Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language

Presenter:

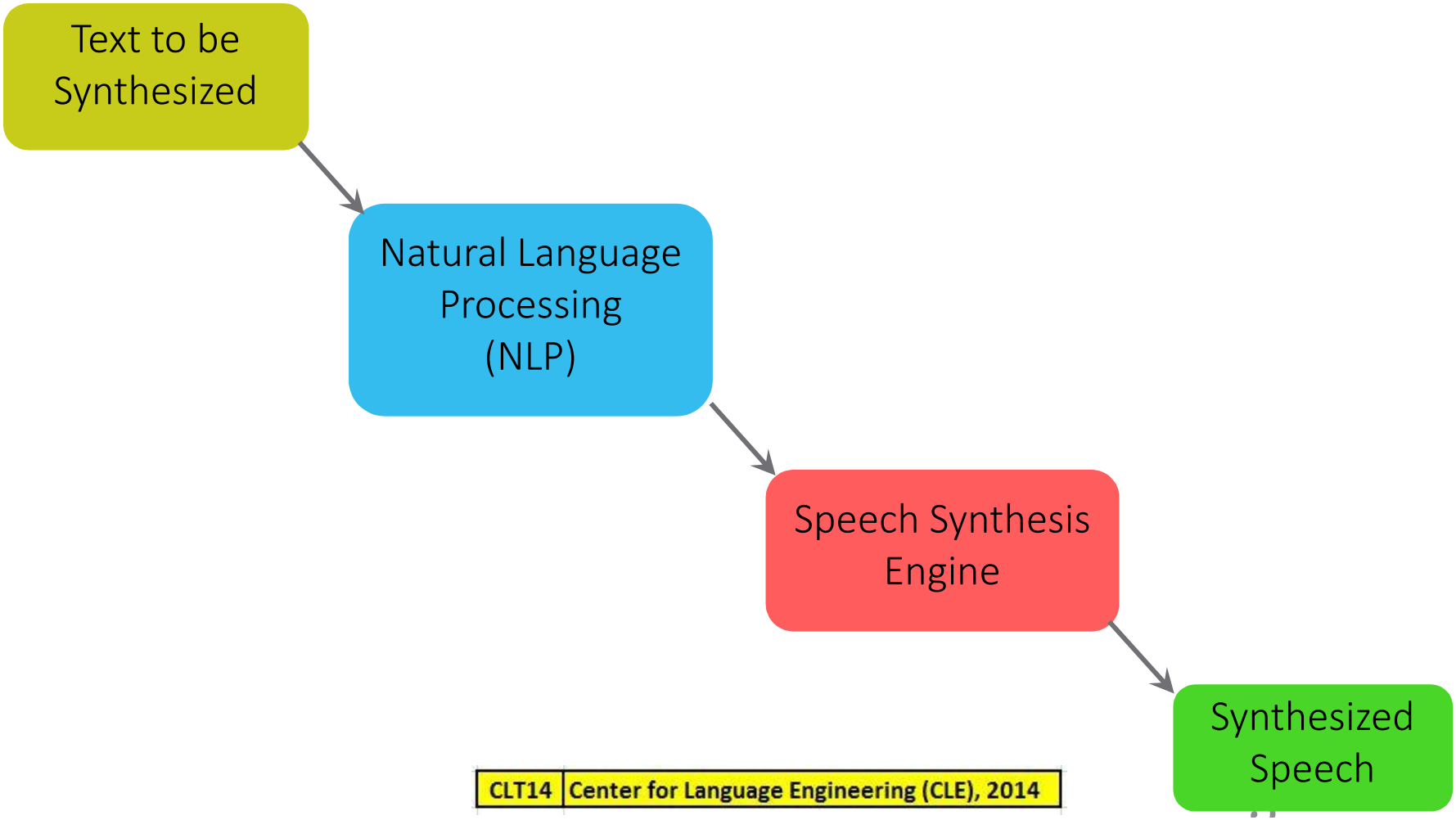
Dr. Tania Habib

Outline:

