms_properties.htn

$$ROC: |z| > a \qquad ROC: |z| < \frac{1}{a}$$

The plot of ROC has two conditions as a > 1 and a < 1, as you do not know a.

天津大学 Tianjin University

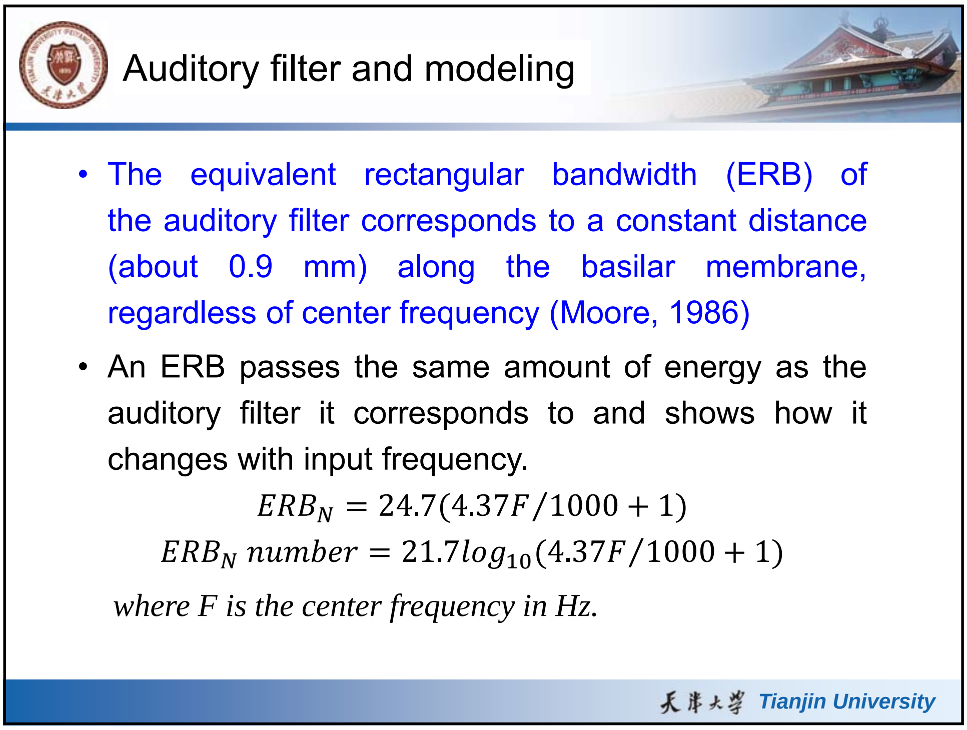


2.2 Neuropsychological based auditory model

Auditory neural pathway

Computational model

天津大学 Tianjin University



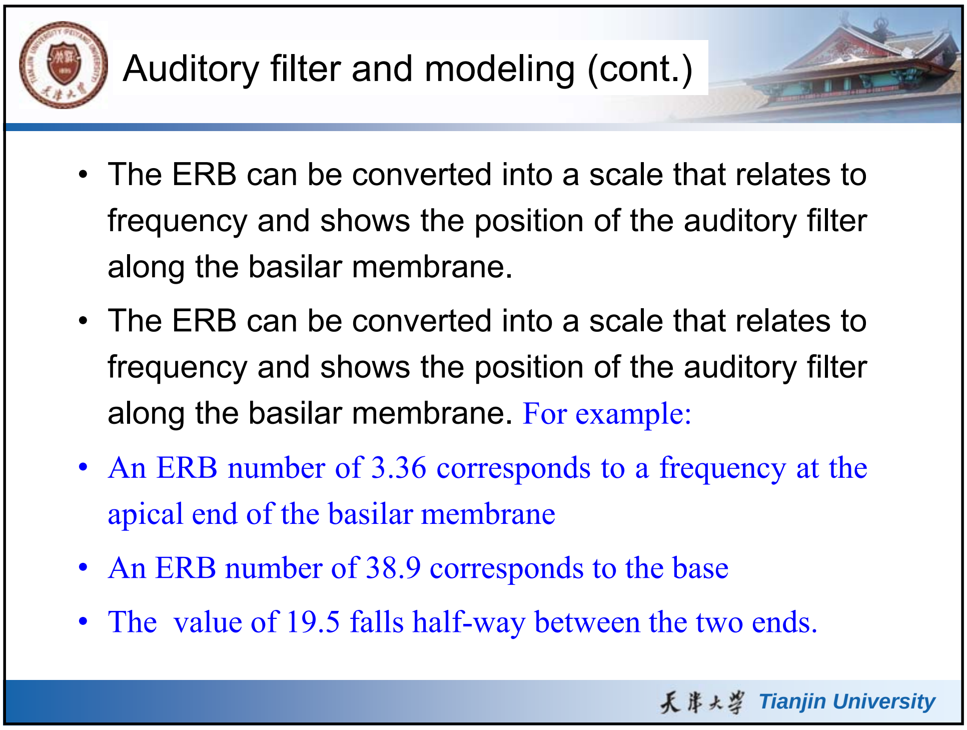
Auditory filter and modeling

- The equivalent rectangular bandwidth (ERB) of the auditory filter corresponds to a constant distance (about 0.9 mm) along the basilar membrane, regardless of center frequency (Moore, 1986)
- An ERB passes the same amount of energy as the auditory filter it corresponds to and shows how it changes with input frequency.

$$ERB_N = 24.7(4.37F/1000 + 1)$$

$$ERB_N\ number = 21.7log_{10}(4.37F/1000 + 1)$$

where F is the center frequency in Hz.



Auditory filter and modeling (cont.)

- The ERB can be converted into a scale that relates to frequency and shows the position of the auditory filter along the basilar membrane.
- The ERB can be converted into a scale that relates to frequency and shows the position of the auditory filter along the basilar membrane. For example:
- An ERB number of 3.36 corresponds to a frequency at the apical end of the basilar membrane
- An ERB number of 38.9 corresponds to the base
- The value of 19.5 falls half-way between the two ends.

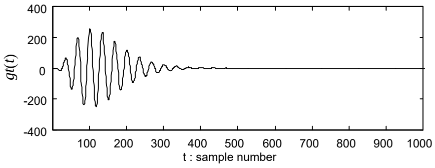
Auditory filter and modeling (cont.)

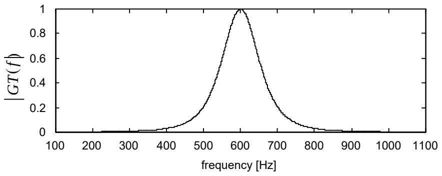
- The Gammatone filter is a model of the auditory filters is. It provides a simple linear filter, which is therefore easy to implement, but cannot by itself account for nonlinear aspects of the auditory system.

$$g_i(t) = at^{n-1}e^{-2\pi B(f_i)t}\,cos(2\pi f_i\,t + \varphi)$$

Where $B\left(f_i\right) = 0.1039f_i + 24.7$， *n=4*

- $at^{n-1}e^{-2\pi B(f_i)t}$ is the same formulation as gamma distribution.





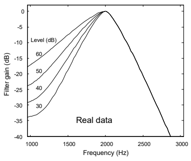
Auditory filter and modeling (cont.)

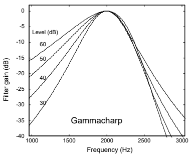
- Gammacharp Filter (Irino & Patterson, 1997)
  - Add sound level dependency to gammatone filter
  - Optimal shape for auditory filter under minimum uncertainty on time-scale representation
- An improved version of the gammatone model is the gammachirp filter,

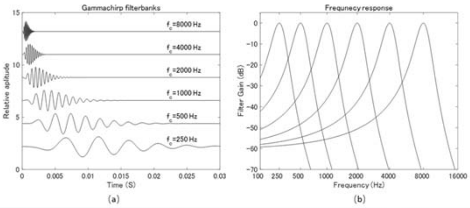
$$g_i^c(t) = at^{n-1}e^{-2\pi B(f_i)t}\,cos(2\pi f_i\,t + cln(t) + \varphi)$$

where n=4, c=-2.96

  - including charp term $cln(t)$.
  - Realizing level dependency by controlling c







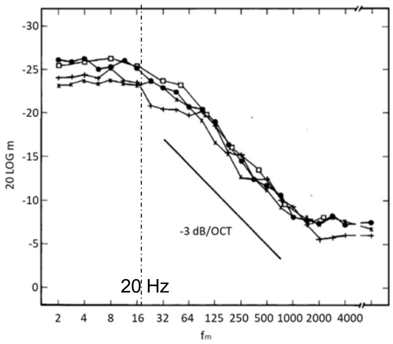
(a) (b)

Temporal modulation transfer function (TMTF)

A temporal modulation transfer function (TMTF) reflects how well a person or a neuron can follow amplitude modulations of a sound.

$$y(t) = [1 + m \cdot sin(2\pi f_m t)]x(t)$$

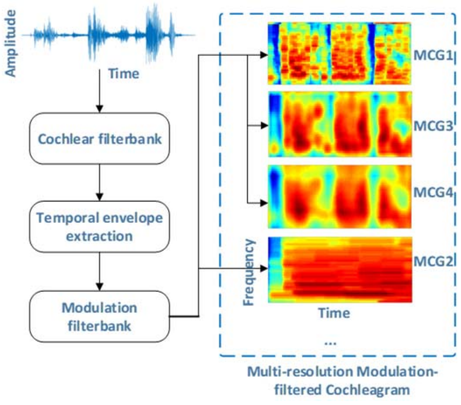
where $x(t)$ is original sound，f_m is modulation frequency，*m* (0~1) is depth of the amplitude modulation (AM). TMTF is the minimal detectable value.



The modulation frequency ranged between 2-16Hz, which is much lower than the minimal resolution frequency (20Hz) of the basilar membrane.

Viemeister, Neal, F., Temporal modulation transfer functions based upon modulation thresholds. Journal of the Acoustical Society of America, 1979. 66(5): p. 1364.

Multi-resolution modulation-filtered cochleagram feature for emotion recognition



MMCG extraction
(Peng et al, 2021)

Modulation unit

$$s_{mu}(c,i) = \mathrm{w}(t_w) \cdot s_m\big(\mathrm{c},(i-1)\cdot Len_s + t_w\big)$$

Where $s_m(m,n) = m_f(m,t) * s_e(n,t)$;

$$c = n * m$$

$$MCG(c,i) = \sum_{i=0}^{L-1} s_{mu}(c,i) * s_{mu}(c,i)$$

$$MCG1(c,i) = log10\Big(\sum_{i=0}^{L-1} s_{mu}(c,i) * s_{mu}(c,i)\Big)$$

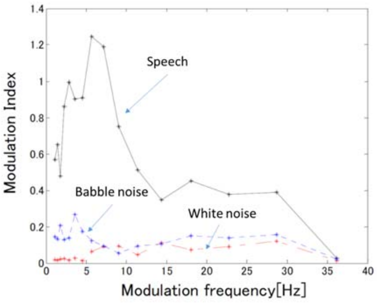
$$MCG3(c,i) = \Big(\sum_{\mathrm{n}=c-2}^{c+2} \sum_{j=i-2}^{i+2} MCG1(c,i)\Big)/(5*5)$$

$$MCG4(c,i) = \Big(\sum_{k=c-5}^{c+5} \sum_{j=i-5}^{i+5} MCG1(c,i)\Big)/(11*11)$$

MCG2 is similar MCG1, but has longer window

Modulation spectrum in speech

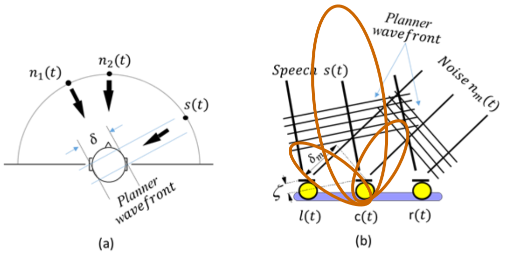
- Time-frequency correlations are given by the modulation spectra. The similarity of the modulation spectra across phonemes as assessed by the spectral correlation index (SCI) was a strong predictor of the confusions made by human listeners. (Gallun, at al. 2008)
- A sparse set of time-averaged patterns of modulation energy can capture a meaningful aspect of the information listeners use to distinguish among speech signals.



Kimhuoch, P., ARAI, T., et al. “Voice activity detection by using modulation spectrum in noise : Investigation on speech frequency band and modulation frequency band”, ASJ meeting, 2009,9

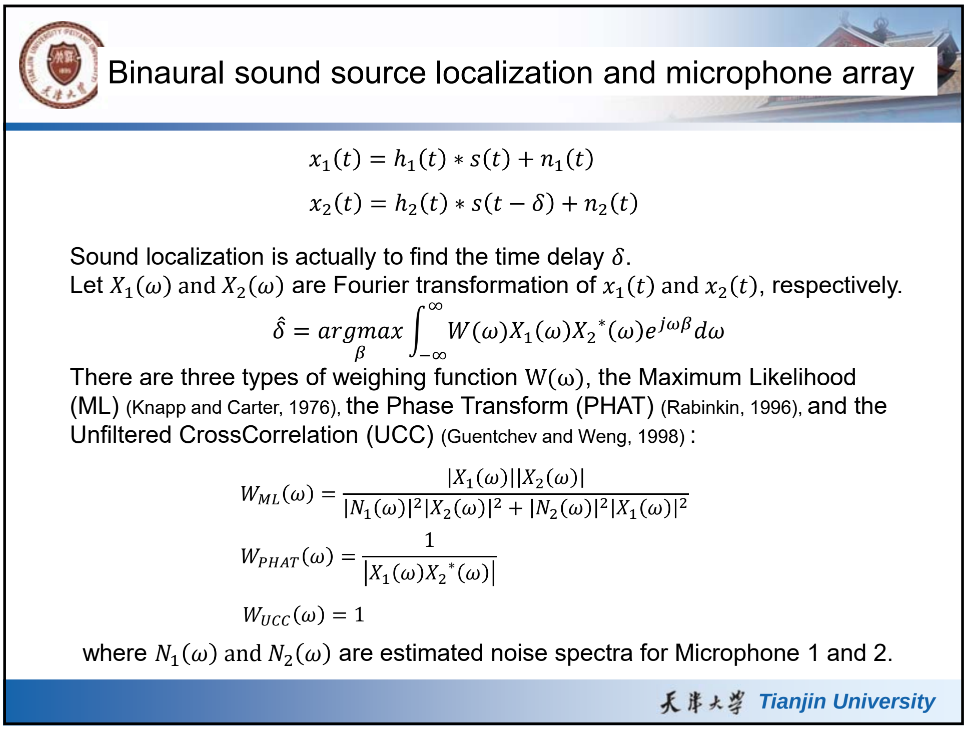
Binaural sound source localization and microphone array

- Sound localization is a listener's ability to identify the location or origin of a detected sound in direction and distance, which is for defensing the enemy.
- Interaural Time Difference (ITD) Sound from the right side reaches the right ear earlier than the left ear. It works for low frequency <800Hz.
- Interaural Intensity Difference (IID) or Interaural Level Difference (ILD) Sound from the right side has a higher level at the right ear than at the left ear, because the head shadows the left ear. These level differences are highly frequency dependent and they increase with increasing frequency. It works for high frequencies >1600Hz.



$$x_1(t) = h_1(t) * s(t) + n_1(t)$$

$$x_2(t) = h_2(t) * s(t - \delta) + n_2(t)$$



Binaural sound source localization and microphone array

$$x_1(t) = h_1(t) * s(t) + n_1(t)$$

$$x_2(t) = h_2(t) * s(t-\delta) + n_2(t)$$

Sound localization is actually to find the time delay δ.
Let $X_1(\omega)$ and $X_2(\omega)$ are Fourier transformation of $x_1(t)$ and $x_2(t)$, respectively.

$$\hat{\delta} = \underset{\beta}{argmax} \int_{-\infty}^{\infty} W(\omega) X_1(\omega) {X_2}^*(\omega) e^{j\omega\beta} d\omega$$

There are three types of weighing function W(ω), the Maximum Likelihood (ML) (Knapp and Carter, 1976), the Phase Transform (PHAT) (Rabinkin, 1996), and the Unfiltered CrossCorrelation (UCC) (Guentchev and Weng, 1998) :

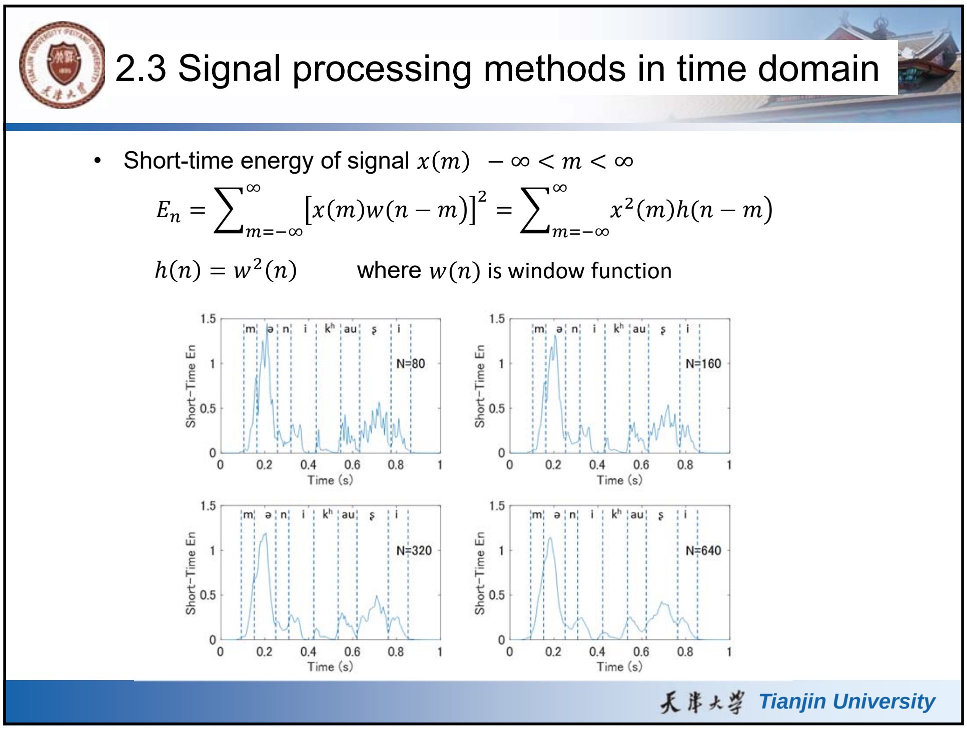
$$W_{ML}(\omega) = \frac{|X_1(\omega)||X_2(\omega)|}{|N_1(\omega)|^2|X_2(\omega)|^2 + |N_2(\omega)|^2|X_1(\omega)|^2}$$

$$W_{PHAT}(\omega) = \frac{1}{\left|X_1(\omega){X_2}^*(\omega)\right|}$$

$$W_{UCC}(\omega) = 1$$

where $N_1(\omega)$ and $N_2(\omega)$ are estimated noise spectra for Microphone 1 and 2.

