Spoken Lang. Processing

Midterm Exam – Fall 2017

Question 1:

(a)

/k/ plosive, Un-Voiced
/sh/ Fricative, Un-Voiced
/AH/ Vowel, Voiced
/M/ Nasal, Voiced

(b)

* Plosives (stops) e.g /k/: Very little power (or no power) followed by sudden explosion and aspiration when constriction is released.
Waveform e.g. 

* Fricatives (e.g. /sh/): Non-periodic sound (friction) - random energy across wide freq. range.

Waveform e.g. 

* Vowels (e.g. /AH/): periodic waveform with fundamental frequency corresponding to vibration rate of vocal-folds.
Vocal Tract enhances some frequencies (Formants) and suppresses others.
Energy of Vowels Sounds is relativily higher than other Sounds.

Waveform e.g. 

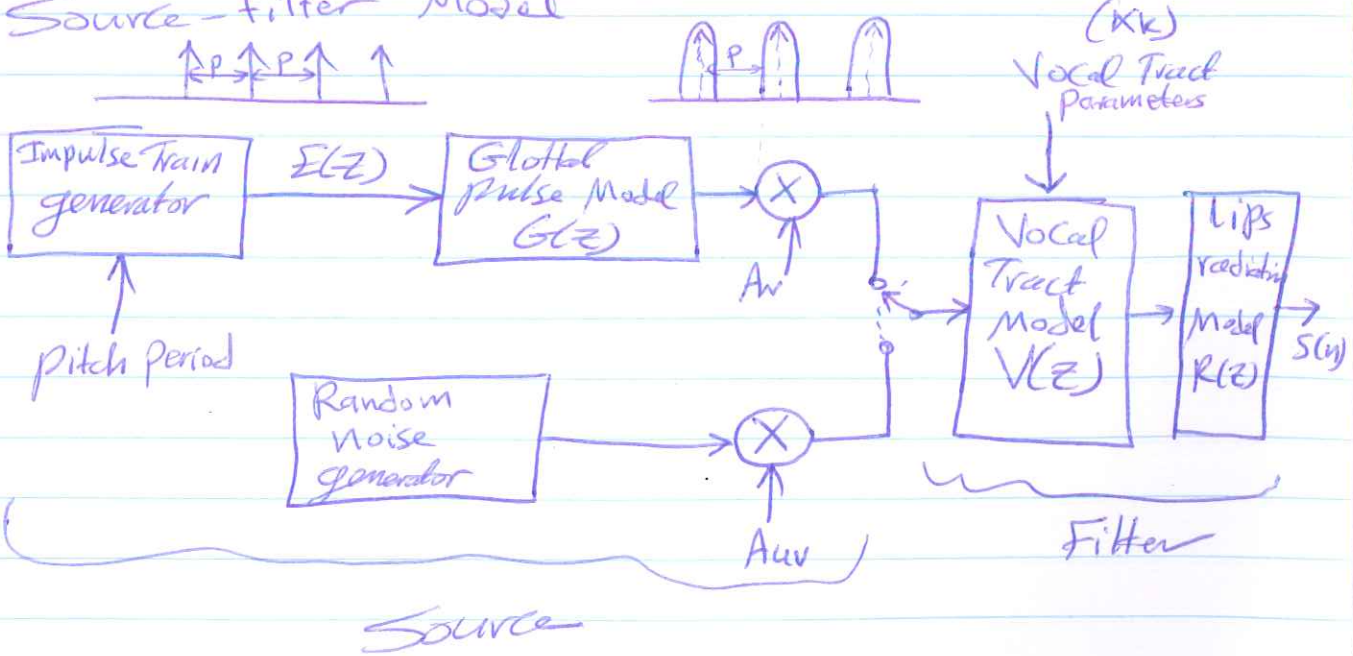
Q1

(c) Spectrogram: Time-frequency representation of speech signal
It is a concatenated spectrums of short frames (10-20 ms) and shows the energy distribution over frequency and Time scale.

