Speech Signal Processing

Reference book: Peter Vary and Rainer Martin, "Digital Speech Transmission – Enhancement, coding and error concealment"

Lectures : 12 Exercises : 12

Projects : 2 Final Examination

Project Assignment 1 : Speech production and Analysis
Submission Deadline : 12 th Nov (submission of presentation slides and sound files)

Project Assignment 2 : Speech Compression and Quantization
Submission Deadline : 7th Dec (submission of report)

Final Exam : 10 th Dec.

SPEECH PRODUCTION ①

As a detailed discussion of the acoustic theory of speech production, phonetics, psychoacoustics, and perception is beyond the scope of this book, the reader is referred to the literature (e.g., [Fant 1970], [Flanagan 1972], [Rabiner, Schafer 1978], [Picket 1980], [Zwicker, Fastl 1999]).

2.1 Organs of Speech Production

The production of speech sounds involves the manipulation of an airstream. The acoustic representation of speech is a sound pressure wave originating from the physiological speech production system. A simplified schematic of the human speech organs is given in Fig. 2.1.

The main components and their functions are:

- lungs: the energy generator,
- trachea: for energy transport,
- larynx with vocal cords: the signal generator, and
- vocal tract: (pharynx, oral and nasal cavities) the acoustic filter.

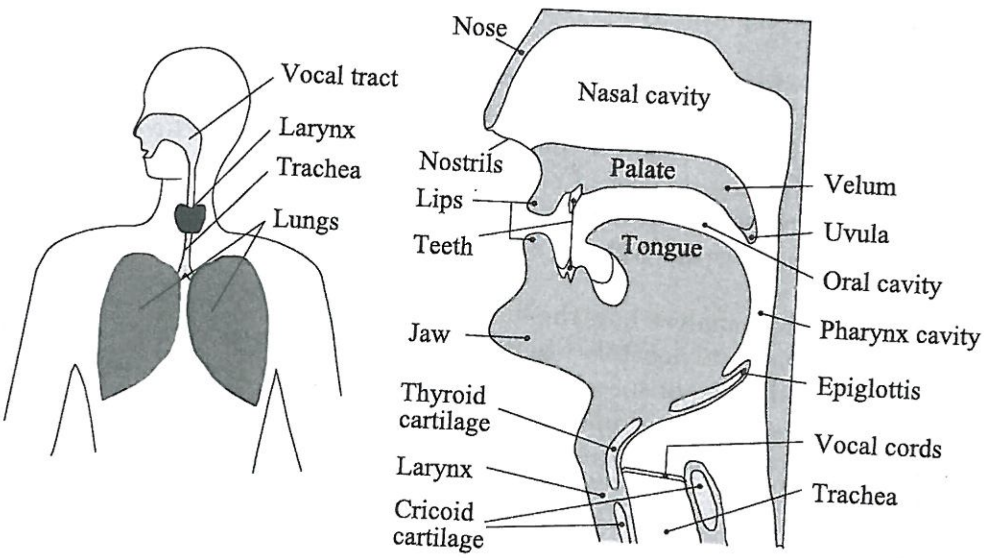


Figure 2.1: Organs of speech production

Speech Sound Types

- Most languages can be described as a set of elementary linguistic units which are called <u>phonemes</u>. A phoneme is defined as the smallest unit which differentiates between two words in one language. The acoustic representation associated with a phoneme is called a <u>phone</u>.

- For example, American English consists of about 42 phonemes.

  - vowels /a/ in 'father'
    (steady sound) /i/ in 'eve'

  - Diphthongs (transition sound one vowel to another vowel)
    /oU/ in 'boat', /ju/ in 'you'

  - Approximants (intermediate between vowels and consonants)
    /w/ in 'wet', /r/ in 'ran'

  - consonants /m/ in more, /t/ in 'tree'

- For Speech Signal Processing, we are more interested in signal characteristics or waveforms. Depending on these aspects, the categories are
  - voiced
  - unvoiced
  - mixed voiced/unvoiced
  - plosive
  - silence