天津大学 Tianjin University



2.3 Signal processing methods in time domain

- Short-time energy of signal $x(m)$ $\quad -\infty < m < \infty$

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(m)w(n-m)\right]^2 = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m)$$

$h(n) = w^2(n)$ where $w(n)$ is window function

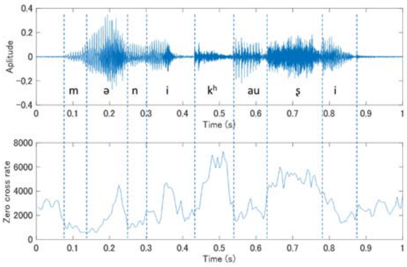
天津大学 Tianjin University

2.3 Signal processing methods in time domain

- Short-time mean zero crossing rate (ZCR) shows spectral properties with frequency transformation

$$Z = \frac{f_s}{N}\sum_{n=0}^{N-1}\left|sign[x(n)] - sign[x(n+1)]\right|$$

Where $sign[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases}$



ZCR reflects the mean frequency of the region including the major energy:

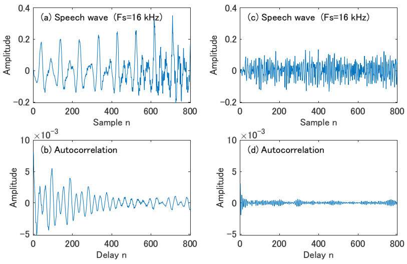
- Voiced sound: about 1400Hz
- Unvoiced sound: about 4500Hz

2.3 Signal processing methods in time domain

- Short-time autocorrelation shows spectral properties with frequency transformation

$$R(i) = \sum_{n=-\infty}^{\infty} x(n)x(n-i), \qquad i = \{0,1,2,\ldots\}$$

$$R(m) = \frac{1}{2N+1}\sum_{n=-N+1}^{N-1} x(n)x(n+|m|),(|m| = 0,1,\cdots,N-1)$$



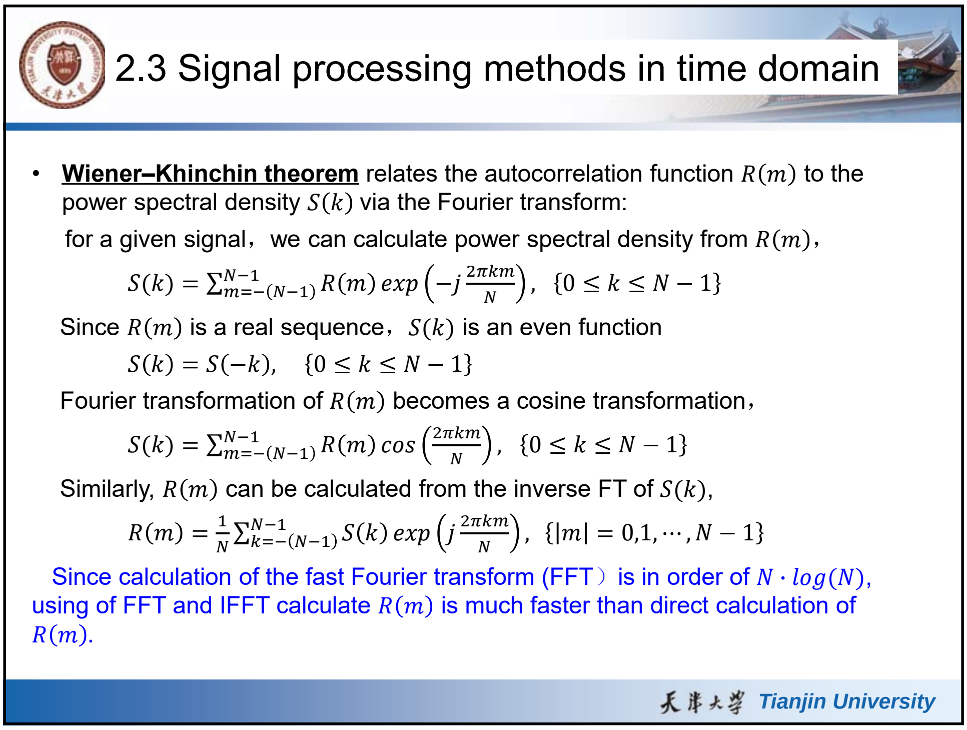
Properties:

- Symmetry property

  $R(m)$=$R(-m)$
- Maximum at zero

  $|R(m)| \leq R(0)$
- *Autocorrelation of white noise*

  The autocorrelation of a continuous-time white noise signal will have a strong peak at $i = 0$ and will be exactly 0 for all other $i \neq 0$.



2.3 Signal processing methods in time domain

- **<u>Wiener–Khinchin theorem</u>** relates the autocorrelation function $R(m)$ to the power spectral density $S(k)$ via the Fourier transform:

for a given signal，we can calculate power spectral density from $R(m)$，

$$S(k) = \sum_{m=-(N-1)}^{N-1} R(m)\, exp\left(-j\frac{2\pi km}{N}\right),\ \ \{0 \le k \le N-1\}$$

Since $R(m)$ is a real sequence，$S(k)$ is an even function

$$S(k) = S(-k),\quad \{0 \le k \le N-1\}$$

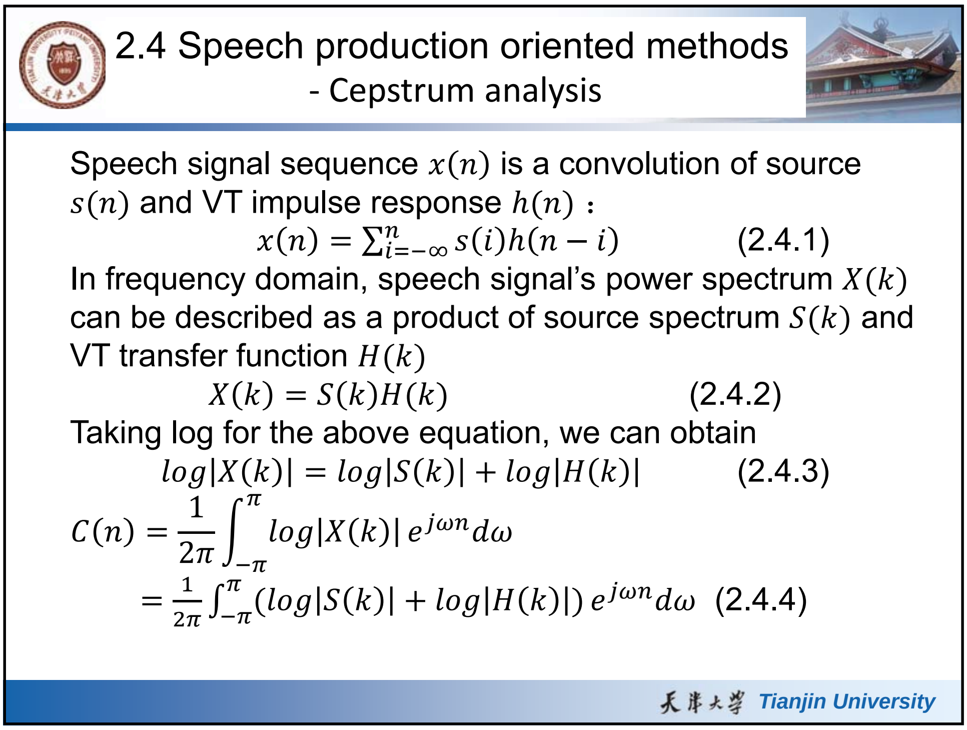
Fourier transformation of $R(m)$ becomes a cosine transformation，

$$S(k) = \sum_{m=-(N-1)}^{N-1} R(m)\, cos\left(\frac{2\pi km}{N}\right),\ \ \{0 \le k \le N-1\}$$

Similarly, $R(m)$ can be calculated from the inverse FT of $S(k)$,

$$R(m) = \frac{1}{N}\sum_{k=-(N-1)}^{N-1} S(k)\, exp\left(j\frac{2\pi km}{N}\right),\ \ \{|m| = 0,1,\cdots,N-1\}$$

Since calculation of the fast Fourier transform (FFT）is in order of $N \cdot log(N)$, using of FFT and IFFT calculate $R(m)$ is much faster than direct calculation of $R(m)$.



2.4 Speech production oriented methods - Cepstrum analysis

Speech signal sequence $x(n)$ is a convolution of source $s(n)$ and VT impulse response $h(n)$ ：

$$x(n) = \sum_{i=-\infty}^{n} s(i)h(n-i) \qquad (2.4.1)$$

In frequency domain, speech signal's power spectrum $X(k)$ can be described as a product of source spectrum $S(k)$ and VT transfer function $H(k)$

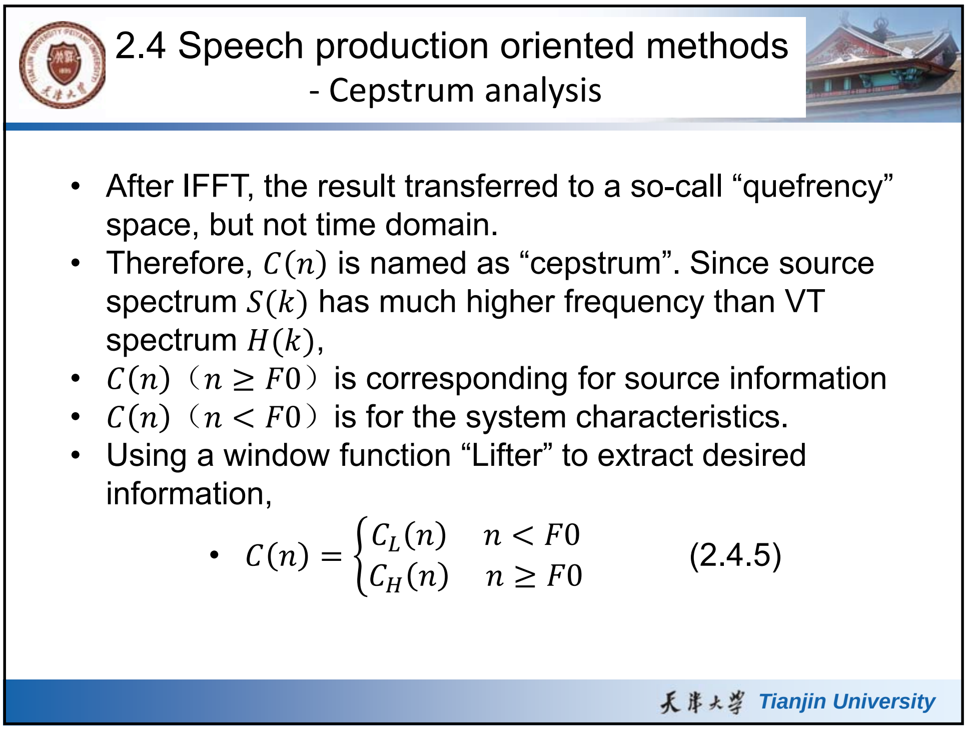
$$X(k) = S(k)H(k) \qquad (2.4.2)$$

Taking log for the above equation, we can obtain

$$log|X(k)| = log|S(k)| + log|H(k)| \qquad (2.4.3)$$

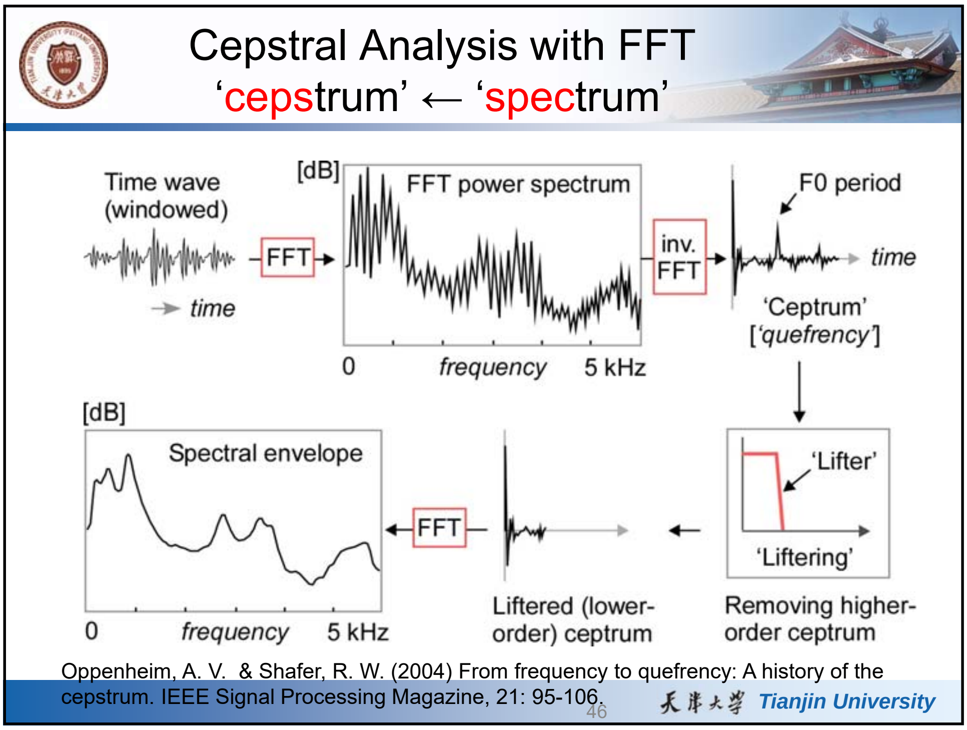
$$C(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} log|X(k)|\, e^{j\omega n} d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} (log|S(k)| + log|H(k)|)\, e^{j\omega n} d\omega \quad (2.4.4)$$



2.4 Speech production oriented methods - Cepstrum analysis

- After IFFT, the result transferred to a so-call “quefrency” space, but not time domain.
- Therefore, $C(n)$ is named as “cepstrum”. Since source spectrum $S(k)$ has much higher frequency than VT spectrum $H(k)$,
- $C(n)$（$n \geq F0$）is corresponding for source information
- $C(n)$（$n < F0$）is for the system characteristics.
- Using a window function “Lifter” to extract desired information,
  - $$C(n) = \begin{cases} C_L(n) & n < F0 \\ C_H(n) & n \geq F0 \end{cases} \qquad (2.4.5)$$



Cepstral Analysis with FFT ‘cepstrum’ ← ‘spectrum’

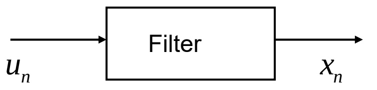
Oppenheim, A. V. & Shafer, R. W. (2004) From frequency to quefrency: A history of the cepstrum. IEEE Signal Processing Magazine, 21: 95-106.

Problems of Cepstral Analysis
on speech signals

- The cepstral analysis is merely smoothing of FFT power spectrum using a moving average technique.
- Frequency resolution of cepstral analysis depends on the population of harmonics. The analysis fails at high-pitched voice due to insufficient data of the cepstrum in the time (quefrency) domain.
- In female voices, the first and second formants (F1, F2) tend to be overlapped to make a single peak in the analysis of back vowels /a/ and /o/.