- Overview
- Unit selection vs HMM based Speech Synthesis (HTS) [1]
- Development
- Requirements for Voice building
- Data Set
- Challenges
- Subjective Evaluation
- Erroneous Words
- Summary

Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language

Speech Synthesis Overview:



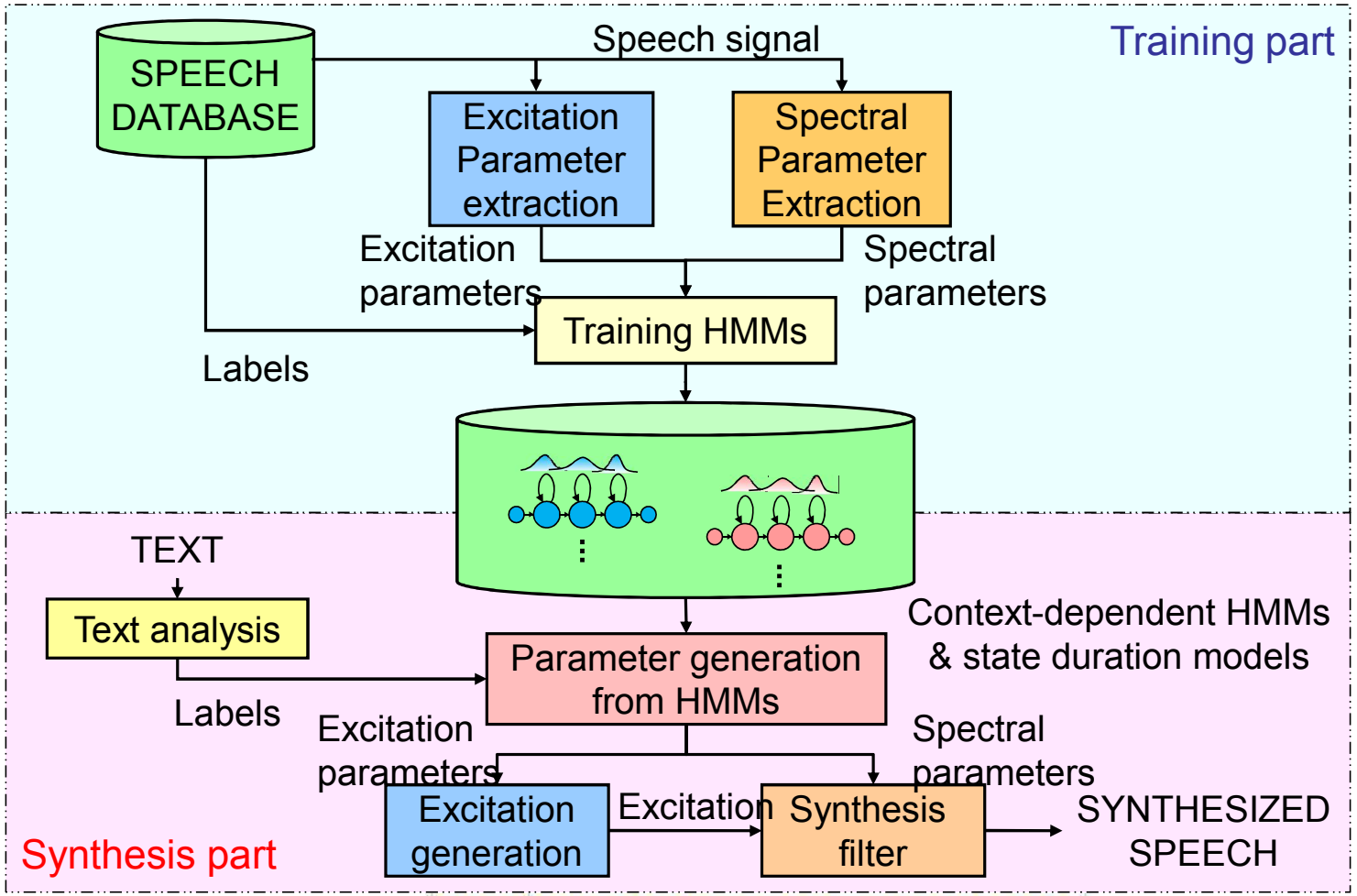
Types of Speech Synthesis:

- Rule-based, *formant synthesis*
  - Hand-crafting each phonetic units by rules
- Corpus based:
  - *Concatenative synthesis* (Unit Selection)
    - High quality speech can be synthesized using waveform concatenation algorithms.
    - To obtain various voices, a large amount of speech data is necessary.
  - *Statistical parametric synthesis* (HMM based)
    - Generate speech parameters from statistical models
    - Voice quality can easily be changed by transforming HMM parameters.

Unit Selection vs. HTS

Unit Selection	HTS
Advantages:	
High Quality at Waveform level (Specific Domain)	• Small Foot Print • Smooth • Stable Quality
Disadvantages:	
• Large footprints • Discontinuous • Unstable quality	Vocoder sound (Domain-independent)

HTS Overview:



[HTS Slides released by HTS Working Group slide no. 21]

Preliminary requirements for the HTS toolkit:

1. Annotated Training data.
2. Define speech features (MFCC, F0 and duration) for model training.
3. Sorting out unique context-dependent as well as context-independent phonemes (from the training data) for model training.
4. Unified question file for spectral, F0 and duration for context clustering.

Data Set Used:

- Source:
  - Paragraphs taken from Urdu Qaida of Grade 2 and 4 respectively
- Duration :
  - 30 minutes
- Total number of utterances:
  - 347
- Recording parameters:
  - Sample rate : 8KHz (up-sampled to 48KHz)
  - Channel : Mono
  - Recording format: .WAV
  - Speaker: Native Urdu female speaker

Challenges:

- Generation of the full-context style labels.
- Addition of Prosodic Layers
  - Segment
  - Stress
  - Syllable
  - Word
- Unbalanced Training Data
- Defining the Question Set (Context Clustering)

Full-Context Format(1/2):

SIL^A-L+I_I=A@ 1_2/A:0_0_1/B:0-0-2@2-1&2-8#1-3$1-1!0-1;0-0|I_I/C:1+0+2/D:0_0/E:content+2@1+5&1+4#0+1/F:content_2/G:0_0/H:9=5^1=2|NONE/I:8=6/J:17+11-2