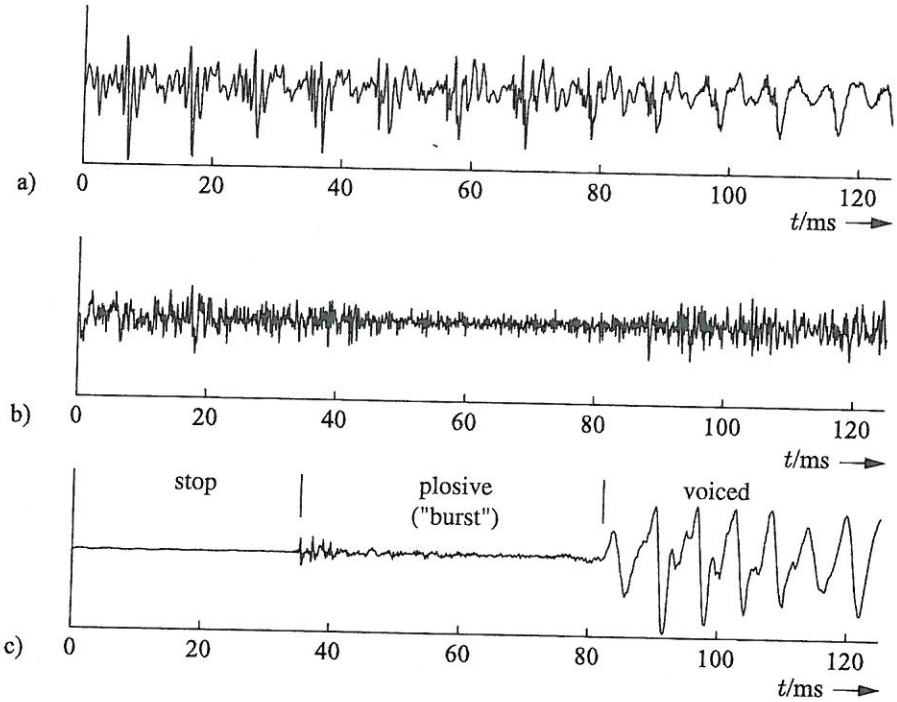


Figure 2.3: Characteristic waveforms of speech signals (each line = 100 ms) [Vary et al. 1998]
a) Voiced (vowel with transition to voiced consonant)
b) Unvoiced (fricative)
c) Transition stop–plosive–vowel

Speech waveforms

3.1

TABLE 2.1. A condensed list of phonetic symbols for American English.

Phoneme	ARPABET	Example	Phoneme	ARPABET	Example
/i/	IY	beat	/ŋ/	NX	sing
/I/	IH	bit	/p/	P	pet
/e/ (eʸ)	EY	bait	/t/	T	ten
/ɛ/	EH	bet	/k/	K	kit
/æ/	AE	bat	/b/	B	bet
/a/	AA	Bob	/d/	D	debt
/ʌ/	AH	but	/g/	H	get
/ɔ/	AO	bought	/h/	HH	hat
/o/ (oʷ)	OW	boat	/f/	F	fat
/U/	UH	book	/θ/	TH	thing
/u/	UW	boot	/s/	S	sat
/ə/	AX	about	/š/ (sh)	SH	shut
/ɨ/	IX	roses	/v/	V	vat
/ɝ/	ER	bird	/ð/	DH	that
/ɚ/	AXR	butter	/z/	Z	zoo
/aʷ/	AW	down	/ž/ (zh)	ZH	azure
/aʸ/	AY	buy	/č/ (tsh)	CH	church
/ɔʸ/	OY	boy	/ǰ/ (dzh, j)	JH	judge
/y/	Y	you	/ʍ/	WH	which
/w/	W	wit	/l̩/	EL	battle
/r/	R	rent	/m̩/	EM	bottom
/l/	L	let	/n̩/	EN	button
/m/	M	met	/ɾ/	DX	batter
/n/	N	net	/ʔ/	Q	(glottal stop)

