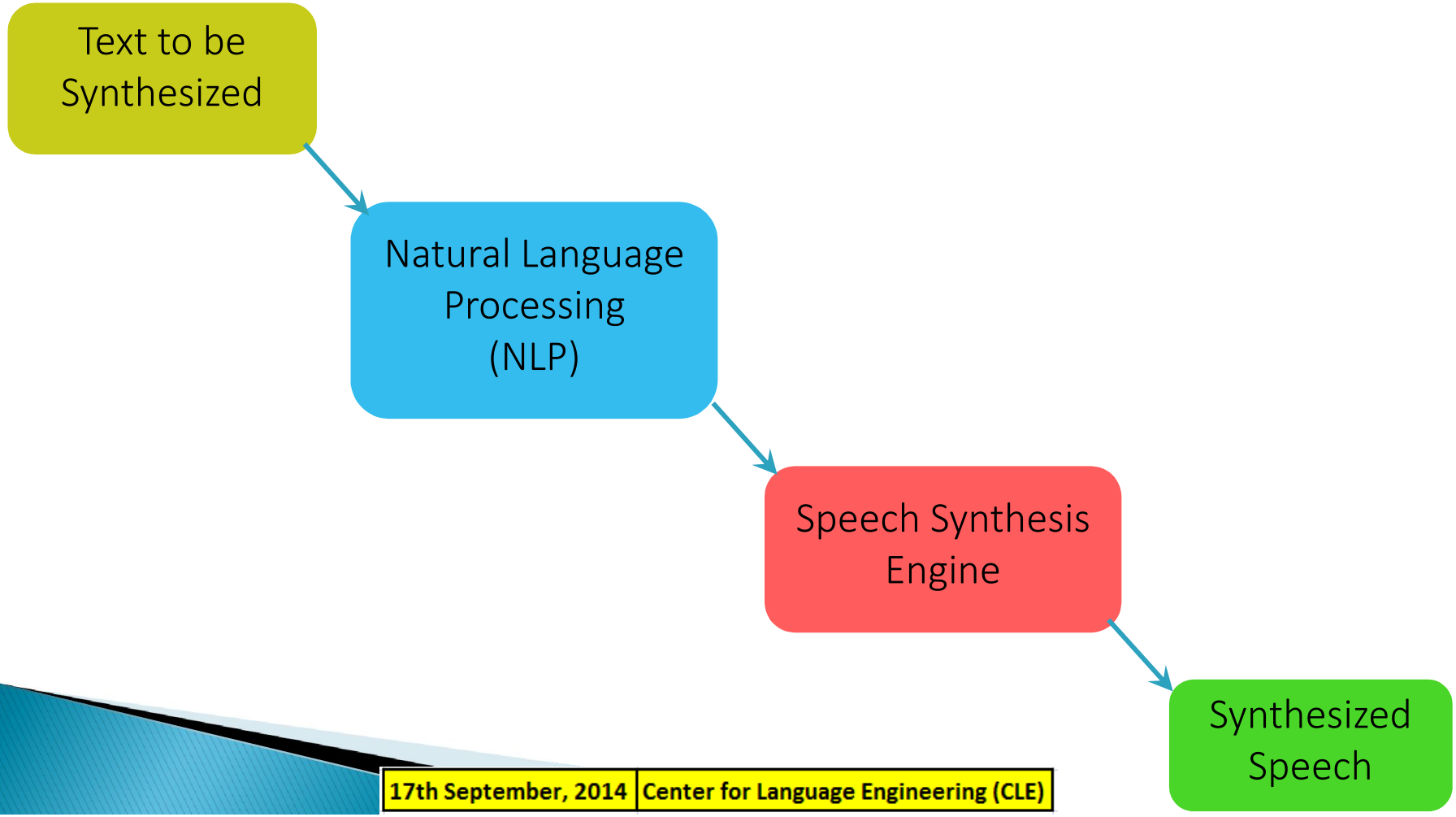
Hidden Markov Model (HMM) based Speech Synthesis using HTS Toolkit.