2.4 Speech production oriented methods --System Overview

- Input/output



Observation is difficult Observable

- Assume impulse or white-noise, and minimum energy
- Auto regressive & Moving average (Pole-Zero model)

$$x_n = -\underset{\boxed{\text{IIR}}}{\sum_{k=1}^{p} a_k x_{n-k}} + G \underset{\boxed{\text{FIR}}}{\sum_{l=0}^{q} b_l u_{n-l}}, \quad b_0 = 1$$

- Transfer function

$$H(Z) = \frac{X(Z)}{U(Z)} = G\frac{1+\sum_{l=1}^{q} b_l Z^{-l}}{1+\sum_{k=1}^{p} a_k Z^{-k}},$$
$$X(Z) = Z\{x_n\}, U(Z) = Z\{u_n\}$$

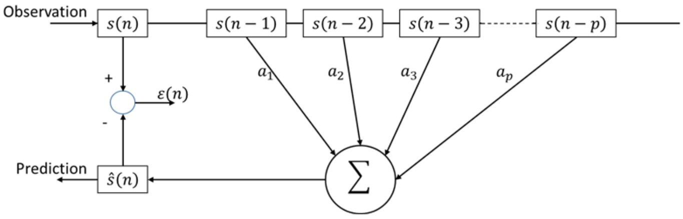
2.4 Speech production oriented methods - Linear predictive coding (LPC)

LPC is a method where the coming value $s(n)$ of a signal is estimated as a linear function of previous samples, e.g. $s(n-i), \{1,2,\dots,p\}$,

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \qquad (2.4.6)$$

$$\varepsilon(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \qquad (2.4.7)$$

Where $\varepsilon(n)$ is the difference between $s(n)$ and predicted $\hat{s}(n)$, $a_i\ (i = 1, 2, \dots, \mathrm{p})$ are the weighting coefficients.



Linear predictive coding (LPC)

Optimal a_i is obtained by minimizing the root mean square of $E[\varepsilon^2(n)]$,

$$E[\varepsilon^2(n)] = E\left[\left[s(n) - \sum_{i=1}^{p} a_i s(n-i)\right]^2\right] \qquad (2.4.8)$$

Let $$\frac{\partial E[\varepsilon^2(n)]}{\partial a_j} = 0, \ \{1 \le j \le p\} \qquad (2.4.9)$$

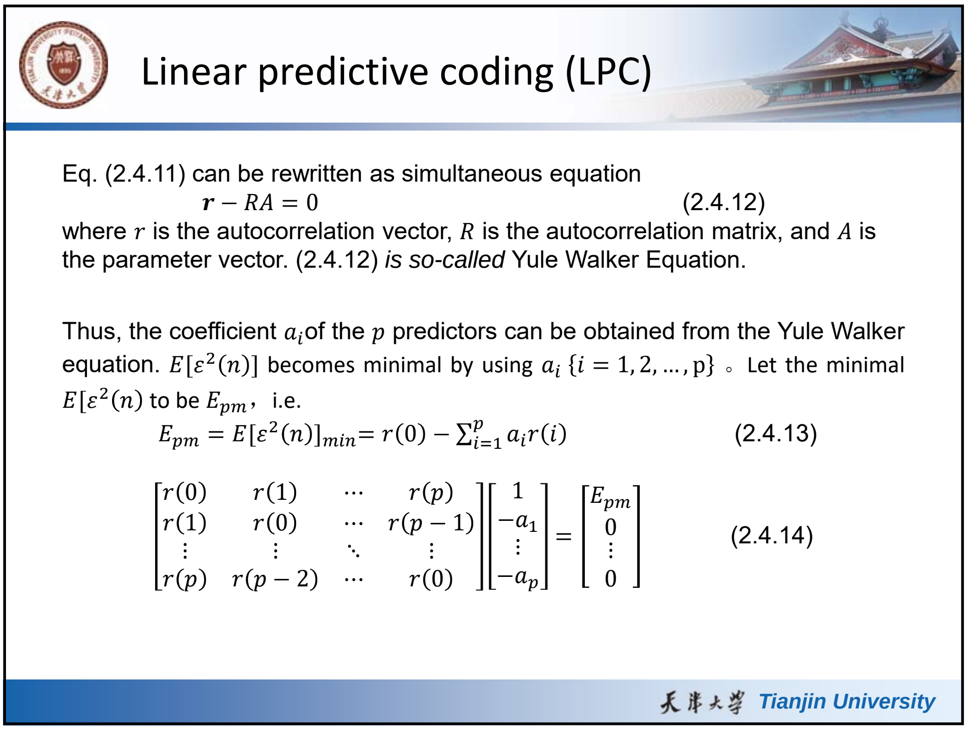
$$\frac{\partial E[\varepsilon^2(n)]}{\partial a_j} = -2E\big[\varepsilon(n)s(n-j)\big] = 0 \qquad (2.4.10)$$

$$\begin{aligned} &E\Big[s(n)s(n-j) - \sum_{i=1}^{p} a_i s(n-i)\, s(n-j)\Big] \\ &= r(j) - \sum_{i=1}^{p} a_i r(j-i) = 0, \qquad \{1 \le j \le p\} \end{aligned} \qquad (2.4.11)$$

where $r(j) = E\big[s(n)s(n-j)\big]$ is autocorrelation of $s(n)$.

Let $\boldsymbol{r} = \big[r(1), r(2), \dots r(p)\big]^{\boldsymbol{T}}$, $\boldsymbol{A} = \big[a_1, a_2, \dots a_p\big]^{\boldsymbol{T}}$,*and*

$$R = \begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(0) & \cdots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix}$$



Linear predictive coding (LPC)

Eq. (2.4.11) can be rewritten as simultaneous equation

$$\boldsymbol{r} - RA = 0 \qquad (2.4.12)$$

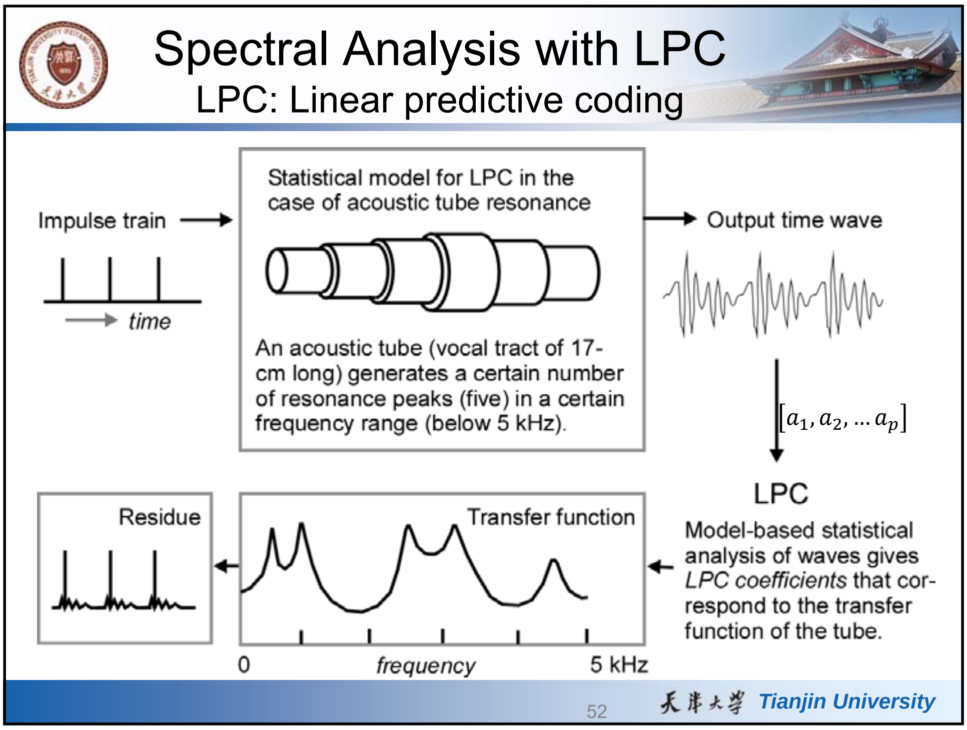
where r is the autocorrelation vector, R is the autocorrelation matrix, and A is the parameter vector. (2.4.12) *is so-called* Yule Walker Equation.

Thus, the coefficient $a_i$of the p predictors can be obtained from the Yule Walker equation. $E[\varepsilon^2(n)]$ becomes minimal by using a_i $\{i = 1, 2, \ldots, \mathrm{p}\}$ 。Let the minimal $E[\varepsilon^2(n)$ to be E_{pm}， i.e.

$$E_{pm} = E[\varepsilon^2(n)]_{min} = r(0) - \sum_{i=1}^{p} a_i r(i) \qquad (2.4.13)$$

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(p) \\ r(1) & r(0) & \cdots & r(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ r(p) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ \vdots \\ -a_p \end{bmatrix} = \begin{bmatrix} E_{pm} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (2.4.14)$$

天津大学 Tianjin University



Spectral Analysis with LPC

LPC: Linear predictive coding

Impulse train

time

Statistical model for LPC in the case of acoustic tube resonance

An acoustic tube (vocal tract of 17-cm long) generates a certain number of resonance peaks (five) in a certain frequency range (below 5 kHz).

Output time wave

$[a_1, a_2, \ldots a_p]$

LPC

Model-based statistical analysis of waves gives *LPC coefficients* that correspond to the transfer function of the tube.

Residue

Transfer function

0 *frequency* 5 kHz

52

天津大学 Tianjin University

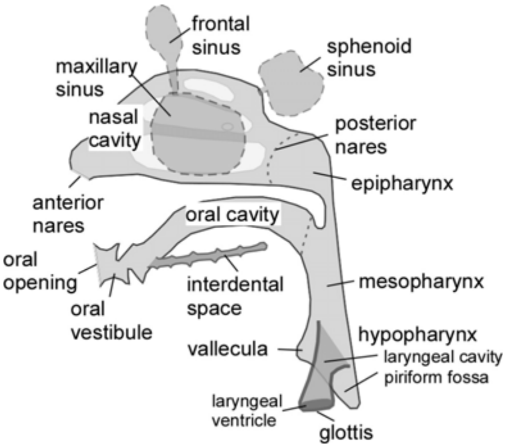
Problems of LPC

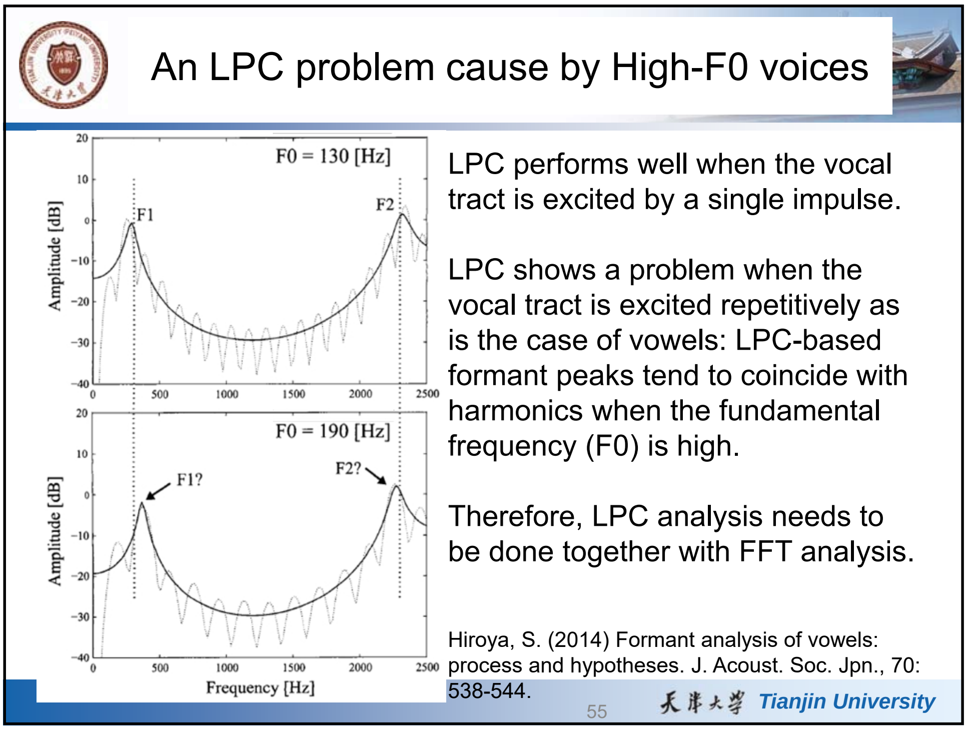
for speech analysis

- LPC follows an assumption of all-pole transfer model (single-tube model of a vocal tract).
- This is obsolete. Vocal tracts are no longer uniaxial. Side branches produce spectral zeros, which are ignored by LPC.
- *Autocorrelation LPC* fails at analyzing transient phenomena in speech signals, such as consonant-vowel junctions. *Covariance LPC* is unstable or requires extra procedures.
- LPC is best performed on speech signals sampled at 10 kHz to analyze 0 – 5 kHz frequency band to extract five formants. Down-sampling is necessary.

Simplicity of LPC vs. Complexity of the VT

Complex shape of the vocal and nasal tracts





An LPC problem cause by High-F0 voices

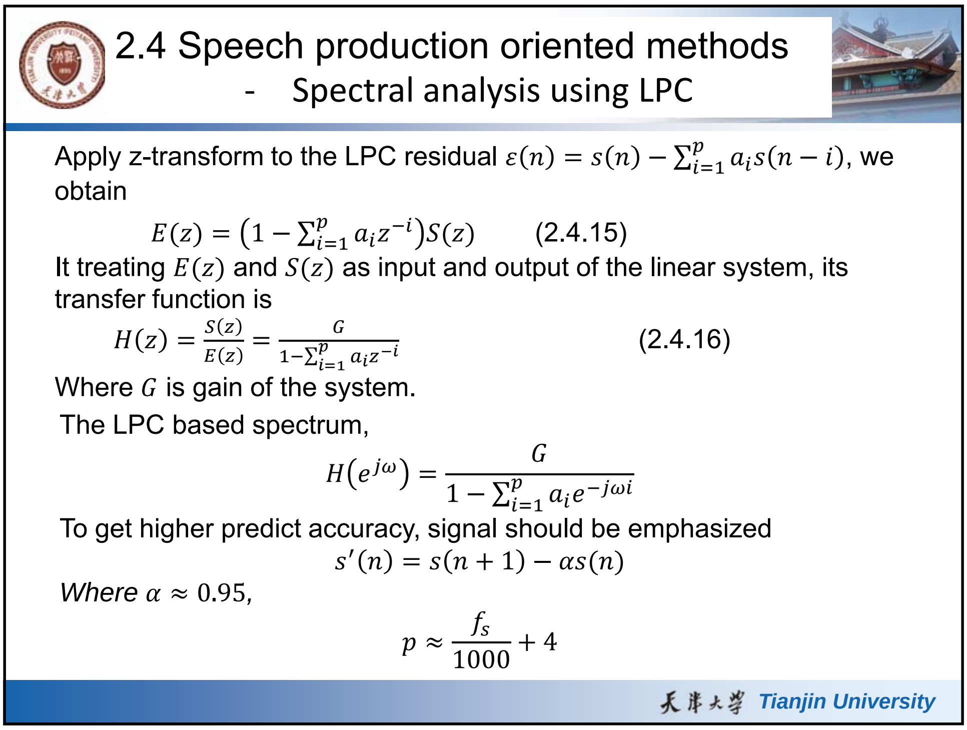
LPC performs well when the vocal tract is excited by a single impulse.

LPC shows a problem when the vocal tract is excited repetitively as is the case of vowels: LPC-based formant peaks tend to coincide with harmonics when the fundamental frequency (F0) is high.

Therefore, LPC analysis needs to be done together with FFT analysis.

Hiroya, S. (2014) Formant analysis of vowels: process and hypotheses. J. Acoust. Soc. Jpn., 70: 538-544.

55 天津大学 Tianjin University



2.4 Speech production oriented methods - Spectral analysis using LPC

Apply z-transform to the LPC residual $\varepsilon(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i)$, we obtain

$$E(z) = \left(1 - \sum_{i=1}^{p} a_i z^{-i}\right) S(z) \qquad (2.4.15)$$

It treating $E(z)$ and $S(z)$ as input and output of the linear system, its transfer function is

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \qquad (2.4.16)$$

Where G is gain of the system.