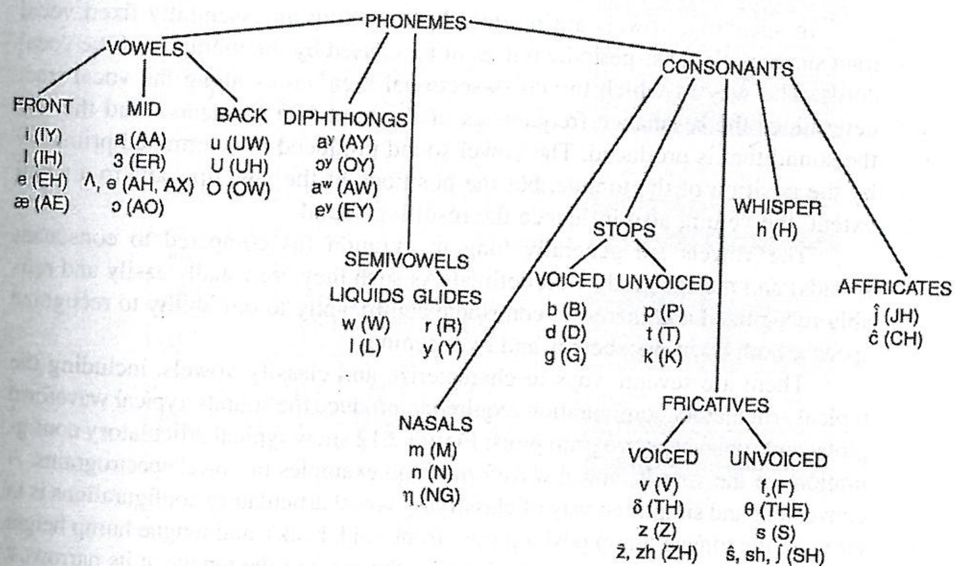
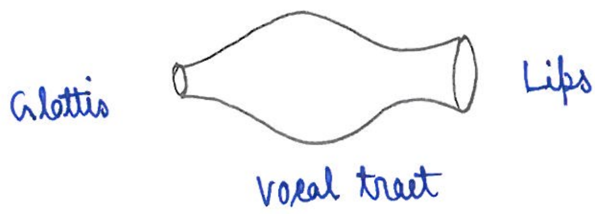
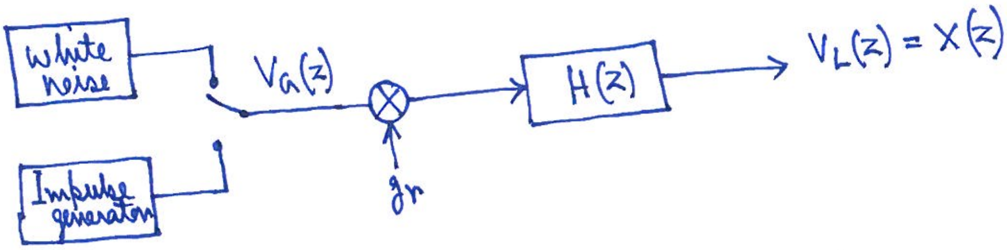


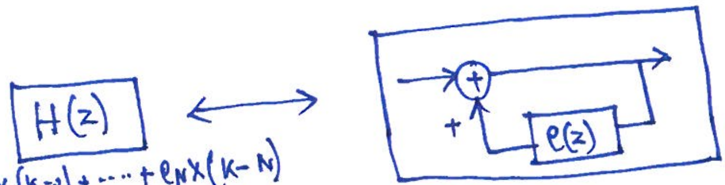
Figure 2.11 Chart of the classification of the standard phonemes of American English into broad sound classes.

Model of Speech Production



Source-filter model (production-filter model)





$$X(z) = V_G(z) + e(z)\,X(z)$$

$$x(k) = V_G(k) + e_1 x(k-1) + e_2 x(k-2) + \cdots + e_N x(k-N)$$

$$e(z) = e_1 z^{-1} + e_2 z^{-2} + \cdots + e_N z^{-N} = \sum_{i=1}^{N} e_i z^{-i}$$

$$\therefore \quad H(z) = \frac{X(z)}{V_G(z)} = \frac{1}{1-e(z)} = \frac{1}{1-\sum_{i=1}^{N} e_i z^{-i}}$$