Presenter:
Omer Nawaz

17th September, 2014 | Center for Language Engineering (CLE)

Speech Synthesis Overview:



Introduction:

- Rule-based, *formant synthesis*
  - Hand-crafting each phonetic units by rules
- CORPUS-BASED:
  - *Concatenative synthesis*
    - 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*
    - Generate speech parameters from statistical models
    - Voice quality can easily be changed by transforming HMM parameters.

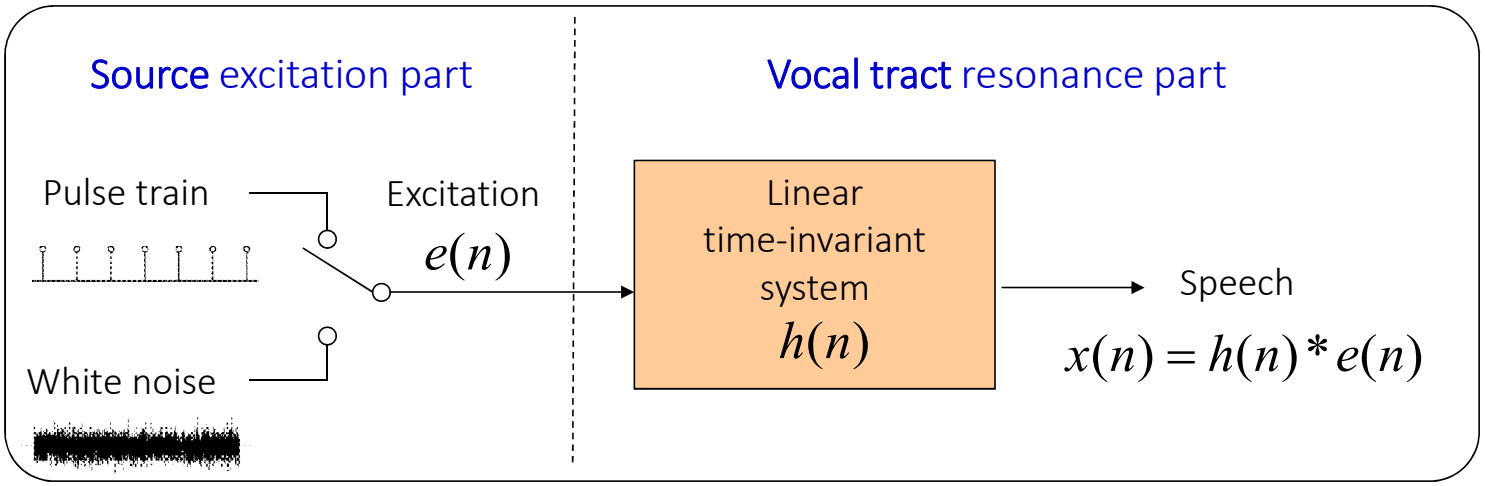
Approaches at CLE:

- CORPUS-BASED:
  - *Unit Selection*
  - *HMM based.*
- Comparison of two Approaches:

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

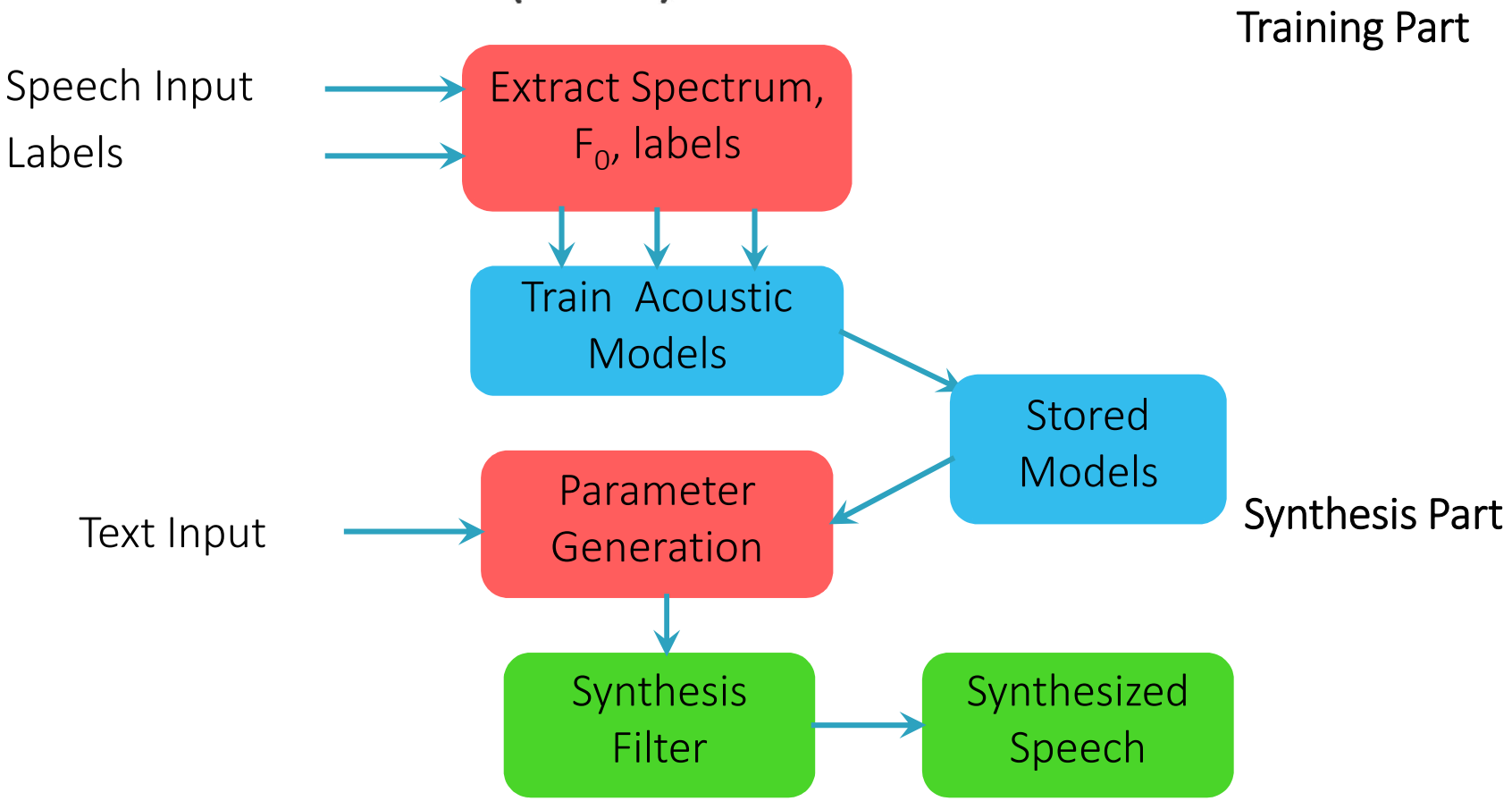
Synthesis Model:

Source Filter Model:



- The *h(n)* is defined by the state output vector of the HMM e.g mel-cepstrum

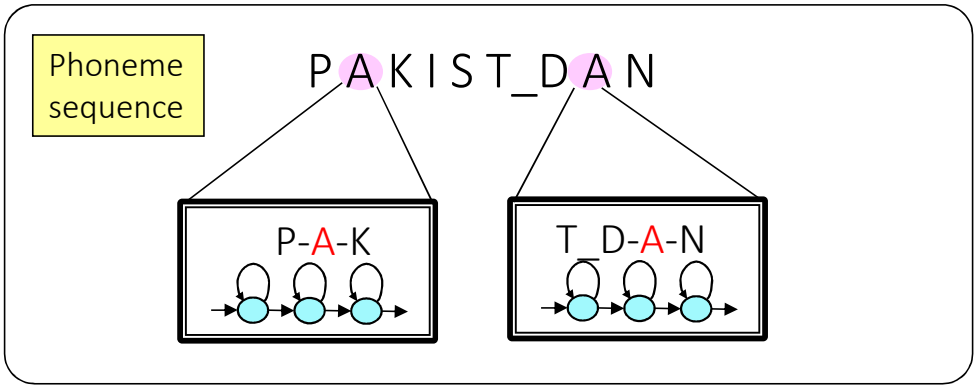
General Overview(HTS):