The LPC based spectrum,

$$H(e^{j\omega}) = \frac{G}{1 - \sum_{i=1}^{p} a_i e^{-j\omega i}}$$

To get higher predict accuracy, signal should be emphasized

$$s'(n) = s(n+1) - \alpha s(n)$$

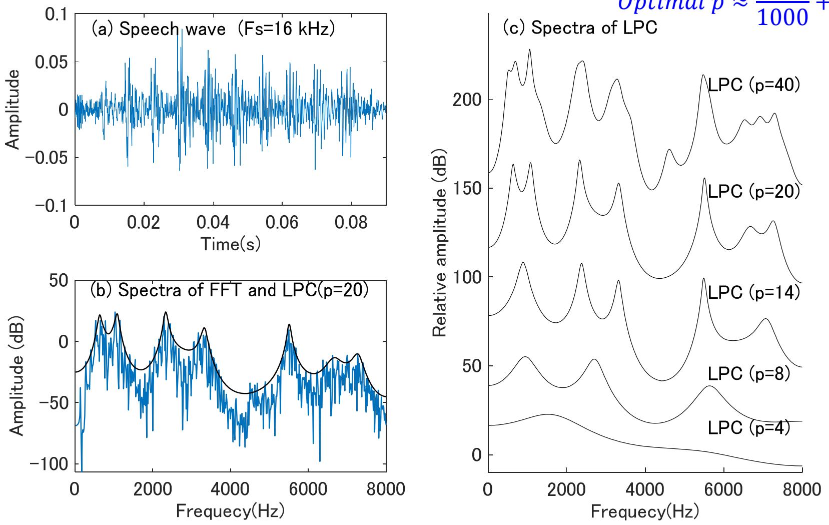
Where $\alpha \approx 0.95$,

$$p \approx \frac{f_s}{1000} + 4$$

天津大学 Tianjin University

2.4 Speech production oriented methods - Spectral analysis using LPC

$Optimal\ p \approx \frac{f_s}{1000} + 4$



2.4 Speech production oriented methods - LPC cepstrum

From LPC based transfer function, we can get z-transformation of system impulse response

$$H(z) = \frac{G}{1-\sum_{i=1}^{p} a_i z^{-i}} = \sum_{n=0}^{\infty} h(n) z^{-n} \qquad (2.4.17)$$

$$h(n) = \sum_{i=1}^{p} a_i h(n-i) + G\delta(n), \quad n \geq 0 \qquad (2.4.18)$$

Then, we can prove the cepstrum $C(n)$ can be calculated using the following recursive formula

$$C(n) = \begin{cases} 0 & (n < 0) \\ log(G), & (n = 0) \\ a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) C(k) a_{n-k}, & (1 \leq n \leq p) \\ \sum_{k=n-p}^{n-1} \left(\frac{k}{n}\right) C(k) a_{n-k}, & (n > p) \end{cases} \qquad (2.4.19)$$

Durbin’s method

- Forward and backward estimation

$$A^{(i)}(Z) = 1 + \sum_{j=1}^{i} a_j^{(i)} Z^{-j} = 1 + \sum_{j=1}^{i-1} \left[a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}\right] Z^{-j} + a_i^{(i)} Z^{-i}$$

$$= 1 + \sum_{j=1}^{i-1} a_j^{(i-1)} Z^{-j} - k_i \sum_{j=1}^{i-1} a_{i-j}^{(i-1)} Z^{-j} - k_i Z^{-i}$$

$$= A^{(i-1)}(Z) - k_i \left[Z^{-i} + \sum_{j=1}^{i-1} a_{i-j}^{(i-1)} Z^{-j}\right]$$

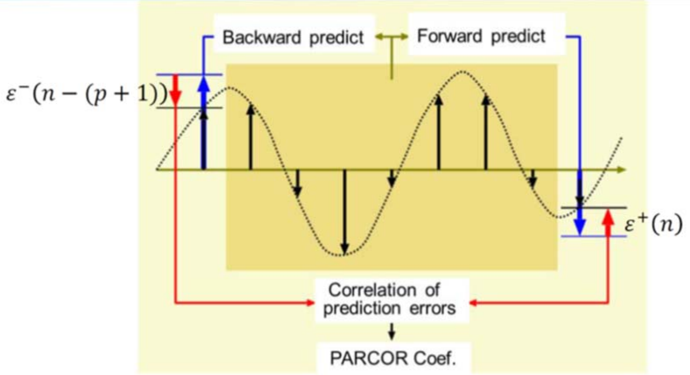
$$= A^{(i-1)}(Z) - k_i B^{(i-1)}(Z) \;\; b_j^{(i)} = a_{i-j}^{(i)}$$

- PARCOR (PARtial CORrelation) parameter

$$k_i = \frac{\sum_{m=0}^{N-1} e_{f,m}^{(i-1)} e_{b,m-1}^{(i-1)}}{\sqrt{\sum_{m=0}^{N-1} \left[e_{f,m}^{(i-1)}\right]^2} \sqrt{\sum_{m=0}^{N-1} \left[e_{b,m-1}^{(i-1)}\right]^2}},$$

Where $e_{f,m}^{(i-1)} = x_m + \sum_{j=1}^{i-1} a_j^{(i-1)} x_{m-j}$, $e_{b,m-1}^{(i-1)} = x_{m-i} + \sum_{j=1}^{i-1} b_j^{(i-1)} x_{m-j}$

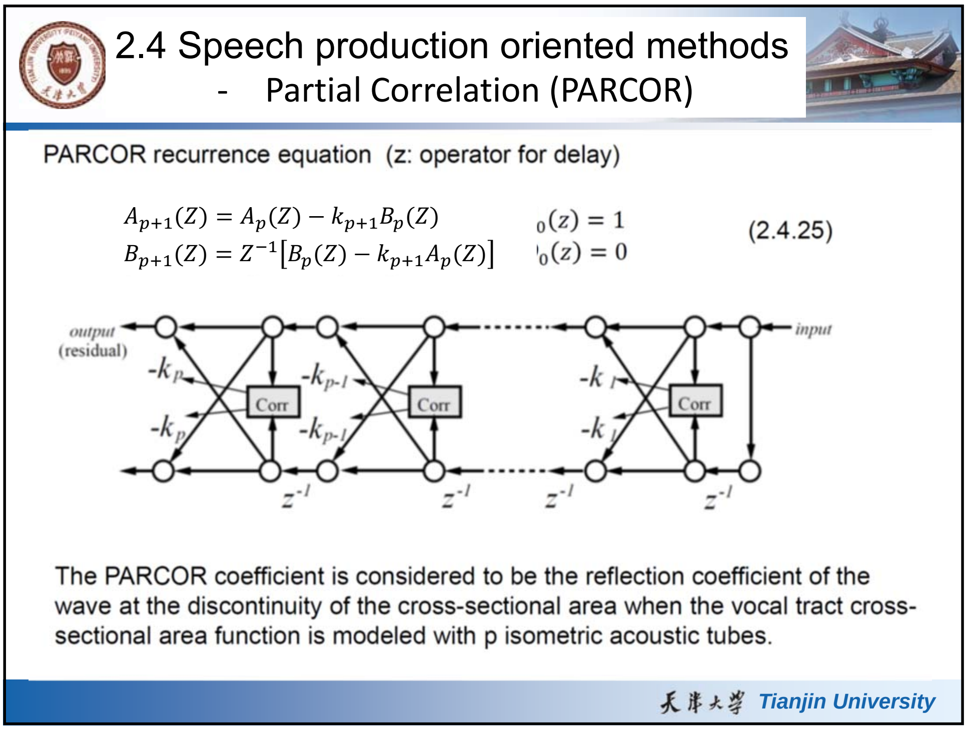
2.4 Speech production oriented methods - Partial Correlation (PARCOR)



$$\varepsilon^+(n) = s(n) - \hat{s}^+(n) = \sum_{i=0}^{p} a_i s(n-i) \Longrightarrow A_p(z) \cdot s(z)$$

$$\varepsilon^-(n-(p+1)) = s(n-(p+1)) - \hat{s}^-(n-(p+1)) = \sum_{i=1}^{p+1} b_i s(n-i) \Longrightarrow B_p(z) \cdot s(z)$$

$$k_{p+1} = \frac{(\varepsilon^+(n), \varepsilon^-(n-(p+1)))}{\|\varepsilon^+(n)\| \|\varepsilon^-(n-(p+1))\|} \qquad -1 < k_i < 1 \qquad (2.4.24)$$



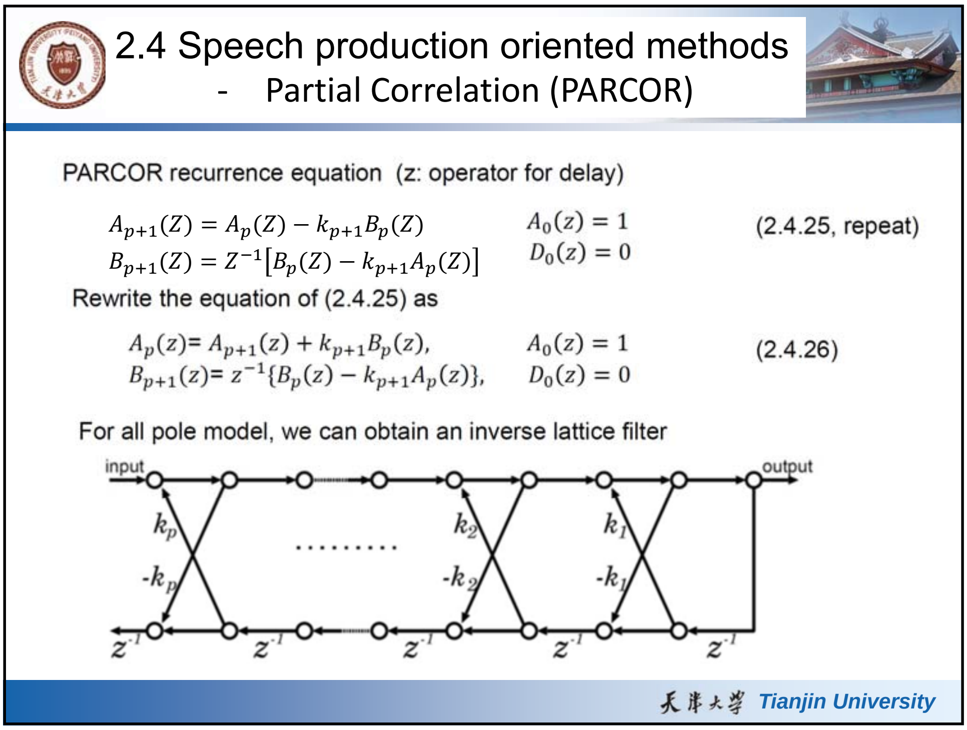
2.4 Speech production oriented methods - Partial Correlation (PARCOR)

PARCOR recurrence equation (z: operator for delay)

$$A_{p+1}(Z) = A_p(Z) - k_{p+1}B_p(Z) \qquad _0(z) = 1$$
$$B_{p+1}(Z) = Z^{-1}[B_p(Z) - k_{p+1}A_p(Z)] \qquad '_0(z) = 0 \qquad (2.4.25)$$

The PARCOR coefficient is considered to be the reflection coefficient of the wave at the discontinuity of the cross-sectional area when the vocal tract cross-sectional area function is modeled with p isometric acoustic tubes.

天津大学 Tianjin University



2.4 Speech production oriented methods - Partial Correlation (PARCOR)

PARCOR recurrence equation (z: operator for delay)

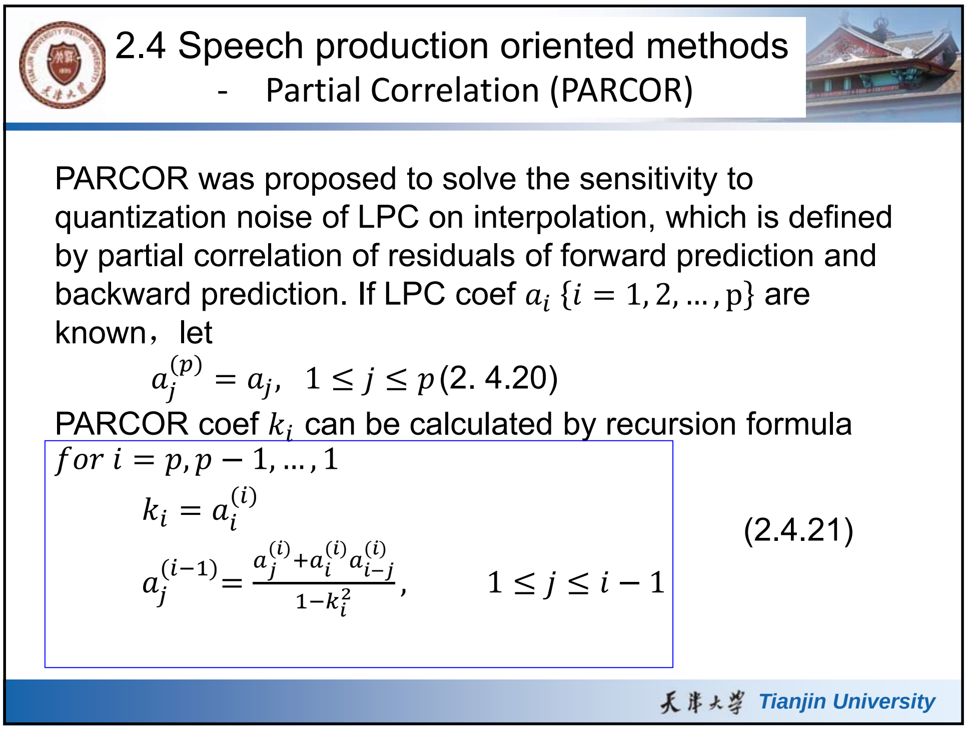
$$A_{p+1}(Z) = A_p(Z) - k_{p+1}B_p(Z) \qquad A_0(z) = 1$$
$$B_{p+1}(Z) = Z^{-1}[B_p(Z) - k_{p+1}A_p(Z)] \qquad D_0(z) = 0 \qquad (2.4.25, repeat)$$

Rewrite the equation of (2.4.25) as

$$A_p(z) = A_{p+1}(z) + k_{p+1}B_p(z), \qquad A_0(z) = 1$$
$$B_{p+1}(z) = z^{-1}\{B_p(z) - k_{p+1}A_p(z)\}, \qquad D_0(z) = 0 \qquad (2.4.26)$$

For all pole model, we can obtain an inverse lattice filter

天津大学 Tianjin University



2.4 Speech production oriented methods
- Partial Correlation (PARCOR)

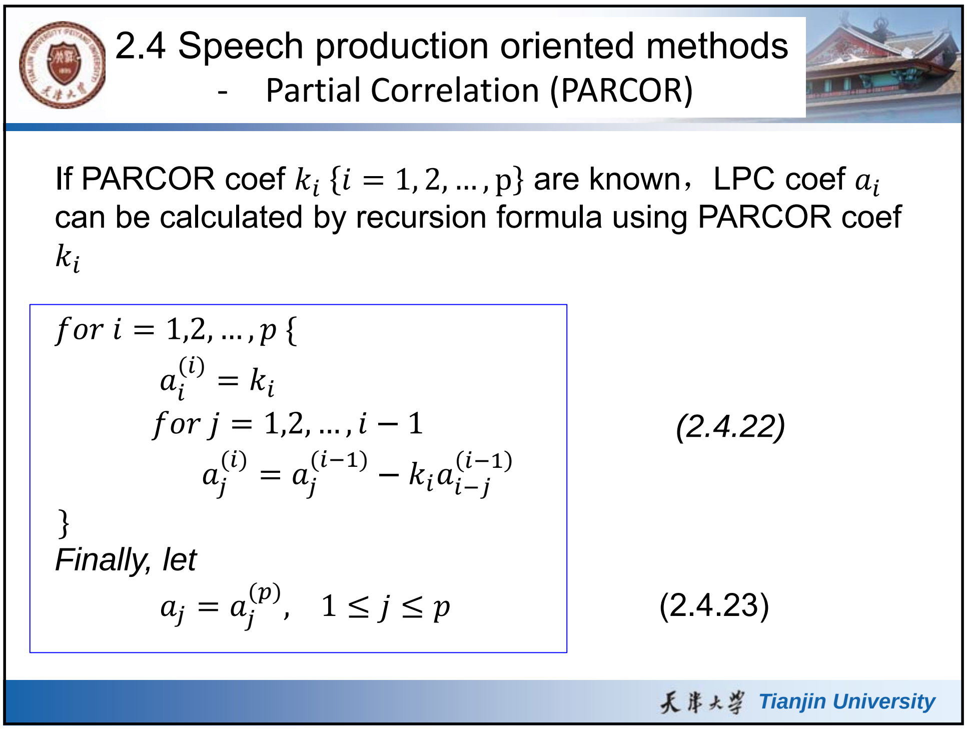
PARCOR was proposed to solve the sensitivity to quantization noise of LPC on interpolation, which is defined by partial correlation of residuals of forward prediction and backward prediction. If LPC coef a_i $\{i = 1, 2, \dots, \mathrm{p}\}$ are known，let

$a_j^{(p)} = a_j,\ \ 1 \le j \le p$(2. 4.20)

PARCOR coef k_i can be calculated by recursion formula

```
$for\ i = p, p-1, \dots, 1$
        $k_i = a_i^{(i)}$
        $a_j^{(i-1)} = \frac{a_j^{(i)} + a_i^{(i)} a_{i-j}^{(i)}}{1-k_i^2},\qquad 1 \le j \le i-1$
```

(2.4.21)



2.4 Speech production oriented methods
- Partial Correlation (PARCOR)

If PARCOR coef k_i $\{i = 1, 2, \dots, \mathrm{p}\}$ are known，LPC coef a_i can be calculated by recursion formula using PARCOR coef k_i

```
$for\ i = 1,2, \dots, p$ {
        $a_i^{(i)} = k_i$
        $for\ j = 1,2, \dots, i-1$
            $a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}$
}
```

(2.4.22)

Finally, let

$$a_j = a_j^{(p)},\quad 1 \le j \le p$$

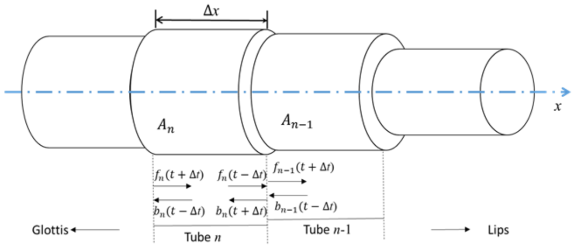
(2.4.23)

Approximation of vocal tract by cascaded tubes （Review）

$f(x,t)$: wave propagate to open-end
$b(x,t)$: wave propagate to close-end
At any point x in the tube:

$$u(x,t) = f(t - x/c) - b(t + x/c)$$

$$p(x,t) = \frac{\rho c}{A}[f(t - x/c) + b(t + x/c)]$$



If $u_n(t)$ and $p_n(t)$ represent the volume velocity and pressure of tube n, Δt is the time of sound from boundary to $\Delta x/2$, then

$$u_n(t) = f_n(t - \Delta t) - b_n(t + \Delta t)$$

$$p_n(t) = \frac{\rho c}{A_n}[f_n(t - \Delta t) + b_n(t + \Delta t)]$$

Approximation of vocal tract by cascaded tubes （review）

Considering the continuity of the pressure and volume velocity at the boundary:

$$p_n(\Delta x/2, t) = p_{n-1}(-\Delta x/2, t)$$

$$u_n(\Delta x/2, t) = u_{n-1}(-\Delta x/2, t)$$

We can obtain recursive equations as

$$f_{n-1}(t + \Delta t) = k_n b_{n-1}(t - \Delta t) + (1 + k_n) f_n(t - \Delta t)$$

$$b_n(t + \Delta t) = (1 - k_n) b_{n-1}(t - \Delta t) - k_n f_n(t - \Delta t)$$

k_n: reflection coefficient

$$k_n = \frac{A_{n-1} - A_n}{A_{n-1} + A_n}$$

PARCOR parameter = reflection coefficient

2.4 Speech production oriented methods
- Line Spectrum Pair (LSP， LSF)

LSP or line spectral frequencies (LSF) are used to represent linear prediction coefficients (LPC) for transmission over a channel. LSPs have several properties (e.g. smaller sensitivity to quantization noise).