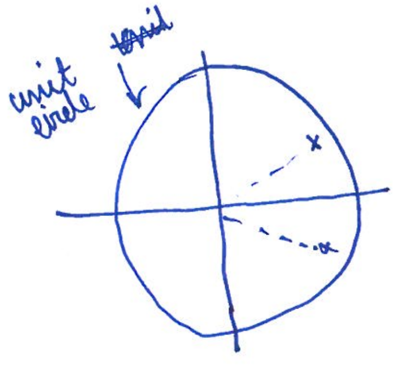
Now, if $\{e_i\}_{i=1}^{N}$ are real, then if we factorize

$$H(z) = \frac{1}{1-e(z)} = \frac{1}{1-\sum_{i=1}^{N} e_i z^{-i}}$$

$$= \frac{1}{\underbrace{(1-p_1 z^{-1})(1-p_1^* z^{-1})}_{\text{conjugate poles.}} \cdots\cdots (1-p_K z^{-K})(1-p_K^* z^{-K})}$$

assuming $N = 2K$

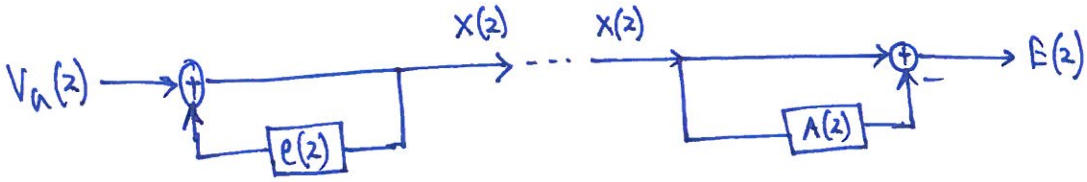
K pole pairs.



$$p_i = r_i e^{\pm j\theta_i}$$

$$p_i^* = r_i e^{\mp j\theta_i}$$

Synthesis – Analysis framework



If $A(z) = e(z)$ then $E(z) = V_a(z)$.

$$X(z) = V_a(z) + e(z)\,X(z)$$
$$E(z) = X(z) - A(z)\,X(z)$$
$$\overline{X(z) + E(z) = V_a(z) + X(z) + \left[\cancelto{0}{e(z)\,X(z) - A(z)\,X(z)}\right]}$$
$$\therefore\ E(z) = V_a(z)$$

- In reality, we do not know $e(z)$. But want to model it as $A(z)$ and estimate the $A(z)$.
- $A(z)$ estimation is done by linear prediction.
- We assume that $E(z)$ ~~and~~ corresponding signal $e(k)$ has a spectrum that is white.

$A(z)$: prediction filter $\qquad \tilde{x}(k) = a_1 x(k-1) + a_2 x(k-2) + \cdots + \cdots + a_p x(k-p)$.

$p \rightarrow$ is the order of filter.

$1 - A(z)$: Analysis filter $\qquad e(k) = x(k) - \tilde{x}(k)$.

$\dfrac{1}{1 - A(z)}$: synthesis filter. $\qquad$ Using this filter, we get $X(z)$ from $e(k)$.

Filter order: what is a typical value of 'p'?

- Vocal tract is assumed to be a lossless tube, closed at the glottis side and open at lips.
- Assume that a male adult has a typical length of vocal tract, $L = 17$ cm.
- Then the wavelength λ_i of each resonance frequency fulfils the standing waveform condition

$$(2i-1) \cdot \frac{\lambda_i}{4} = L \quad ; \quad i = 1, 2, 3, \cdots$$

For $L = 17$ cm, we compute the resonance frequencies

$$f_i = \frac{c}{\lambda_i} = (2i-1) \cdot \frac{c}{4L} \in \{500, 1500, 2500, 3500, \cdots\} \text{ Hz}.$$

where the sound velocity $c = 340$ m/sec.

proof: $$\frac{34000}{4 \times 17} = \frac{1000}{2} = 500$$

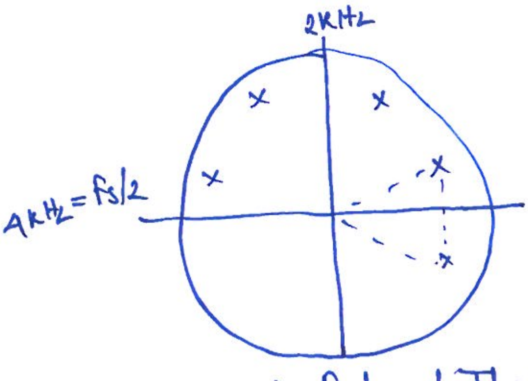
$$\therefore f_i = (2i-1) \times 500$$

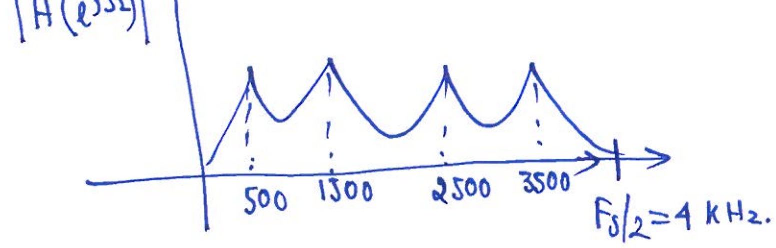
- If sampling frequency of speech signal $F_s = 8$ KHz (telephone quality).

then $\frac{F_s}{2} = 4$ KHz.