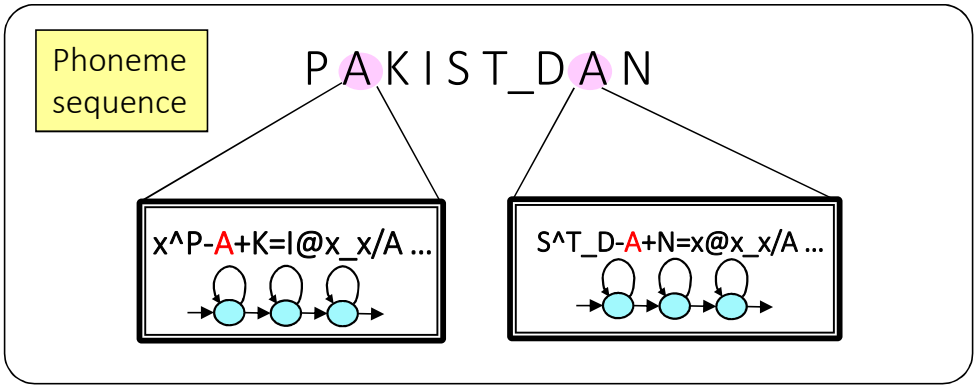
Challenges:

- Generation of the full-context style labels.
- Addition of Stress/Syllable Layer.
- Defining the Question Set.

Full-Context Label Style:



Tri-phone context dependent model



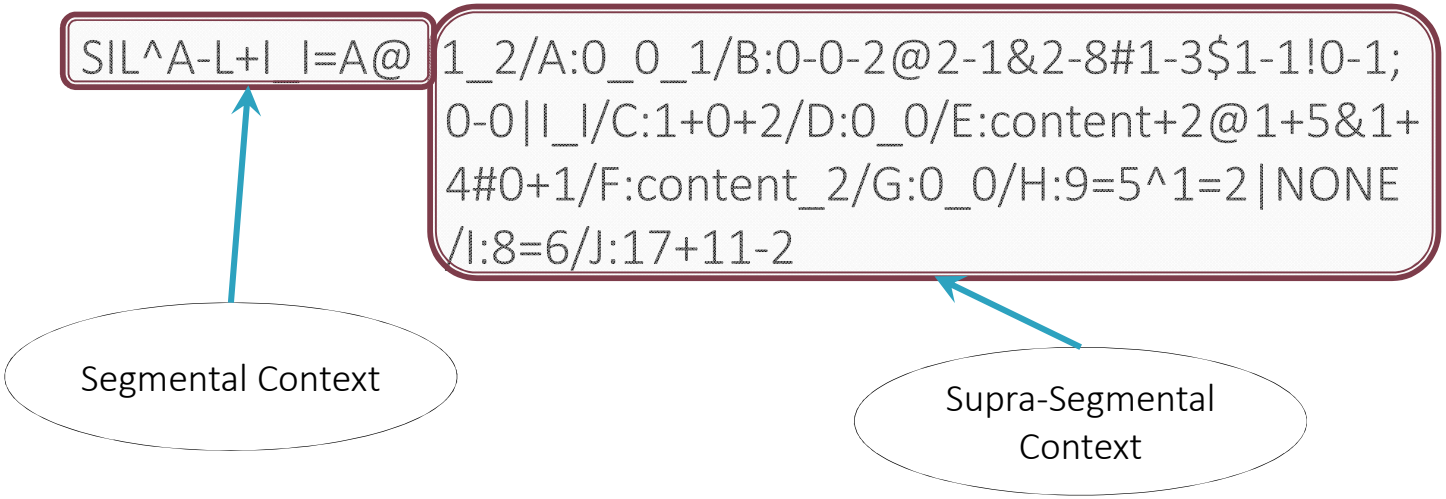
Full-context style context dependent model

