SAL Assignment-2

① Language to analyzed: Hindi

Word	Variant	ARPAbet Transcription	Stress Marking
शहर	1	SH AH0 H ER1	Unstressed in variant 1 (SH AH0) (HER1) Stressed in variant 1
	2	SH UH1 H ER0	
काम	1	K AA1 M	KAA1 (Stressed in variant 1)
	2	K AH0 M	
दोस्त	1	D OW1 S T	DOW1 (Stressed in variant 1)
	2	D OW0 S T	
सड़क	1	S AH0 D AH0 K	SAH0 (unstressed in variant 1)
	2	S AH1 D AH0 K	D AH0 (unstressed in both variants)
दिल्ली	1	D IH1 L L IY0	D IH1 (Stressed in variant 1)
	2	D IH0 L L IY1	L IY0 (unstressed in variant 1)

② Pronunciation by Tamil, Bengali and Malyalam Speakers-

a) खुश (Khush) → Happy

i) Tamil - /Kush/
- As tamil speakers lacks the aspirated sounds. So, aspiration is not observed.

ii) Bengali - /K^{h}US/
- Aspiration will be in first sound but removed in second sound

iii) Malyalam - /K^{h}us/
- Aspiration in the first sound but not present in second sound

b) ध्यान (dhyān) → attention

i) Tamil - /TAYAN/
- As absence of nasalization and retroflex & aspirated sound gets substituted

ii) Bengali - /dyan/
- They lacks 'dh' sound, which they replaced with 'd'.

iii) Malyalam - /dyan/
- They usually have absence of nasalization

c) रज़ाई (razāi) → quilt

i) Tamil - razai
- Nasalization is reduced in vowels

ii) Bengali - rajāi
- They pronunced it as same

iii) Malyalam - rayai
- Nasalization is possible in vowel

(4) Sentene - "The cat sat on the mat"

1) Unigrams - the, cat, sat, on, the, mat
Bigrams - the cat, cat sat, sat on, on the, the mat
Trigrams - the cat sat, cat sat on, sat on the, on the mat

2) Frequenies

a)

Unigrams	Frequency
the	2
cat	1
sat	1
on	1
mat	1

b)

Bigrams	Frequency
the cat	1
cat sat	1
sat on	1
on the	1
the mat	1

c)

Trigrams	Frequency
the cat sat	1
cat sat on	1
sat on the	1
on the mat	1

3) Probability is calculated as -
Unigram prob. = freq. of unigram / total no. of unigram
Bigram prob. = freq. of bigram / freq. of 1st word in bigram
Trigram prob. = freq. of trigram / freq. of first two word in the trigram

a)

Unigram	freq.	Probability
the	2	2/6 = 0.333
cat	1	1/6 = 0.167
sat	1	1/6 = 0.167
on	1	1/6 = 0.167
mat	1	1/6 = 0.167

b)

Bigram	frequency (1st)	Probability
the cat	2	1/2 = 0.5
cat sat	1	1/1 = 1
sat on	1	1/1 = 1
on the	1	1/1 = 1
the mat	2	1/2 = 0.5

c)

Trigram	Freq. (First two words)	Probability
the cat sat	1	1
cat sat on	1	1
sat on the	1	1
on the mat	1	1

4) Here we can see that the repetition of word "the" increases its unigram probability as compared to the other unigrams. In Bigrams, "the" splits its frequency between "the cat" and "the mat" which reduces its individual probability. In 'trigrams', the repetition in two different trigrams "the cat sat" and "the mat" have probability to 1, as both occur exactly once. So, the repetition shifts the overall distribution making sequence involving "the" more probable (in longer texts)

(5) States - |g|, |un|, |d|

Transition probability

	|g|	|un|	|d|
|g|	0.5	0.4	0.1
|un|	0.2	0.6	0.2
|d|	0.1	0.3	0.6

Posterior Probability

	1	2	3	4	5	6	7	8	9	10
P(g|O)	0.7	0.6	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1
P(un|O)	0.2	0.1	0.8	0.7	0.6	0.8	0.7	0.3	0.1	0.1
P(d|O)	0.1	0.3	0.1	0.2	0.3	0.1	0.2	0.6	0.7	0.8

To find best sequence we use viterbi algorithm ~~with~~:

~~Read probabilities for all states~~

1) Initialise first column using initial probability = Posterior prob.

2) For each observation, we compute max prob. of reaching each state from each previous state using value of previous state & transmission probability and posterior probability.

val(S, t) = max (val(S-1, t-1) x trans(S-1, S) x posterior(S))

3) In the end, select the max value from last column to get the sequence

Path Probabilities for all states

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
|g|	0.7	0.21	0.0105	0.00134	5.64×10^{-4}	2.08×10^{-4}	9.75×10^{-5}	4.1×10^{-5}	1.47×10^{-5}	1.03×10^{-6}
|un|	0.2	0.028	0.0672	0.0282	1.02×10^{-3}	4.86×10^{-3}	2.05×10^{-3}	3.69×10^{-4}	2.21×10^{-5}	3.1×10^{-6}
|d|	0.1	0.021	0.0021	0.00269	1.69×10^{-3}	2.03×10^{-3}	1.95×10^{-4}	2.64×10^{-4}	1.03×10^{-4}	4.96×10^{-5}

∴ Sequence is - |g|, |g|, |un|, |un|, |un|, |un|, |un|, |d|, |d|, |d|

(6) The phoneme sequence in the phase "Speech Analysis and Linguistics" /S/, /P/, /IY/, /CH/, /AE/, /N/, /D/.

/AH/, /N/, /AE/, /K/, /IX/, /S/, /IH/, /S/, /N/, /IH/, /NG/, /G/, /W/, /UH/, /IH/, /S/, /T/, /IH/, /K/, /S/

⇒ Articulatory features of consonants

Phoneme	Place	Manner	Voicing
/S/	Alveolar	Fricatives	Unvoiced
/P/	Bilabial	Plosives	Unvoiced
/CH/	Alveolar	Affricatives	Unvoiced
/N/	Alveolar	Nasal	Voiced
/K/	Alveolar	Lateral	Voiced
/D/	Alveolar	Plosive	Unvoiced
/NG/	Velar	Nasal	Voiced
/G/	Velar	Plosive	Voiced
/W/	Labio-dental	Approximant	Voiced
/T/	Alveolar	Plosive	Unvoiced
/K/	Velar	Plosive	Unvoiced