Now for each resonance, we require one pole-pair.

So, in total 4 pole-pair, leading to filter order $p = 8$.





- Rule of Thumb $\longrightarrow$ One resonance per KHz.
- To be safe side $\longrightarrow$ use $p = 10$.

Spectral Transformations

- Spectral Transformations are important. But why? The purpose of spectral transformations is to represent a signal in a domain where 'analysis and interpretation' becomes more effective. For example, certain signal properties becomes more visible or accessible.

Fourier Transform (FT) of Continuous Signals

The FT provides Spectra.

$$X_a(j\omega) = \int_{-\infty}^{+\infty} x_a(t)\, e^{-j\omega t}\, dt$$ where $\omega = 2\pi f$ in radian

Inverse FT (IFT) is

$$x_a(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X_a(j\omega)\, e^{j\omega t}\, d\omega$$

The FT converges in the mean square if (energy sense)

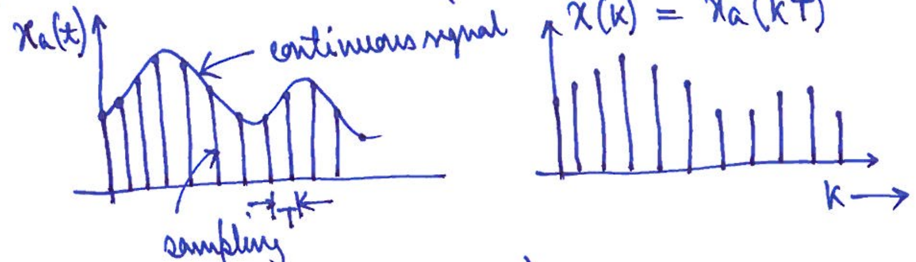
$$\int_{-\infty}^{+\infty} |x_a(t)|^2\, dt < \infty$$

	Property	Time domain	Frequency Domain
	Definition	$x(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X(j\omega)\, e^{j\omega t}\, d\omega$	$X(j\omega) = \int_{-\infty}^{+\infty} x(t)\, e^{-j\omega t}\, dt$
prove:	Linearity	$a x_1(t) + b x_2(t)$	$a X_1(j\omega) + b X_2(j\omega)$
	conjugation	$x^*(t)$	$X^*(-j\omega)$
	convolution	$x_1(t) * x_2(t)$	$X_1(j\omega) \cdot X_2(j\omega)$
prove:	Time shift	$x(t-t_0)$	$e^{-j\omega t_0} X(j\omega)$
	Modulation	$x(t)\, e^{j\omega_m t}$	$X(j(\omega-\omega_M))$
	Scaling	$x(at)$, $a \in R$, $a \neq 0$	$\frac{1}{\|a\|} X\left(j\frac{\omega}{a}\right)$

Fourier Transform of Discrete Signals

- This is well known as "Discrete Time Fourier Transform" (DTFT)

- Here we ~~normalize~~ deal with a discrete signal which is a sampled version of a continuous signal.



(sampling frequency $= f_s$)

$f \longrightarrow$ normal frequency

$f_s \longrightarrow$ sampling frequency

$T \rightarrow$ sampling period

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

normalized radian frequency $= \omega T = 2\pi f T = 2\pi \frac{f}{f_s}$.