	Window Length	Freq. resolution	Temporal Resolution
Narrow-Band	Wide	Fine	poor
Wide-Band	Narrow	poor	Fine

Question 2:

(a) Source-filter Model



* Impulse Train generator corresponds to Vocal folds Vibration with frequency of f_0 (fundamental) For Voiced Sounds

* Glottal Model: Corresponds to "opening / closing" of the Vocal Folds in the larynx which shaps the impulses to be like pulses (puffs)

* for Un-Voiced Sounds (eg. /sh/) the output of the larynx is noise-like signal. This is modeled by a random noise generator.

* In general, Voiced Sounds have greater amplitude than Un-Voiced. This is represented by Multiplying output of larynx, by two constants A_V (Voiced) and A_{UV} (Un-voiced)
$A_V > A_{UV}$.

* Vocal Tract is modeled by a filter with sufficient number of poles (All-pole filter). Usually 10-14 poles for both Voiced and Un-Voiced Sounds.

* Lip radiation is modeled by a high-pass filter of a transfer function $R(z)$.

((b)) Impulse Train: $E(z) = \frac{1}{1 - z^{P}}$, P: pitch period.

Glottal model:

$$G(z) = \begin{cases} \frac{1}{(1 - e^{-\frac{1}{T}} z^{-1})^2} \approx \frac{1}{(1 - z^{-1})^2}, & \text{for Voiced Sounds} \\ 1, & \text{for Un-Voiced Sounds} \end{cases}$$

Vocal Tract model:

$$V(z) = \begin{cases} \dfrac{1}{1 + \sum_{k=1}^{P} a_k z^{-k}}, & \text{for Voiced sound} \\ \dfrac{1}{1 + \sum_{k=1}^{P+2L} a_k z^{-k}}, & \text{for Un-Voiced sound} \end{cases}$$

P = no. of poles
L = no. of zeros.

OR

$$V(z) = \frac{1}{1 + \sum_{k=1}^{P} a_k' z^{-k}}, \quad P \tilde{=} 10 - 14$$

* For Both Voiced and Un-voiced Speech.

Lips radiation:

$$R(z) = 1 - z^{-1}, \quad \text{first-order HPf.}$$

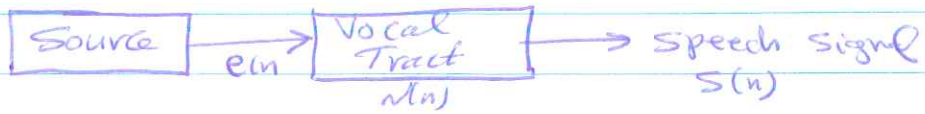
* overall

$$\frac{S(z)}{E(z)} = \frac{G}{1 + \sum_{k=1}^{q} a_k' z^{-k}}, \quad 10 \leq q \leq 14$$

Question 2

(c) Right-hand graph is spectrum of speech frame
Left-hand graph is the response of the vocal tract

* Left-hand graph (Vocal tract) is the envelop of the speech spectrum, or low frequency Componets of the speech Spectrum.



$$S(n) = e(n) * v(n)$$
$$S(z) = E(z) \cdot V(z)$$

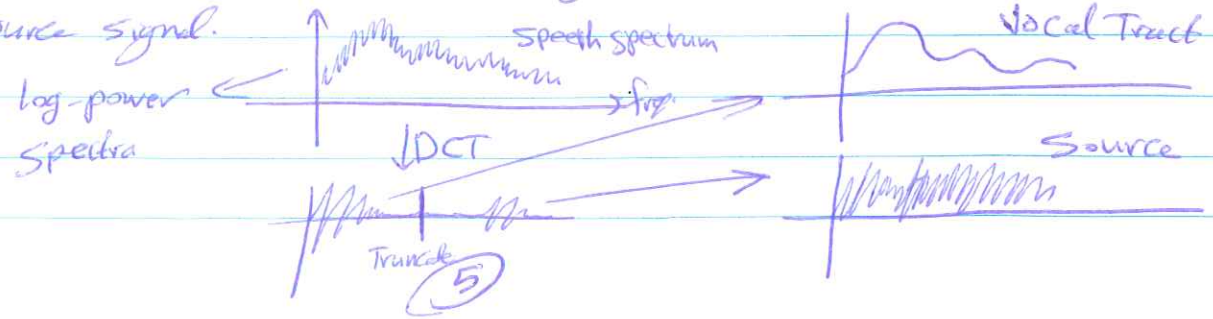
* This looks like a voiced sound because the power is highier in the low-frequency components i.e. power at F_1 > power at F_2 > ...