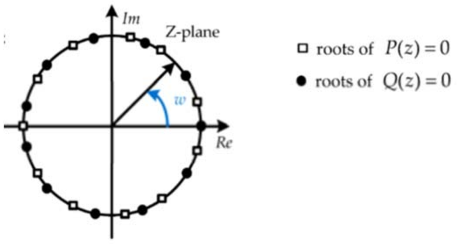
$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} = \frac{1}{2}[P(z) + Q(z)] \quad (2.\,4.24)$$

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1})$$
$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \quad (2.\,4.25)$$

By construction, P is a symmetric polynomial and Q an antisymmetric polynomial; physically

- P(z) corresponds to the vocal tract with the glottis closed and
- Q(z) with the glottis open, i.e.,

$P(z)\big|_{z=-1} = 0,\ Q(z)\big|_{z=1} = 0.$

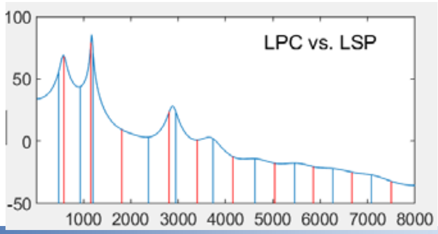


2.4 Speech production oriented methods
- Line Spectrum Pair (LSP， LSF)

- Roots of P (ω_i) and Q (θ_i) lie on the unit circle in the complex plane, and alternate with each other, $0 < \omega_1 < \theta_1 < \omega_2 < \theta_2 < \cdots < \omega_{\frac{p}{2}} < \theta_{\frac{p}{2}} < \pi$
- As the coefficients of P and Q are real, the roots occur in conjugate pairs, while a pair reflects a resonance peak.

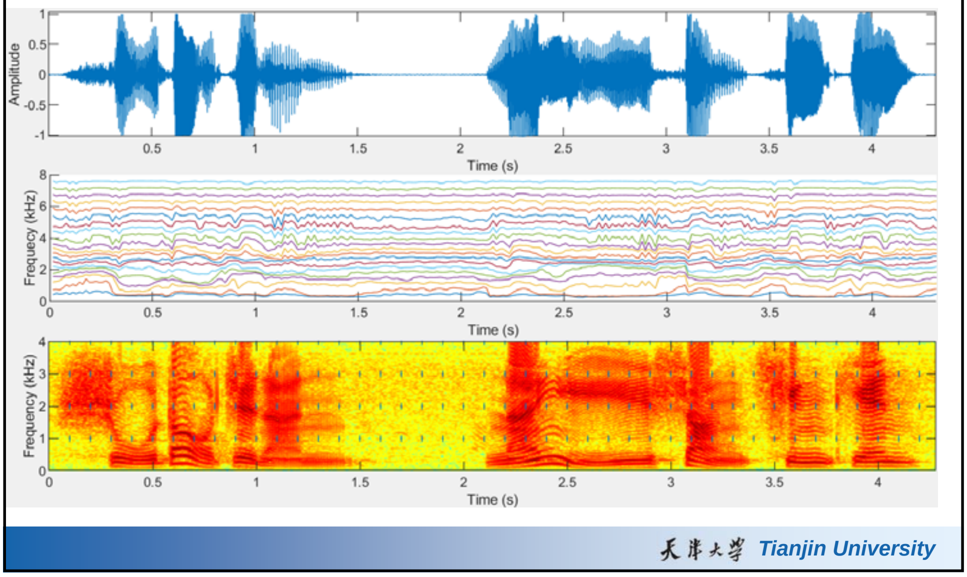
$$\left|H(e^{j\omega})\right| = G/\left|A(e^{j\omega})\right| = 2G/\left|P(e^{j\omega}) + Q(e^{j\omega})\right|$$
$$= \frac{2^{(1-p)/2}G}{\sin^2(\frac{\omega}{2})\prod_{i=1}^{\frac{p}{2}}(\cos\omega - \cos\theta_i)^2 + \cos^2(\frac{\omega}{2})\prod_{i=1}^{p/2}(\cos\omega - \cos\omega_i)^2}$$

(2. 4.26)



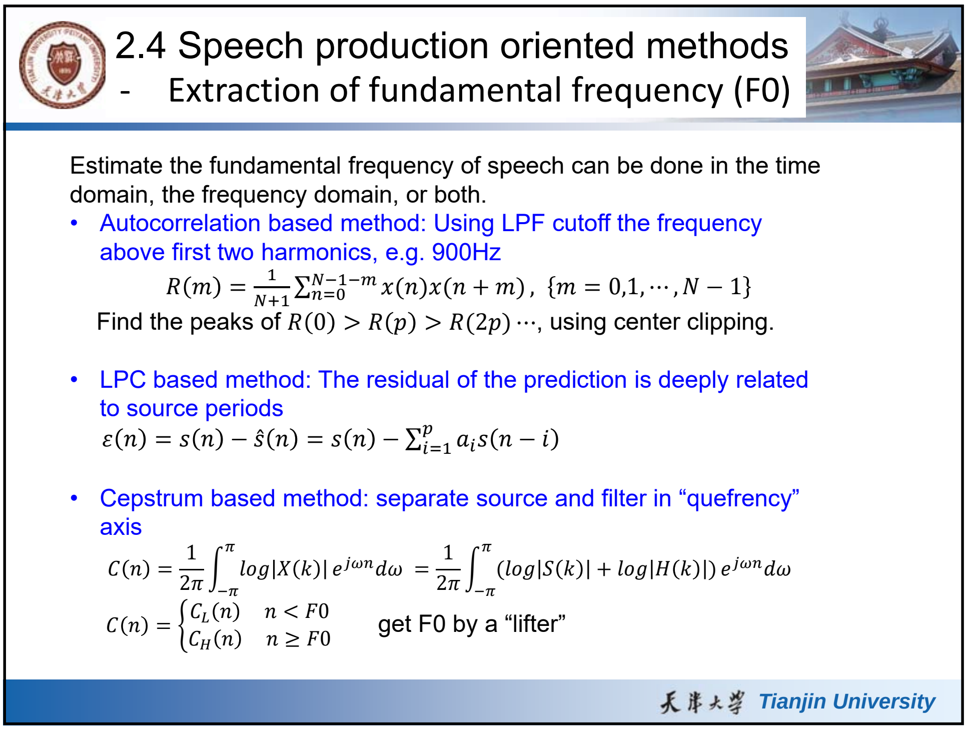
2.4 Speech production oriented methods

- Line Spectrum Pair (LSP)



Summary

- LPC to PARCOR or LSF
  - LPC:
    - Sensitive to quantization -> unstable
    - Difficult to interpolation -> short frame shift (about 10 ms)
  - PARCOR:
    - Stable when $|k_i| < 1$
    - Difficult to interpolation -> short frame shift (about 10 ms)
  - LSF:
    - Robust to quantization $0 \le \omega_{q,0} < \omega_{p,0} < \omega_{q,1} < \omega_{p,1} < \ldots.. \le \pi$
    - Easy to interpolate -> frame shift (about 15 - 20 ms)



2.4 Speech production oriented methods - Extraction of fundamental frequency (F0)

Estimate the fundamental frequency of speech can be done in the time domain, the frequency domain, or both.

- Autocorrelation based method: Using LPF cutoff the frequency above first two harmonics, e.g. 900Hz

$$R(m) = \frac{1}{N+1}\sum_{n=0}^{N-1-m} x(n)x(n+m),\ \{m = 0,1,\cdots,N-1\}$$

Find the peaks of $R(0) > R(p) > R(2p)\cdots$, using center clipping.

- LPC based method: The residual of the prediction is deeply related to source periods

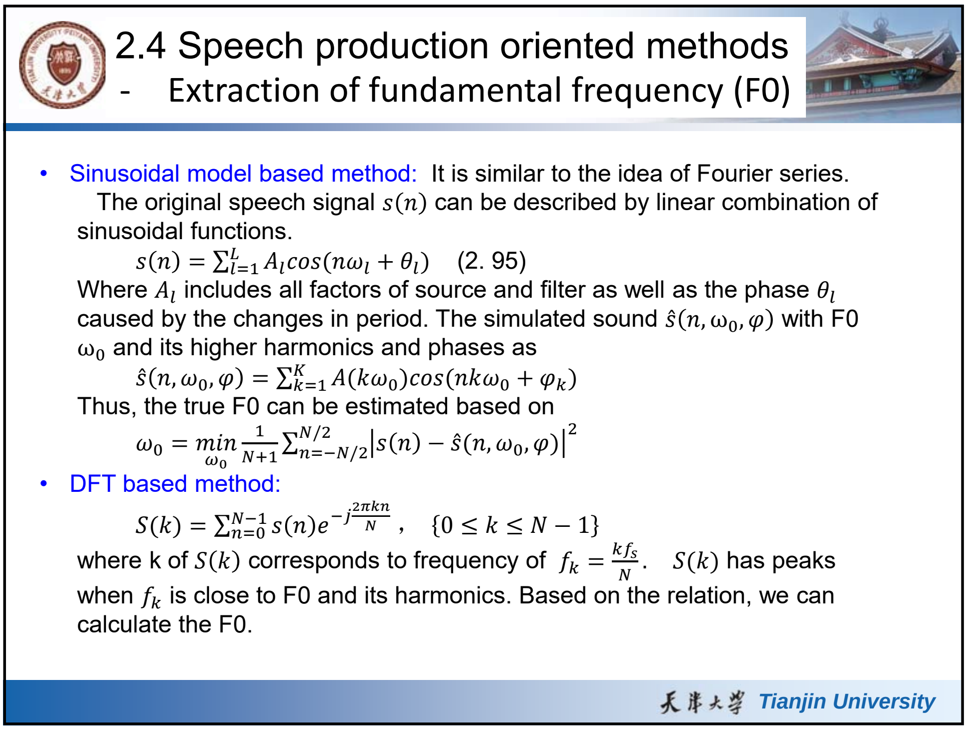
$$\varepsilon(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i)$$

- Cepstrum based method: separate source and filter in "quefrency" axis

$$C(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} log|X(k)|\, e^{j\omega n} d\omega\ = \frac{1}{2\pi}\int_{-\pi}^{\pi} (log|S(k)| + log|H(k)|)\, e^{j\omega n} d\omega$$

$$C(n) = \begin{cases} C_L(n) & n < F0 \\ C_H(n) & n \geq F0 \end{cases}$$ get F0 by a "lifter"

天津大学 Tianjin University



2.4 Speech production oriented methods - Extraction of fundamental frequency (F0)

- Sinusoidal model based method: It is similar to the idea of Fourier series.

The original speech signal $s(n)$ can be described by linear combination of sinusoidal functions.

$$s(n) = \sum_{l=1}^{L} A_l cos(n\omega_l + \theta_l) \quad (2.\ 95)$$

Where A_l includes all factors of source and filter as well as the phase θ_l caused by the changes in period. The simulated sound $\hat{s}(n, \omega_0, \varphi)$ with F0 ω_0 and its higher harmonics and phases as

$$\hat{s}(n, \omega_0, \varphi) = \sum_{k=1}^{K} A(k\omega_0) cos(nk\omega_0 + \varphi_k)$$

Thus, the true F0 can be estimated based on

$$\omega_0 = \min_{\omega_0} \frac{1}{N+1}\sum_{n=-N/2}^{N/2} |s(n) - \hat{s}(n, \omega_0, \varphi)|^2$$

- DFT based method:

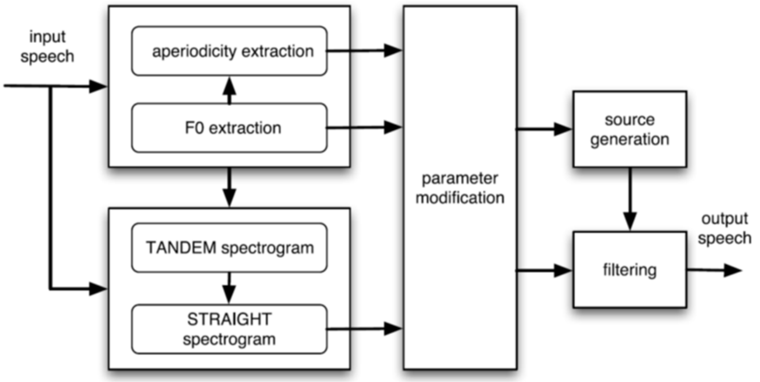
$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-j\frac{2\pi kn}{N}},\quad \{0 \leq k \leq N-1\}$$

where k of $S(k)$ corresponds to frequency of $f_k = \frac{kf_s}{N}$. $S(k)$ has peaks when f_k is close to F0 and its harmonics. Based on the relation, we can calculate the F0.

天津大学 Tianjin University

2.4 Speech production oriented methods - Extraction of fundamental frequency (F0)

- TANDEM-STRAIGHT (Kawahara, et al. 2009)



Schematic diagram of TANDEM-STRAIGHT. Manipulation block modifies source parameters (F0 and aperiodicity) and STRAIGHT spectrogram as well as their coordinates (time and frequency axes).

2.4 Speech production oriented methods - Extraction of fundamental frequency (F0)

TANDEM is for extracting temporally static power spectra.

$$P_T(\omega,t) = \left[\left|S(\omega,t-T_0/4)\right|^2 + \left|S(\omega,t+T_0/4)\right|^2\right]/2 \qquad (2.\ 4.27)$$

where $P_T(\omega,t)$ represents the TANDEM spectrogram. $S(\omega,t)$ is spectrum of speech signal at time t. The interval T_0 represents the reciprocal of fundamental frequency f_0. Temporally varying components in $\left|S(\omega,t)\right|^2$ are virtually cancelled by (2.4.27).

Spectral envelope estimation consists of two stages: spectral smoothing and compensation for consistent sampling.

$$C(\omega,t) = \int_{\omega_L}^{\omega} P_T(\gamma,t)d\gamma$$

$$L(\omega,t) = log(C(\omega+\omega_0/2\,,t) - C(\omega-\omega_0/2\,,t))$$

$$P_{ST}(\omega,t) = exp\left[q_0 L(\omega,t) + q_1\left(L(\omega+\omega_0,t) + L(\omega-\omega_0,t)\right)\right]$$

where $P_{ST}(\omega,t)$ represents the STRAIGHT spectrogram and $\omega_0 = 2\pi f_0$ is fundamental angular frequency. The initial two lines are an implementation of spectral smoothing using a rectangular smoothing kernel of width f_0. The last equation approximately applies the consistent sampling concept by introducing compensation constants q_0 and q_1, which are calculated from autocorrelation of the Fourier transform of the time windowing function.

2.4 Speech production oriented methods

- Extraction of fundamental frequency (F0)

This implementation assures positivity of the resultant STRAIGHT spectrogram.