- DTFT Definition

$$X(e^{j\Omega}) = \sum_{k=-\infty}^{+\infty} x(k)\, e^{-j\Omega k}$$

and $$x(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\Omega})\, e^{j\Omega k}\, d\Omega$$

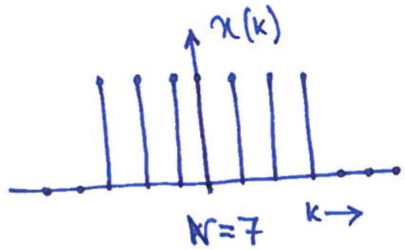
$$X(e^{j\Omega}) = |X(e^{j\Omega})|\, e^{j\phi(\Omega)}$$

$e^{j\phi(\Omega)} \leftarrow$ phase spectrum

$|X(e^{j\Omega})| \downarrow$ amplitude spectrum

Example: Consider a discrete rectangular pulse

$$x(k) = \sum_{l=-(N-1)/2}^{(N-1)/2} \delta(k-l)$$



$$\delta(k) = \begin{cases} 1 & k=0 \\ 0 & k \neq 0 \end{cases}$$

$$X(e^{j\Omega}) = \sum_{k=-\infty}^{+\infty} x(k)\, e^{-j\Omega k}$$

$$= \sum_{k=-(N-1)/2}^{(N-1)/2} e^{-j\Omega k} = e^{j\Omega (N-1)/2} \sum_{k'=0}^{N-1} e^{-j\Omega k'}$$

$$= e^{j\Omega (N-1)/2} \cdot \frac{\left(e^{-j\Omega}\right)^{N} - 1}{e^{-j\Omega} - 1}$$

$$= \frac{e^{j\Omega \frac{N}{2}}}{e^{j\Omega \cdot \frac{1}{2}}} \cdot \frac{\cancel{\left(e^{-j\Omega}\right)}\, e^{-j\Omega N} - 1}{e^{-j\Omega} - 1}$$

$$= \frac{e^{-j\Omega \frac{N}{2}} - e^{j\Omega \frac{N}{2}}}{e^{-j\frac{\Omega}{2}} - e^{+j\frac{\Omega}{2}}}$$

$$= \frac{e^{j\Omega \frac{N}{2}} - e^{-j\Omega \frac{N}{2}}}{e^{j\frac{\Omega}{2}} - e^{-j\frac{\Omega}{2}}} = \frac{\sin(\Omega N/2)}{\sin(\Omega/2)}$$

$$x(k) \xrightarrow{\text{DTFT}} X(e^{j\Omega})$$

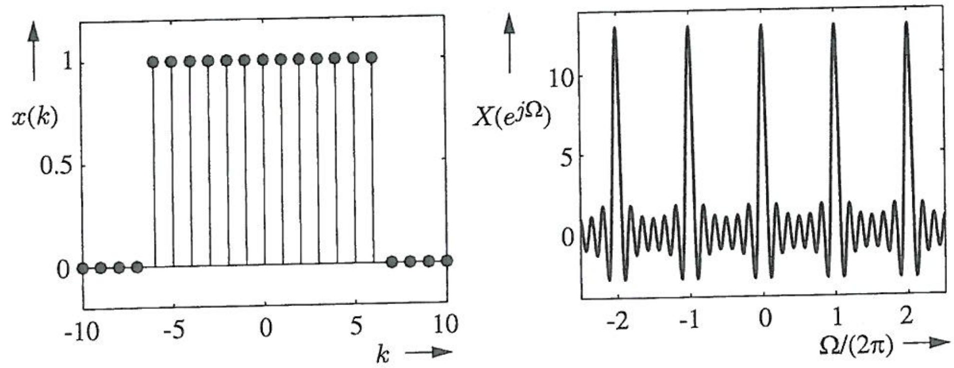


Figure 3.1: FT of a discrete rectangular pulse ($N = 13$)

DTFT of a rectangular pulse

Property	Time Domain	Frequency Domain
Definition	$x(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\Omega})e^{j\Omega k}\,d\Omega$	$X(e^{j\omega}) = \sum_{k=-\infty}^{+\infty} x(k)e^{-j\Omega k}$
Linearity	$a x_1(k) + b x_2(k)$	$a X_1(e^{j\Omega}) + b X_2(e^{j\Omega})$
Conjugation	$x^*(k)$	$X^*(e^{-j\Omega})$
prove Convolution	$x_1(k) * x_2(k)$	$X_1(e^{j\Omega}) \cdot X_2(e^{j\Omega})$
Time Shift	$x(k-k_0)$	$e^{-j\Omega k_0} X(e^{j\Omega})$
Modulation	$x(k)e^{j\Omega_M k}$	$X(e^{j(\Omega-\Omega_M)})$