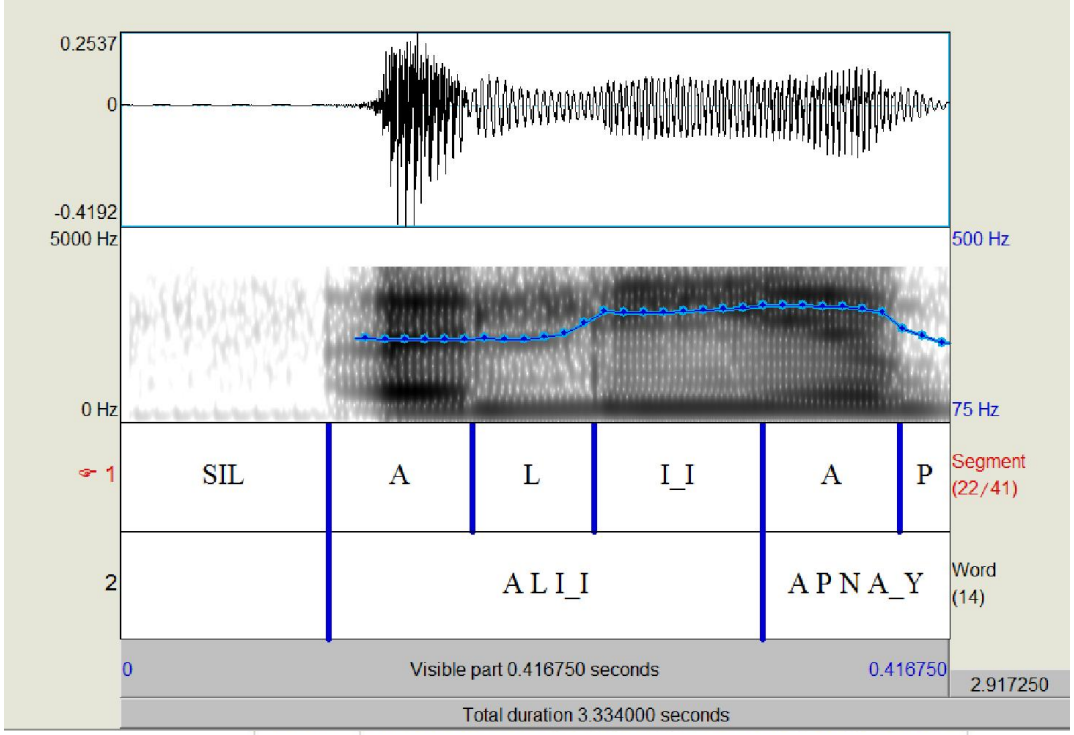
Segmental Context

Supra-Segmental Context

Segmental	Supra-Segmental
• Current Phoneme • Previous two Phonemes • Next two Phonemes	• Syllable • Stress • Word • Phrase • POS

Full-Context Format(2/2):

x^x-**SIL**+A=L@1_0/A:0_0_0/B:0-0-0@1-0&1-1#1-1$1-1!0-0;0- …
x^SIL-**A**+L=I_I@1_1/A:0_0_0/B:0-0-1@1-2&1-9#1-3$1-1!0-2;0- …
SIL^A-**L**+I_I=A@1_2/A:0_0_1/B:0-0-2@2-1&2-8#1-3$1-1!0-1;0-0 …
A^L-**I_I**+A=P@2_1/A:0_0_1/B:0-0-2@2-1&2-8#1-3$1-1!0-1;0- …



علی اپنے ۔۔۔

Questions on Segmental/Prosodic Layers:

Phoneme

- {preceding, succeeding} two phonemes
- current phoneme

Syllable

- # of phonemes at {preceding, current, succeeding} syllable
- {accent, stress} of {preceding, current, succeeding} syllable
- Position of current syllable in current word
- # of {preceding, succeeding} {accented, stressed} syllable in current phrase
- # of syllables {from previous, to next} {accented, stressed} syllable
- Vowel within current syllable

Word

- Part of speech of {preceding, current, succeeding} word
- # of syllables in {preceding, current, succeeding} word
- Position of current word in current phrase
- # of {preceding, succeeding} content words in current phrase
- # of words {from previous, to next} content word