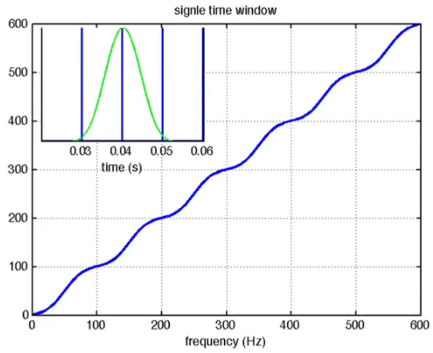
1) F0 extraction:

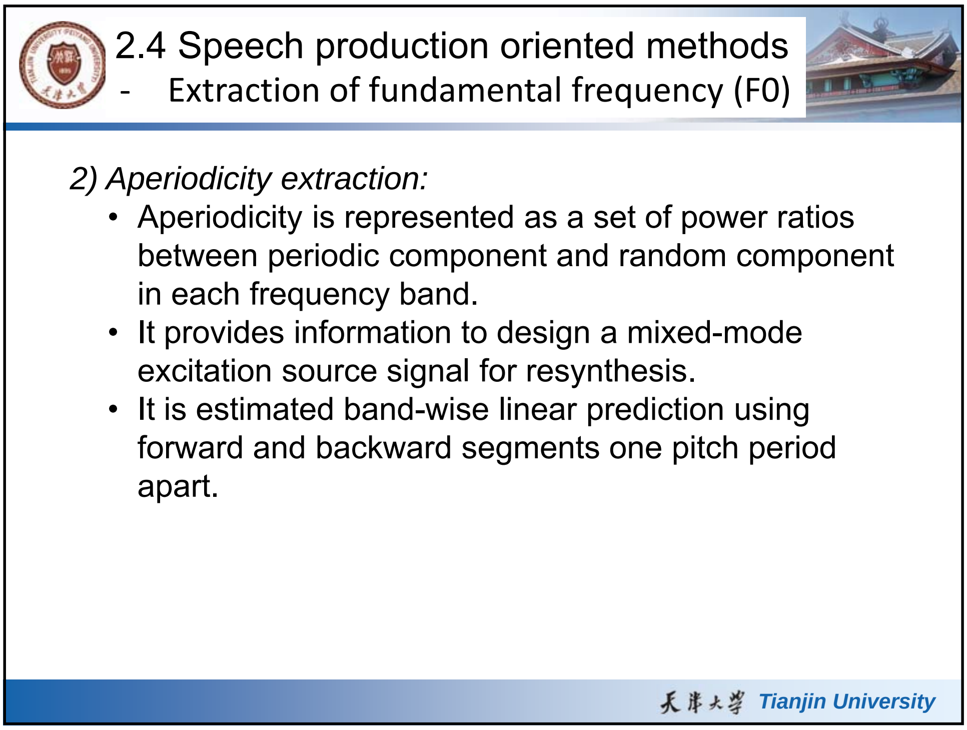
- The default F0 extractor of TANDEM-STRAIGHT is based on spectral division using two power spectral representations.
- Dividing a power spectrum consisting of periodicity information by its envelope leaves a power spectrum with periodicity information only and yields a periodicity detector specialized to its assumed F0.
- The default F0 extractor consists of a set of these specialized detectors setting assumed F0 candidates equidistant on the log-frequency axis covering normal F0 range.

TANDEM

TANDEM is for extracting temporally static power spectra.

$$P_T(\omega,t) = \left[\left|S(\omega,t-T_0/4)\right|^2 + \left|S(\omega,t+T_0/4)\right|^2\right]/2$$



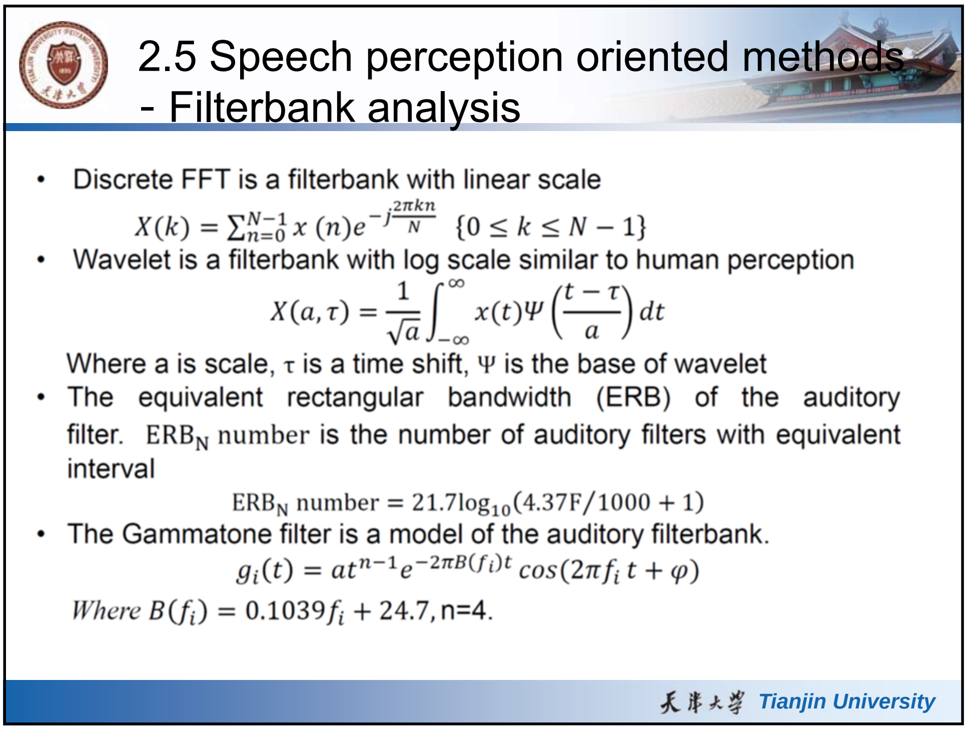


2.4 Speech production oriented methods - Extraction of fundamental frequency (F0)

2) Aperiodicity extraction:

- Aperiodicity is represented as a set of power ratios between periodic component and random component in each frequency band.
- It provides information to design a mixed-mode excitation source signal for resynthesis.
- It is estimated band-wise linear prediction using forward and backward segments one pitch period apart.

天津大学 Tianjin University



2.5 Speech perception oriented methods - Filterbank analysis

- Discrete FFT is a filterbank with linear scale

$$X(k) = \sum_{n=0}^{N-1} x\,(n) e^{-j\frac{2\pi kn}{N}} \quad \{0 \le k \le N-1\}$$

- Wavelet is a filterbank with log scale similar to human perception

$$X(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \Psi\left(\frac{t-\tau}{a}\right) dt$$

Where a is scale, τ is a time shift, Ψ is the base of wavelet

- The equivalent rectangular bandwidth (ERB) of the auditory filter. $\mathrm{ERB_N}$ number is the number of auditory filters with equivalent interval

$$\mathrm{ERB_N\ number} = 21.7\log_{10}(4.37\mathrm{F}/1000 + 1)$$

- The Gammatone filter is a model of the auditory filterbank.

$$g_i(t) = at^{n-1} e^{-2\pi B(f_i)t} \cos(2\pi f_i\, t + \varphi)$$

$Where\ B(f_i) = 0.1039 f_i + 24.7$, n=4.

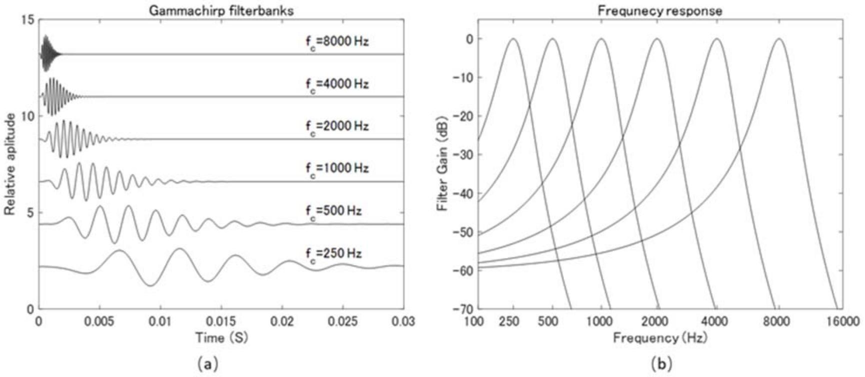
天津大学 Tianjin University

Auditory filter and modeling (cont.)

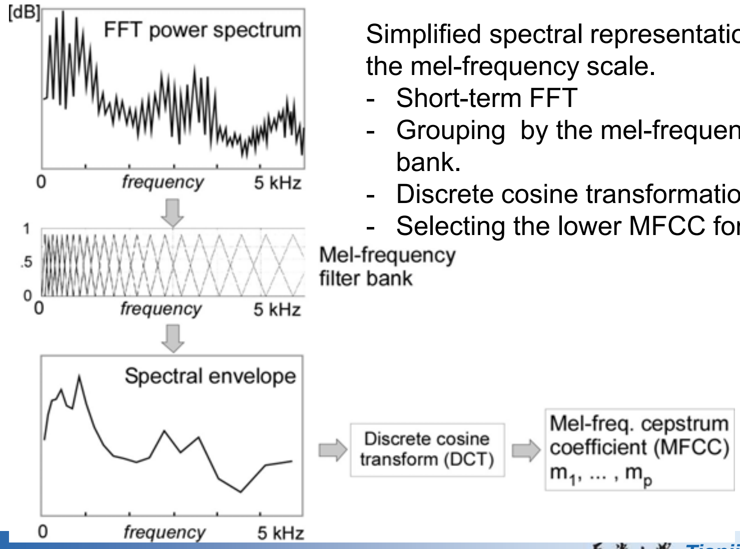
- An improved version of the gammatone model is the gammachirp filter,

$$g_i^c(t) = at^{n-1}e^{-2\pi B(f_i)t}\, cos(2\pi f_i\, t + cln(t) + \varphi)$$

where n=4, x=-2,96

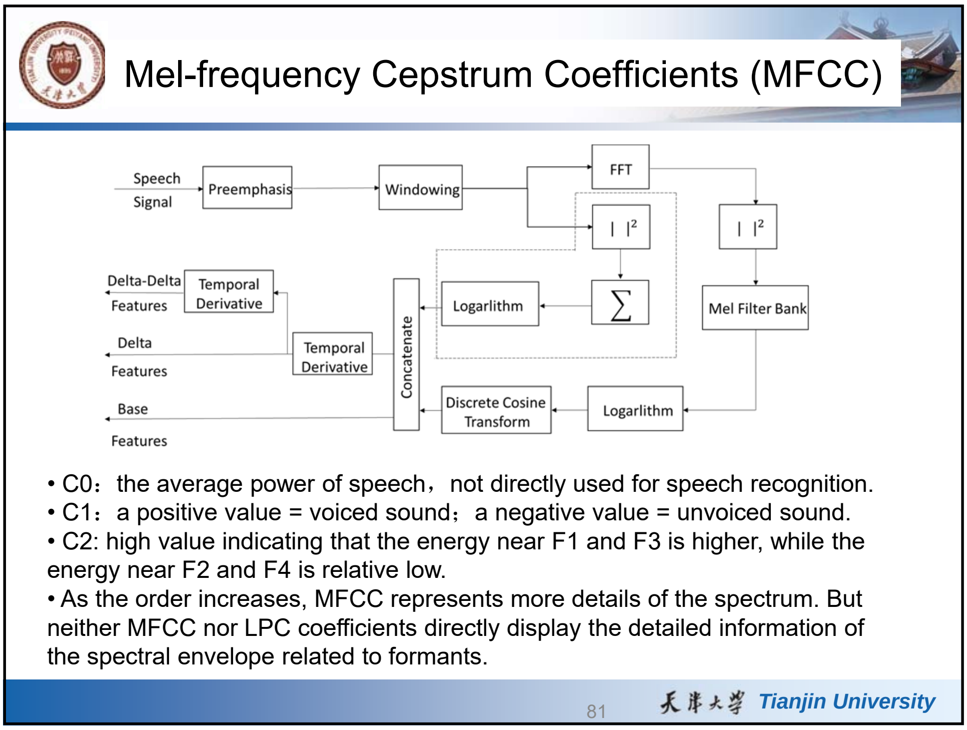


Mel-frequency Cepstrum Coefficients (MFCC)


[dB]
FFT power spectrum
0
frequency
5 kHz
1
.5
0
0
frequency
5 kHz
Mel-frequency filter bank
Spectral envelope
0
frequency
5 kHz
Discrete cosine transform (DCT)
Mel-freq. cepstrum coefficient (MFCC) $m_1, \dots, m_p$

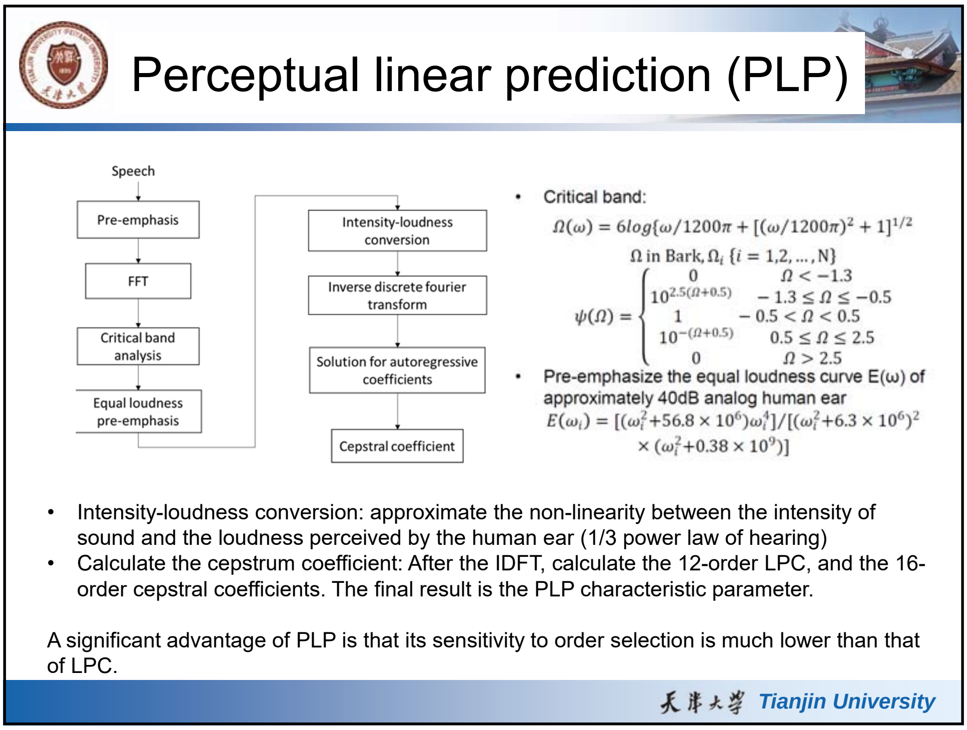
Simplified spectral representation along the mel-frequency scale.

- Short-term FFT
- Grouping by the mel-frequency filter bank.
- Discrete cosine transformation (DCT)
- Selecting the lower MFCC for ASR



Mel-frequency Cepstrum Coefficients (MFCC)

- C0：the average power of speech，not directly used for speech recognition.
- C1：a positive value = voiced sound；a negative value = unvoiced sound.
- C2: high value indicating that the energy near F1 and F3 is higher, while the energy near F2 and F4 is relative low.
- As the order increases, MFCC represents more details of the spectrum. But neither MFCC nor LPC coefficients directly display the detailed information of the spectral envelope related to formants.