(d) Two benefits of using DCT in the MFCC technique.

1. Un-correlate features, so diagonal Covariance Matrix Can be used in the Modelling.

2. Separate source signal (excitation) from the vocal tract filter response by truncating DCT coefficients.

The low order coefficients (≈ first 12) Corresponds to Vocal tract filter response, and high order coefficients Corresponds to Source signal.



Question 3

(a) $F_s = 600$ sample/sec

$s(n) = \{-0.35, 0.0, 3.15, -2.5, -3.54, 2.8, 0.0, -0.28\}$, $N = 8$

(i) $\text{Energy} = \frac{1}{N}\sum_{n=0}^{N-1} s(n)^2 = \frac{1}{8}\sum_{n=0}^{7} s(n)^2$

$= \frac{1}{8}\left[(-0.35)^2 + (0)^2 + (3.15)^2 + (-2.5)^2 + (-3.5)^2 + (2.8)^2 + (0)^2 + \right.$

$\left. + (-0.28)^2\right] = 36.4634$

$E_{dB} = 10 \log_{10} E = 15.618$ dB.

(ii) $ZCC = \frac{1}{N}\sum_{n=0}^{N-1} \frac{1}{2}\left|\text{sign}(s(n)) - \text{sign}(s(n-1))\right|$

$= \frac{1}{8}\cdot(4) = \frac{1}{2}$

(iii)

$R(k) = \frac{1}{N}\sum_{n=0}^{N-1} s(n)\,s(n-k)$

$150 \le F_0 \le 300 \Rightarrow \frac{600}{300} \le k \le \frac{600}{150} \Rightarrow 2 \le k \le 4$

We compute $R(2)$, $R(3)$ and $R(4)$ $\Rightarrow$ The pitch period corresponds to k which gives $R(k)$ max. value.

Q3. (a) (cii) Continue - - -

$s(n)$	-0.35	0	3.15	-2.5	-3.54	2.8	0	-0.28
$s(n-2)$	0	0	-0.35	0	3.15	-2.5	-3.54	2.8
$s(n-3)$	0	0	0	-0.35	0	3.15	-2.5	-3.54
$s(n-4)$	0	0	0	0	-0.35	0	3.15	-2.5

$R(2) = \frac{1}{8}(3.15)(-0.35) + (3.15)(-3.54) + (2.8)(-2.5) + (2.8)(-0.28) = -2.673$

$R(3) = \frac{1}{8}(-2.5)(-0.35) + (2.8)(3.15) + (-0.28)(-3.54) = 1.335$

$R(4) = \frac{1}{8}(-3.54)(-0.35) + (-0.28)(-2.5) = 0.2425$

So, $R(3)$ is the max. value $\Rightarrow k=3$ corresponds to pitch period.

$\Rightarrow$ pitch period $= 3 * \frac{1}{600} = \frac{1}{200}$ sec. $\Rightarrow f_0 = 200$ Hz.

Q3 (b)

High energy + low zcc + nonzero $f_0 \Rightarrow$ Voiced speech

Low energy + High zcc + zero $f_0 \Rightarrow$ Unvoiced speech

This speech segment is voiced because Fundamental frequency (f_0) is not zero ($f_0 = 200$ Hz).

Question 3 (c)

(1) No, because min. Sound pressure level for 100Hz is around 25 dB

(2) a tone of 1000 Hz at 20 dB because human ear is more sensitive to 1000 Hz than 100 Hz.

(3) B 2 kHz – 5 kHz.