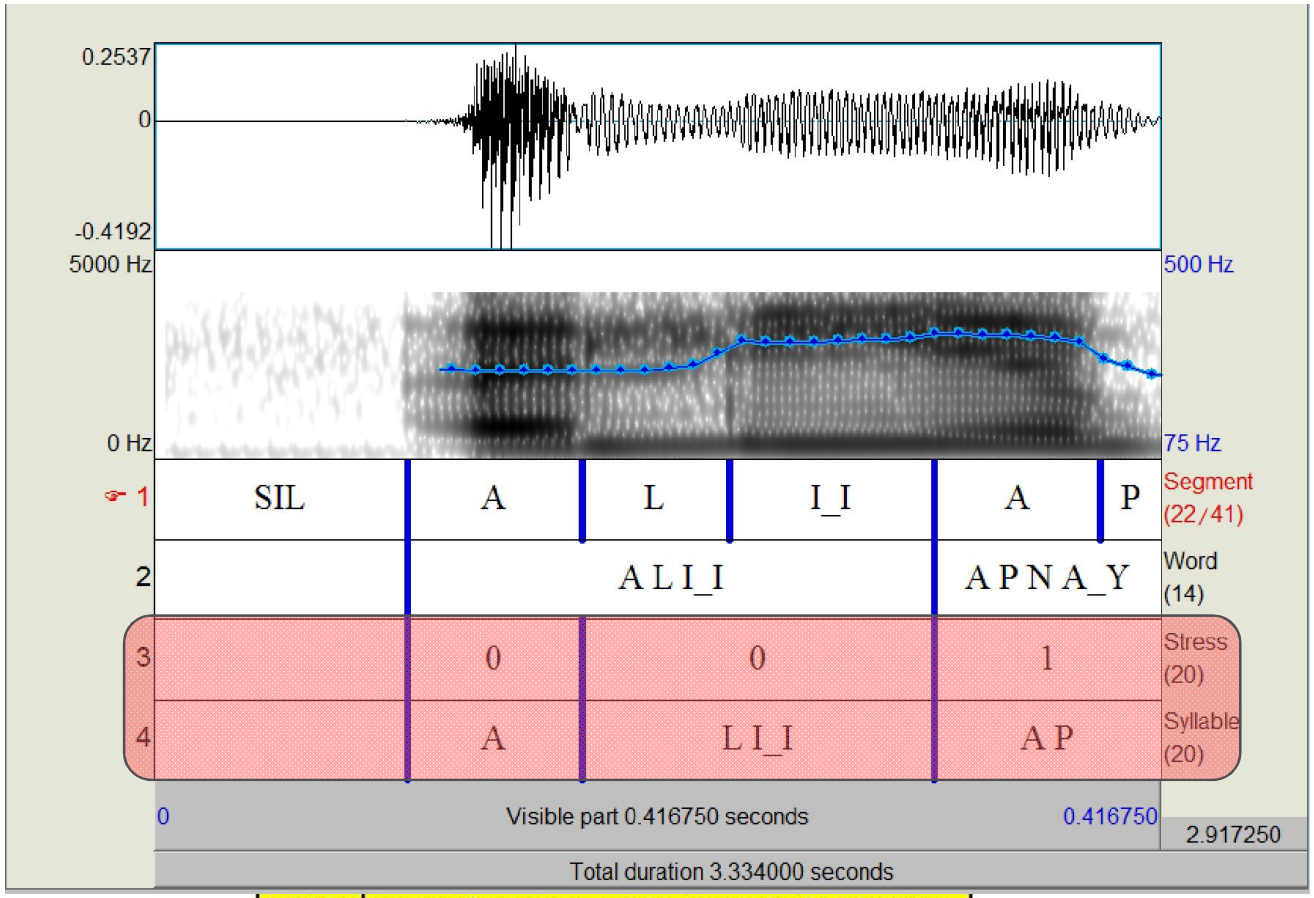
Phrase

- # of syllables in {preceding, current, succeeding} phrase

.....

Addition of Stress/Syllable Layer:

- Added layers:
  - Stress
  - Syllable



Unbalanced Training data:

- High occurrence for vowels
- Some of the phonemes were completely ignored
  - {J_H, L_H, M_H, N_G_H, R_H, Y, Z_Z} [2]