Perceptual linear prediction (PLP)

- Critical band:

$$\Omega(\omega) = 6log\{\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{1/2}$$

$$\Omega \text{ in Bark}, \Omega_i\ \{i = 1,2,\dots,\mathrm{N}\}$$

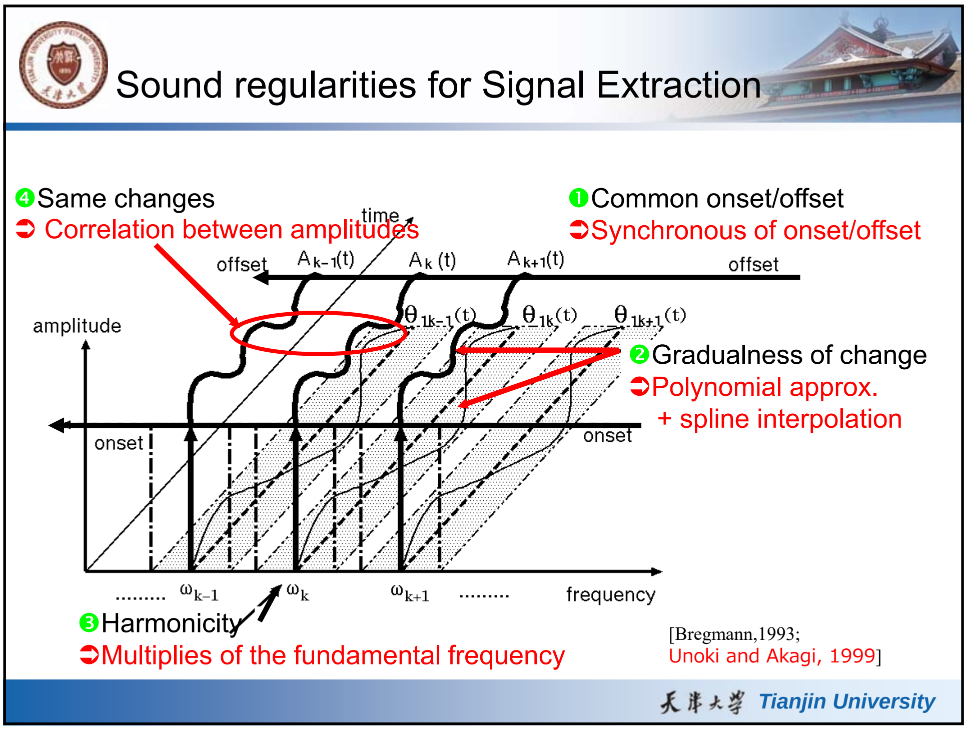
$$\psi(\Omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & -1.3 \le \Omega \le -0.5 \\ 1 & -0.5 < \Omega < 0.5 \\ 10^{-(\Omega+0.5)} & 0.5 \le \Omega \le 2.5 \\ 0 & \Omega > 2.5 \end{cases}$$

- Pre-emphasize the equal loudness curve E(ω) of approximately 40dB analog human ear

$$E(\omega_i) = [(\omega_i^2 + 56.8 \times 10^6)\omega_i^4]/[(\omega_i^2 + 6.3 \times 10^6)^2 \times (\omega_i^2 + 0.38 \times 10^9)]$$

- Intensity-loudness conversion: approximate the non-linearity between the intensity of sound and the loudness perceived by the human ear (1/3 power law of hearing)
- Calculate the cepstrum coefficient: After the IDFT, calculate the 12-order LPC, and the 16-order cepstral coefficients. The final result is the PLP characteristic parameter.

A significant advantage of PLP is that its sensitivity to order selection is much lower than that of LPC.



Sound regularities for Signal Extraction

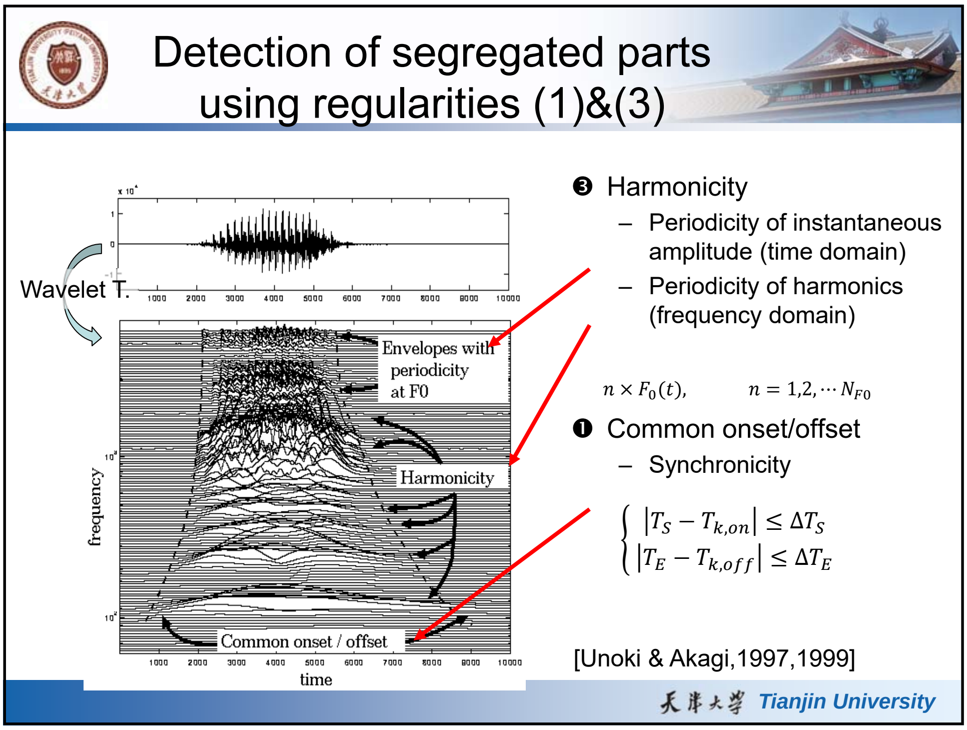
❹Same changes
➲ Correlation between amplitudes

❶Common onset/offset
➲Synchronous of onset/offset

❷Gradualness of change
➲Polynomial approx.
+ spline interpolation

❸Harmonicity
➲Multiplies of the fundamental frequency

[Bregmann,1993;
Unoki and Akagi, 1999]



Detection of segregated parts using regularities (1)&(3)

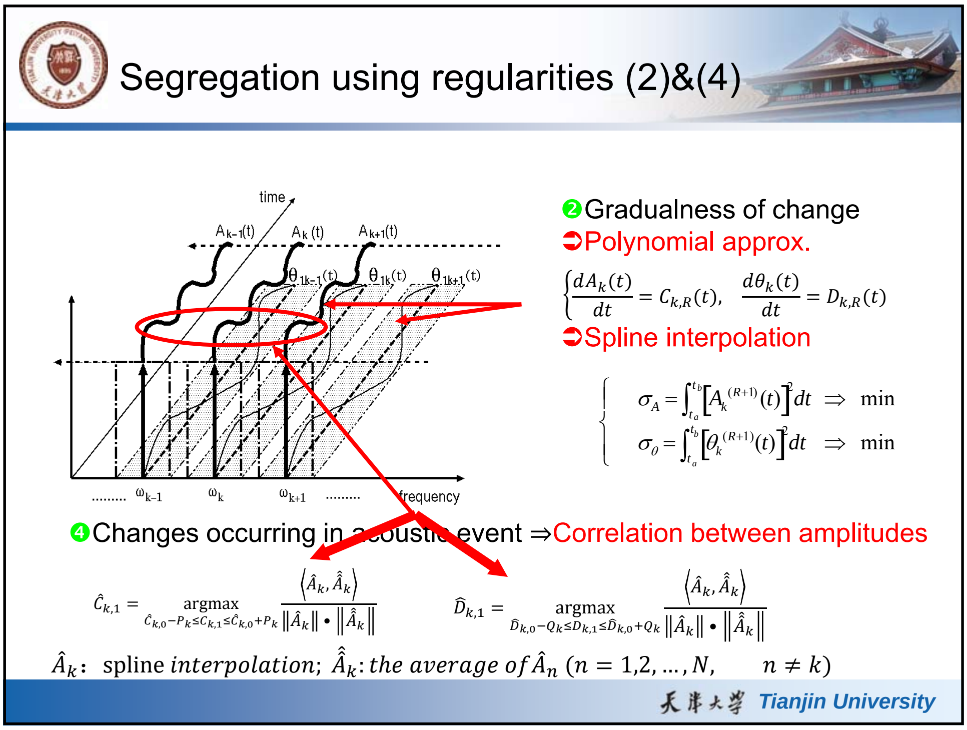
❸ Harmonicity
- Periodicity of instantaneous amplitude (time domain)
- Periodicity of harmonics (frequency domain)

$n \times F_0(t), \qquad n = 1,2,\cdots N_{F0}$

❶ Common onset/offset
- Synchronicity

$$\begin{cases} |T_S - T_{k,on}| \le \Delta T_S \\ |T_E - T_{k,off}| \le \Delta T_E \end{cases}$$

[Unoki & Akagi,1997,1999]



Segregation using regularities (2)&(4)

❷Gradualness of change

➲Polynomial approx.

$$\begin{cases}\dfrac{dA_k(t)}{dt} = C_{k,R}(t), \quad \dfrac{d\theta_k(t)}{dt} = D_{k,R}(t)\end{cases}$$

➲Spline interpolation

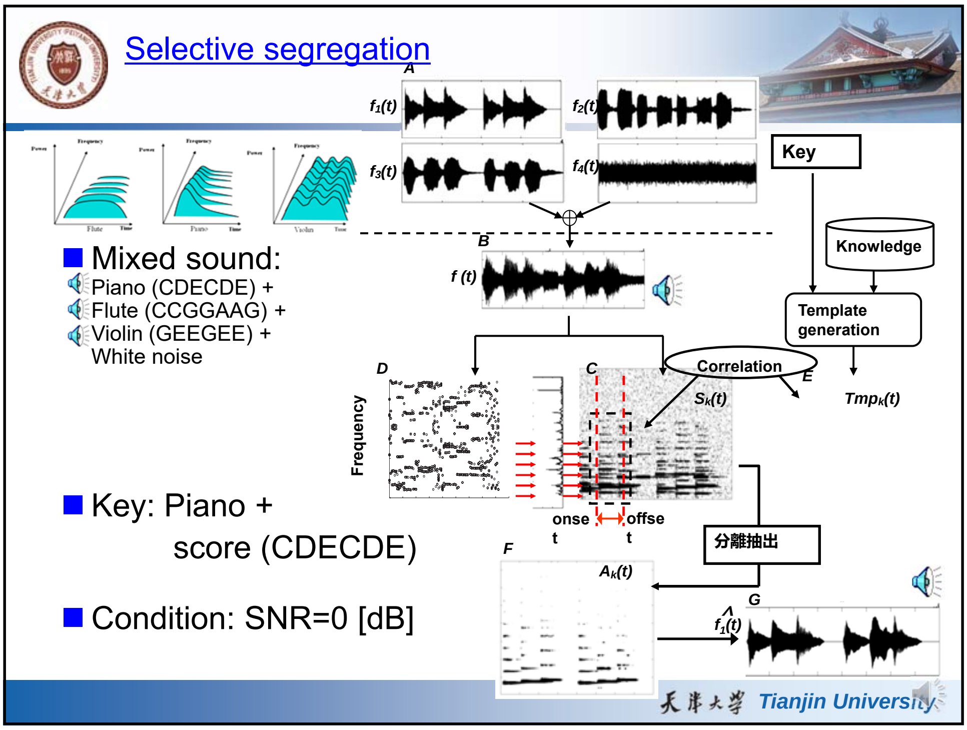
$$\begin{cases}\sigma_A = \int_{t_a}^{t_b}\left[A_k^{(R+1)}(t)\right]^2 dt \Rightarrow \min \\ \sigma_\theta = \int_{t_a}^{t_b}\left[\theta_k^{(R+1)}(t)\right]^2 dt \Rightarrow \min\end{cases}$$

❹Changes occurring in acoustic event ⇒Correlation between amplitudes

$$\hat{C}_{k,1} = \underset{\hat{C}_{k,0}-P_k \le C_{k,1} \le \hat{C}_{k,0}+P_k}{\operatorname{argmax}} \frac{\left\langle \hat{A}_k, \hat{\hat{A}}_k \right\rangle}{\left\|\hat{A}_k\right\| \bullet \left\|\hat{\hat{A}}_k\right\|} \qquad \hat{D}_{k,1} = \underset{\hat{D}_{k,0}-Q_k \le D_{k,1} \le \hat{D}_{k,0}+Q_k}{\operatorname{argmax}} \frac{\left\langle \hat{A}_k, \hat{\hat{A}}_k \right\rangle}{\left\|\hat{A}_k\right\| \bullet \left\|\hat{\hat{A}}_k\right\|}$$

$\hat{A}_k$: spline $interpolation$; $\hat{\hat{A}}_k$: $the\ average\ of\ \hat{A}_n$ $(n = 1,2,\dots,N, \quad n \neq k)$

天津大学 Tianjin University



Selective segregation

- Mixed sound:
  Piano (CDECDE) +
  Flute (CCGGAAG) +
  Violin (GEEGEE) +
  White noise
- Key: Piano +
  score (CDECDE)
- Condition: SNR=0 [dB]

天津大学 Tianjin University

2.6 Statistical base methods

Wiener-Khintchine theorem

Power spectrum <==> autocorrelation function

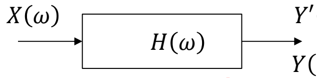
$$\Phi_{xx}(\omega) \quad = \quad F[\phi_{xx}(\tau)]$$

$$\begin{aligned}
F^{-1}[|X(\omega)|^2] &= \frac{1}{2\pi}\int_{-\infty}^{\infty}|X(\omega)|^2 e^{j\omega t}d\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\{X^*(\omega)\int_{-\infty}^{\infty}x(\tau)e^{-j\omega\tau}d\tau\}e^{j\omega t}d\omega \\
&= \int_{-\infty}^{\infty}\{x(\tau)\frac{1}{2\pi}\int_{-\infty}^{\infty}X(-\omega)e^{-j\omega(\tau-t)}d\omega\}d\tau \\
&= \int_{-\infty}^{\infty}\{x(\tau)\frac{1}{2\pi}\int_{-\infty}^{\infty}X(\omega')e^{j\omega'(\tau-t)}d\omega'\}d\tau \\
&= \int_{-\infty}^{\infty}x(\tau)x(\tau-t)d\tau
\end{aligned}$$

Noise reduction by cross power spectrum

Cross power spectrum $S_{xy}(\omega)$ of two signals, $x(t)$ and $y(t)$, is

$$S_{xy}(\omega) = \int_{-\infty}^{\infty}R_{xy}(\tau)e^{-j\omega\tau}d\tau, \qquad R_{xy}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}S_{xy}(\omega)e^{-j\omega\tau}d\omega$$



$Y'(\omega) = H(\omega)X(\omega)$ In ideal case

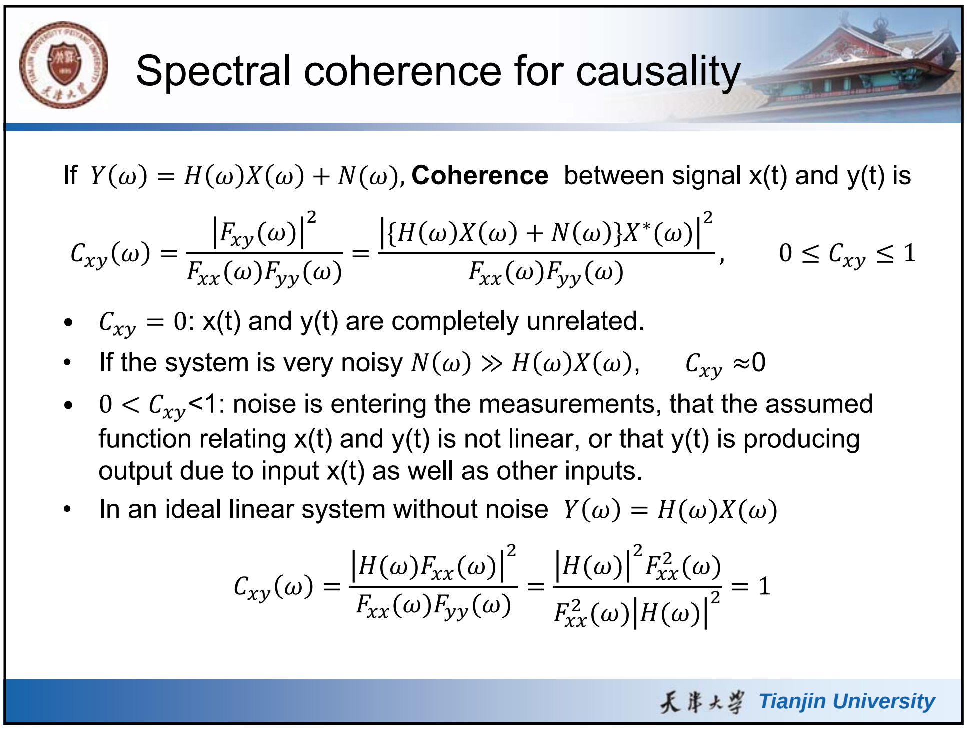
$Y(\omega) = H(\omega)X(\omega) + N(\omega)$

$$H(\omega) = \frac{Y(\omega)}{X(\omega)} + \frac{N(\omega)}{X(\omega)}$$

Not system property

$$\frac{Y(\omega)}{X(\omega)} = \frac{\{H(\omega)X(\omega) + N(\omega)\}X^*(\omega)}{X(\omega)X^*(\omega)} = \frac{F_{xy}(\omega)}{F_{xx}(\omega)} = H(\omega)$$

$N(\omega)X^*(\omega) = \int_{-\infty}^{\infty}R_{nx}(\tau)e^{-j\omega\tau}d\tau \approx 0$ If x(t) and noise n(t) are unrelated



Spectral coherence for causality

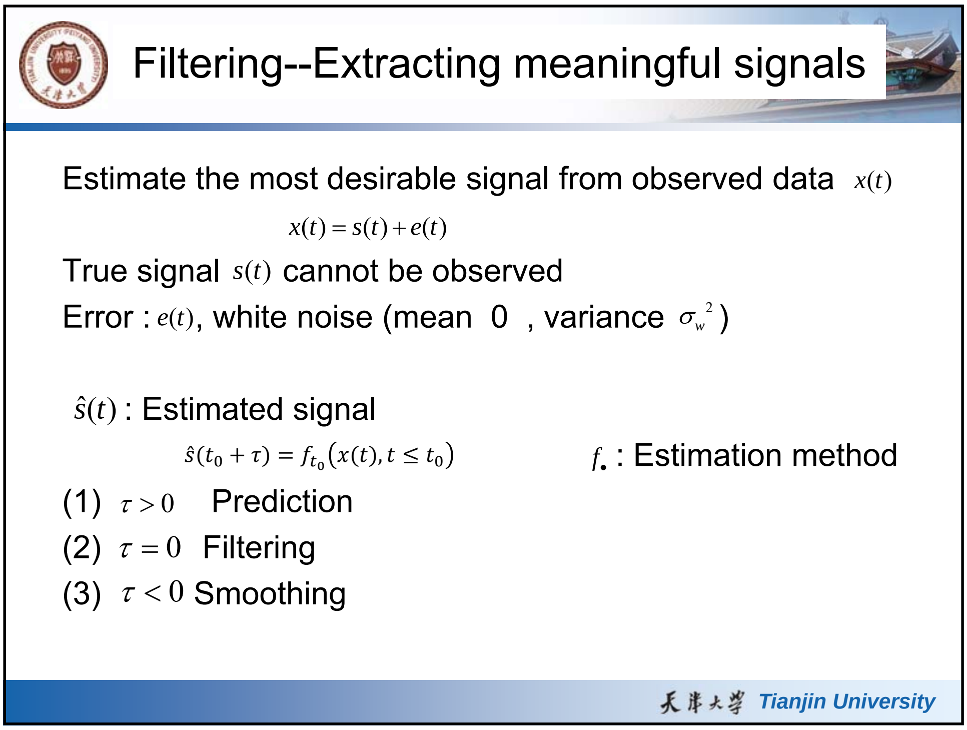
If $Y(\omega) = H(\omega)X(\omega) + N(\omega)$, **Coherence** between signal x(t) and y(t) is

$$C_{xy}(\omega) = \frac{\left|F_{xy}(\omega)\right|^2}{F_{xx}(\omega)F_{yy}(\omega)} = \frac{\left|\{H(\omega)X(\omega) + N(\omega)\}X^*(\omega)\right|^2}{F_{xx}(\omega)F_{yy}(\omega)}, \qquad 0 \le C_{xy} \le 1$$

- $C_{xy} = 0$: x(t) and y(t) are completely unrelated.
- If the system is very noisy $N(\omega) \gg H(\omega)X(\omega)$, $C_{xy} \approx 0$
- $0 < C_{xy} < 1$: noise is entering the measurements, that the assumed function relating x(t) and y(t) is not linear, or that y(t) is producing output due to input x(t) as well as other inputs.
- In an ideal linear system without noise $Y(\omega) = H(\omega)X(\omega)$

$$C_{xy}(\omega) = \frac{\left|H(\omega)F_{xx}(\omega)\right|^2}{F_{xx}(\omega)F_{yy}(\omega)} = \frac{\left|H(\omega)\right|^2 F_{xx}^2(\omega)}{F_{xx}^2(\omega)\left|H(\omega)\right|^2} = 1$$

天津大学 Tianjin University



Filtering--Extracting meaningful signals

Estimate the most desirable signal from observed data $x(t)$

$$x(t) = s(t) + e(t)$$

True signal $s(t)$ cannot be observed

Error : $e(t)$, white noise (mean 0 , variance σ_w^2)

$\hat{s}(t)$: Estimated signal

$$\hat{s}(t_0 + \tau) = f_{t_0}(x(t), t \le t_0)$$

$f_{\bullet}$: Estimation method

(1) $\tau > 0$ Prediction

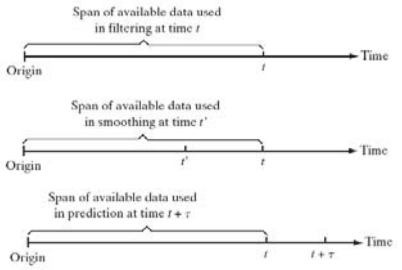
(2) $\tau = 0$ Filtering

(3) $\tau < 0$ Smoothing

天津大学 Tianjin University

The types of signal processing

- Filters may be used for three information-processing tasks
  - Filtering
  - Smoothing
  - Prediction


Span of available data used
in filtering at time t
Origin
t
Time
Span of available data used
in smoothing at time t'
Origin
t'
t
Time
Span of available data used
in prediction at time t + τ
Origin
t
t + τ
Time

- Given an optimality criteria we often can design optimal filters
  - Requires a priori information about the environment
  - Example: Under certain conditions the so called Wiener filter is optimal in the mean-squared sense
- Adaptive filters are self-designing using a recursive algorithm
  - Useful if complete knowledge of environment is not available a priori

Orthogonality principle for estimation

Let the true signal $s(t)$ be estimated by a weighted sum of observed signal values $x(t)$ as

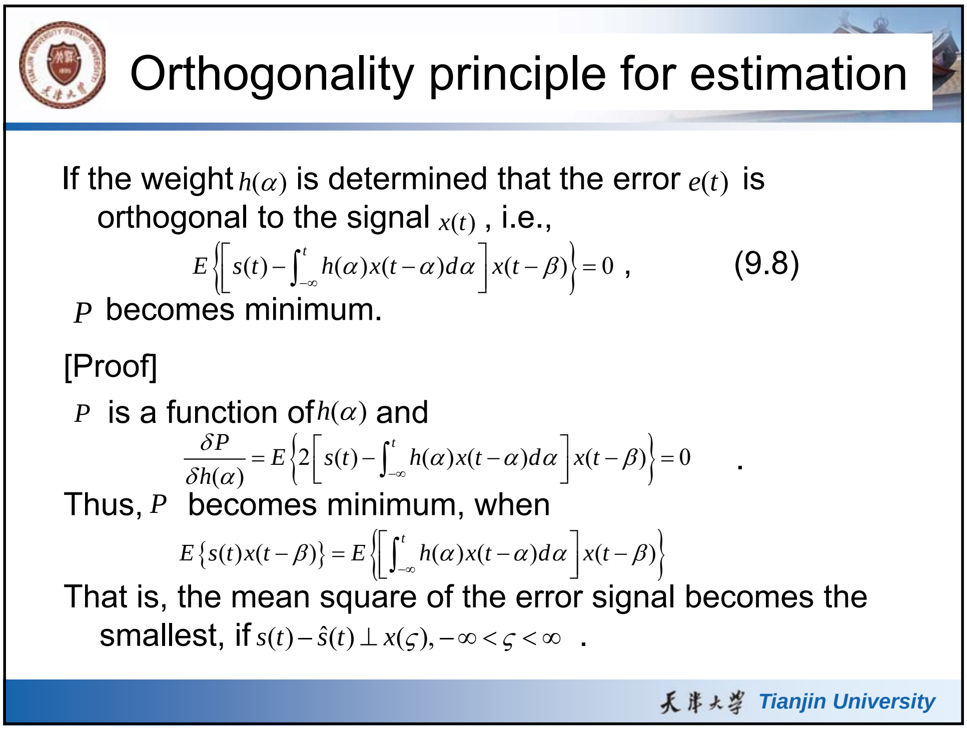
$$\hat{s}(t)=\int_{-\infty}^{\infty}h(\alpha)x(t-\alpha)d\alpha \qquad (9.6)$$

and require that a mean square of the error signal P becomes minimum,

$$P=E\left\{\left|s(t)-\hat{s}(t)\right|^2\right\}=E\left\{\left|s(t)-\int_{-\infty}^{\infty}h(\alpha)x(t-\alpha)d\alpha\right|^2\right\} \qquad (9.7)$$

where the error is given as

$$e(t)=s(t)-\hat{s}(t)=s(t)-\int_{-\infty}^{\infty}h(\alpha)x(t-\alpha)d\alpha$$



Orthogonality principle for estimation

If the weight $h(\alpha)$ is determined that the error $e(t)$ is orthogonal to the signal $x(t)$, i.e.,

$$E\left\{\left[s(t)-\int_{-\infty}^{t}h(\alpha)x(t-\alpha)d\alpha\right]x(t-\beta)\right\}=0\ , \qquad (9.8)$$

P becomes minimum.

[Proof]

P is a function of $h(\alpha)$ and

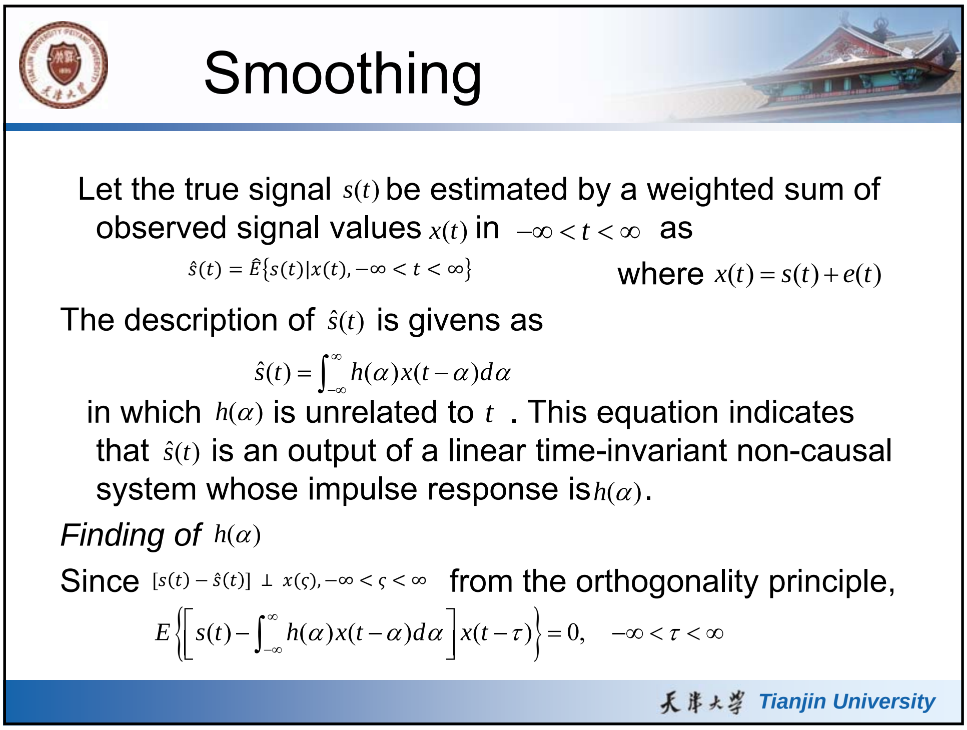
$$\frac{\delta P}{\delta h(\alpha)}=E\left\{2\left[s(t)-\int_{-\infty}^{t}h(\alpha)x(t-\alpha)d\alpha\right]x(t-\beta)\right\}=0 \quad .$$

Thus, P becomes minimum, when

$$E\{s(t)x(t-\beta)\}=E\left\{\left[\int_{-\infty}^{t}h(\alpha)x(t-\alpha)d\alpha\right]x(t-\beta)\right\}$$

That is, the mean square of the error signal becomes the smallest, if $s(t)-\hat{s}(t)\perp x(\varsigma),-\infty<\varsigma<\infty$.

天津大学 Tianjin University



Smoothing

Let the true signal $s(t)$ be estimated by a weighted sum of observed signal values $x(t)$ in $-\infty<t<\infty$ as

$$\hat{s}(t)=\hat{E}\{s(t)|x(t),-\infty<t<\infty\}$$

where $x(t)=s(t)+e(t)$

The description of $\hat{s}(t)$ is givens as

$$\hat{s}(t)=\int_{-\infty}^{\infty}h(\alpha)x(t-\alpha)d\alpha$$

in which $h(\alpha)$ is unrelated to t . This equation indicates that $\hat{s}(t)$ is an output of a linear time-invariant non-causal system whose impulse response is $h(\alpha)$.

Finding of $h(\alpha)$

Since $[s(t)-\hat{s}(t)]\perp x(\varsigma),-\infty<\varsigma<\infty$ from the orthogonality principle,

$$E\left\{\left[s(t)-\int_{-\infty}^{\infty}h(\alpha)x(t-\alpha)d\alpha\right]x(t-\tau)\right\}=0,\quad -\infty<\tau<\infty$$

天津大学 Tianjin University

Wiener integral equation and Wiener filter

Then,

$$R_{sx}(\tau)=\int_{-\infty}^{\infty}h(\alpha)R_{xx}(\tau-\alpha)d\alpha$$

This equation is called Wiener integral equation.

If the both sides of this equation are transformed by Fourier transformation,

$$S_{sx}(\omega)=H(\omega)S_{xx}(\omega)$$

Then, $$H(\omega)=\frac{S_{sx}(\omega)}{S_{xx}(\omega)}$$

This is a non-causal Wiener filter. The mean square estimation error P is given by

Wiener integral equation and Wiener filter

$$P=E\left\{\left(s(t)-\hat{s}(t)\right)s(t)\right\}=E\left\{\left[s(t)-\int_{-\infty}^{\infty}h(\alpha)x(t-\alpha)d\alpha\right]s(t)\right\}$$

$$=R_{SS}(0)-\int_{-\infty}^{\infty}h(\alpha)R_{sx}(\alpha)d\alpha=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left[S_{ss}(\omega)-H^{*}(\omega)S_{sx}(\omega)\right]e^{j\omega\tau}d\omega\Big|_{\tau=0}$$

$$H^{*}(\omega)S_{sx}(\omega)=\left|H(\omega)\right|^{2}S_{xx}(\omega)$$

If $s(t)$ and $e(t)$ are orthogonal,

$$E\left\{s(t)e(t)\right\}=0$$

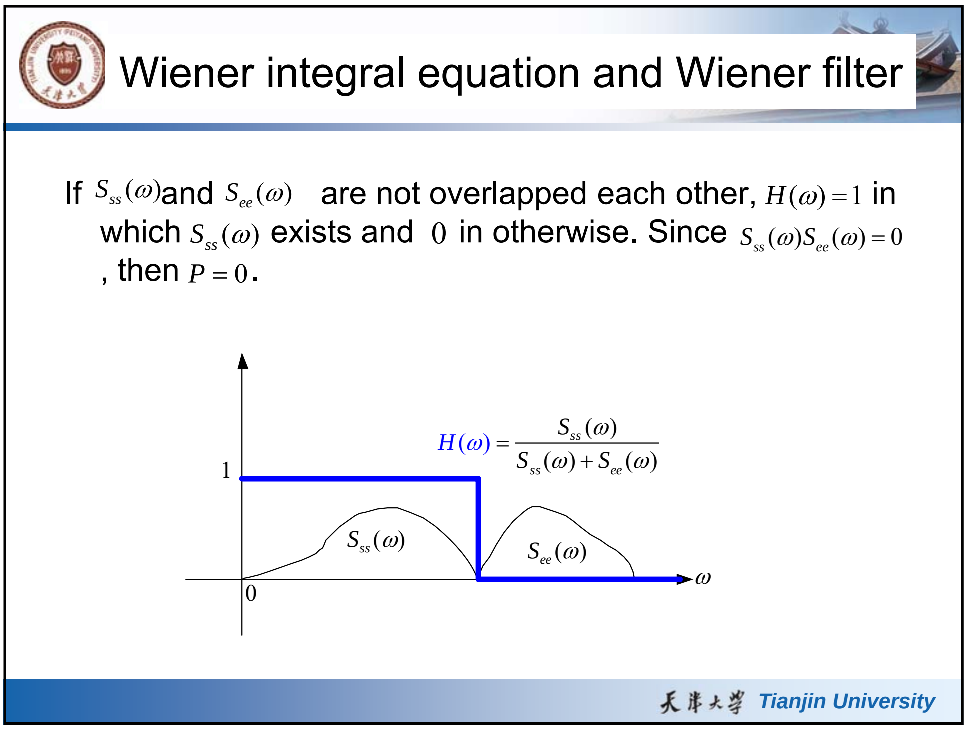
$$E\left\{s(t)x(t)\right\}=E\left\{s(t)\left[s(t)+e(t)\right]\right\}=E\left[s^{2}(t)\right]\rightarrow\qquad S_{sx}(\omega)=S_{ss}(\omega)$$

And $S_{xx}(\omega)=S_{ss}(\omega)+S_{ee}(\omega)$

Then $$H(\omega)=\frac{S_{ss}(\omega)}{S_{ss}(\omega)+S_{ee}(\omega)}$$

$$P=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left[S_{ss}(\omega)-H(\omega)S_{sx}(\omega)\right]d\omega$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left[S_{ss}(\omega)-\frac{S_{ss}(\omega)}{S_{ss}(\omega)+S_{ee}(\omega)}S_{ss}(\omega)\right]d\omega=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{S_{ss}(\omega)S_{ee}(\omega)}{S_{ss}(\omega)+S_{ee}(\omega)}d\omega$$



Wiener integral equation and Wiener filter

If $S_{ss}(\omega)$and $S_{ee}(\omega)$ are not overlapped each other, $H(\omega)=1$ in which $S_{ss}(\omega)$ exists and 0 in otherwise. Since $S_{ss}(\omega)S_{ee}(\omega)=0$, then $P=0$.

$$H(\omega)=\frac{S_{ss}(\omega)}{S_{ss}(\omega)+S_{ee}(\omega)}$$

天津大学 Tianjin University