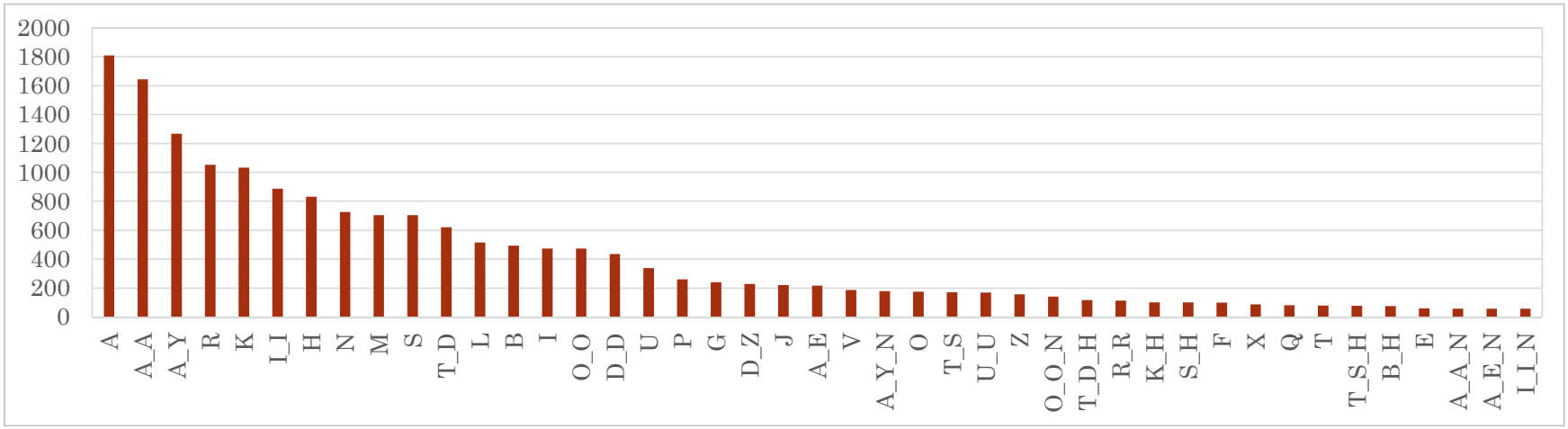


Figure. Phoneme Coverage for the 30-min speech data

Context Clustering (Question Set):

- Number of possible combinations are quite enormous with 53 different questions.
- Possible contexts = C^n

  where C = Total count of basic phonetic units,

  n = Total number of Questions

  - With only Segmental Context (n=5) Possible models are:

$$66^5 \approx 1252 \text{ million}$$

- If we consider all the context, it will be practically infinite.

Solution:

- Record data having maximum phoneme coverage at tri-phone or di-phone level.
- Apply context clustering technique to classify and share acoustically similar models

Subjective Evaluation:

- **Testing Methodology:**
  - Mean Opinion Score (MOS)[3] for:
    - Naturalness
    - Intelligibility
- **Naturalness:**

  How close it seems to be produced by a human?
- **Intelligibility:**

  How much conveniently the word was recognized?

Subjective Testing (Results):

Listener Type	MOS Naturalness	MOS Intelligibility
Technical 1	3.23	3.65
Linguistic 1	2.82	3.66
Linguistic 2	2.86	3.58
Linguistic 3	3.48	3.52

Table 1. Mean Opinion Score (MOS) results of four listeners

Erroneous words:

Nastalique Style	CISAMPA (Correct)	Listened (Incorrect)	Coverage (%)
طرف	T_DARAF	T_DALAF	5.92
گا	GA_A	D_DA_A	1.35
معلوم	MAYLU_UM	MAT_DLU_UM	0.00
تھے	T_D_HA_Y	T_SA_Y	0.66
رزی	RAZI_I	RAD_DI_I	0.88
کیونکہ	KIU_U_NKA_Y	T_SU_NKA_Y	0.15
حق	HAQ	HABS	0.46
بعد	BAYD_D	BAD_D	0.00
خیال	XAJA_AL	FIJA_AL	0.50

Table 2. Synthesized words with errors

Some Synthesized Examples:

	Synthesized:	Training Set:
Seen Context:		
Un-seen Context:		
Different Carrier Word:		

Summary:

- Text to Speech Synthesis (TTS):
  - Concatenative
  - Parametric (Hmm based)
- Requirement for Voice building
  - Annotated speech corpus
  - Speech features
  - Question file
- Challenges
  - Full context style labels
  - Addition of prosodic layers
  - Question file for context clustering
- Subjective Evaluation
- Erroneous words

References:

- 1. H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in proc. of Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany, August, 2007.
- 2. “IPA to CISAMPA Conversion Chart," Center for Language Engineering, UET, Lahore, [Online]. Available: http://www.cle.org.pk/resources/CISAMPA.pdf. [Accessed 3 March 2014].
- 3. M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," Computer Speech & Language, vol. 19, no. 1, pp. 55-83, 2005.