Full-Context Format:

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- …

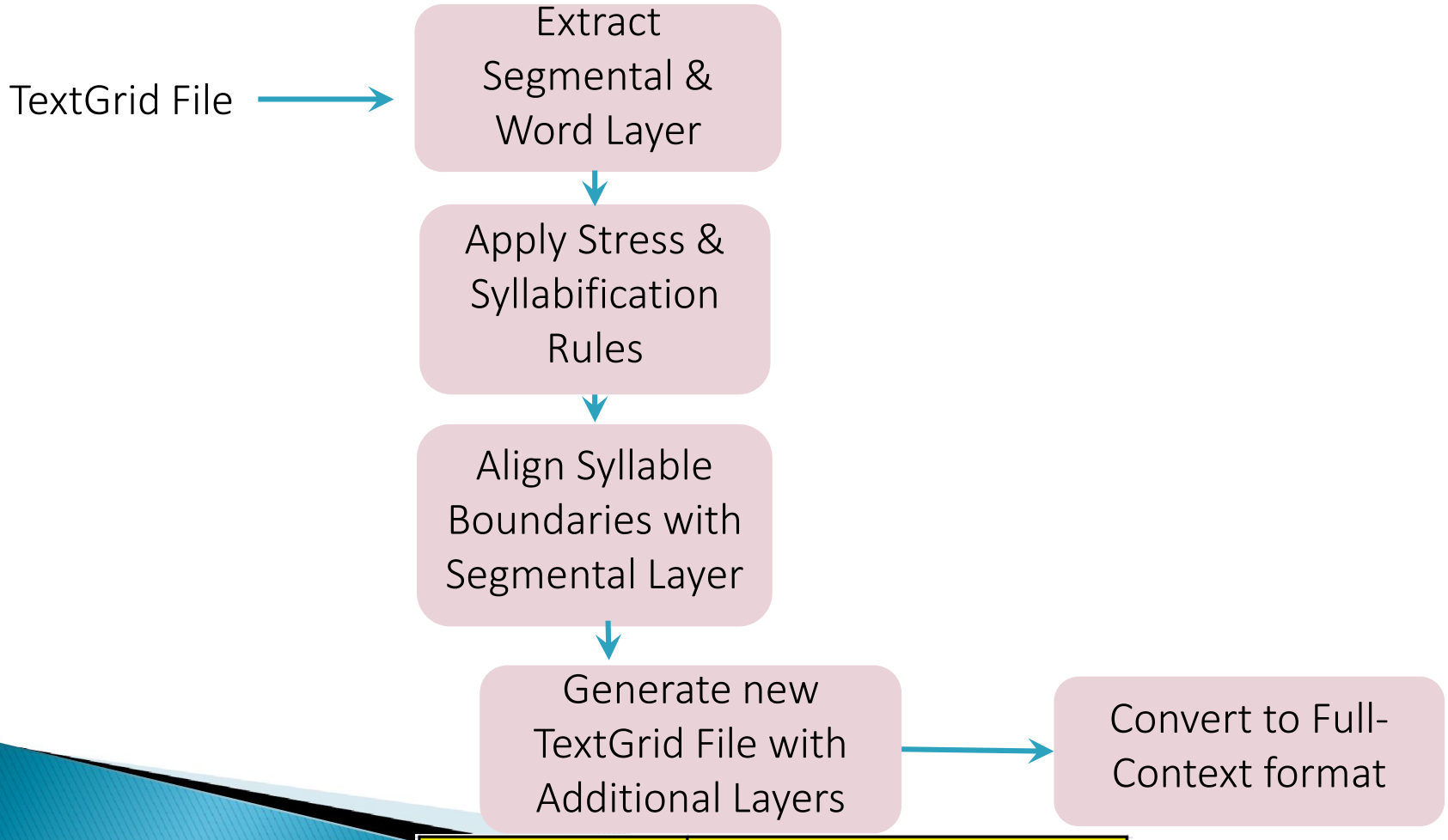
علی اپنے ۔۔۔

Full-Context Format:

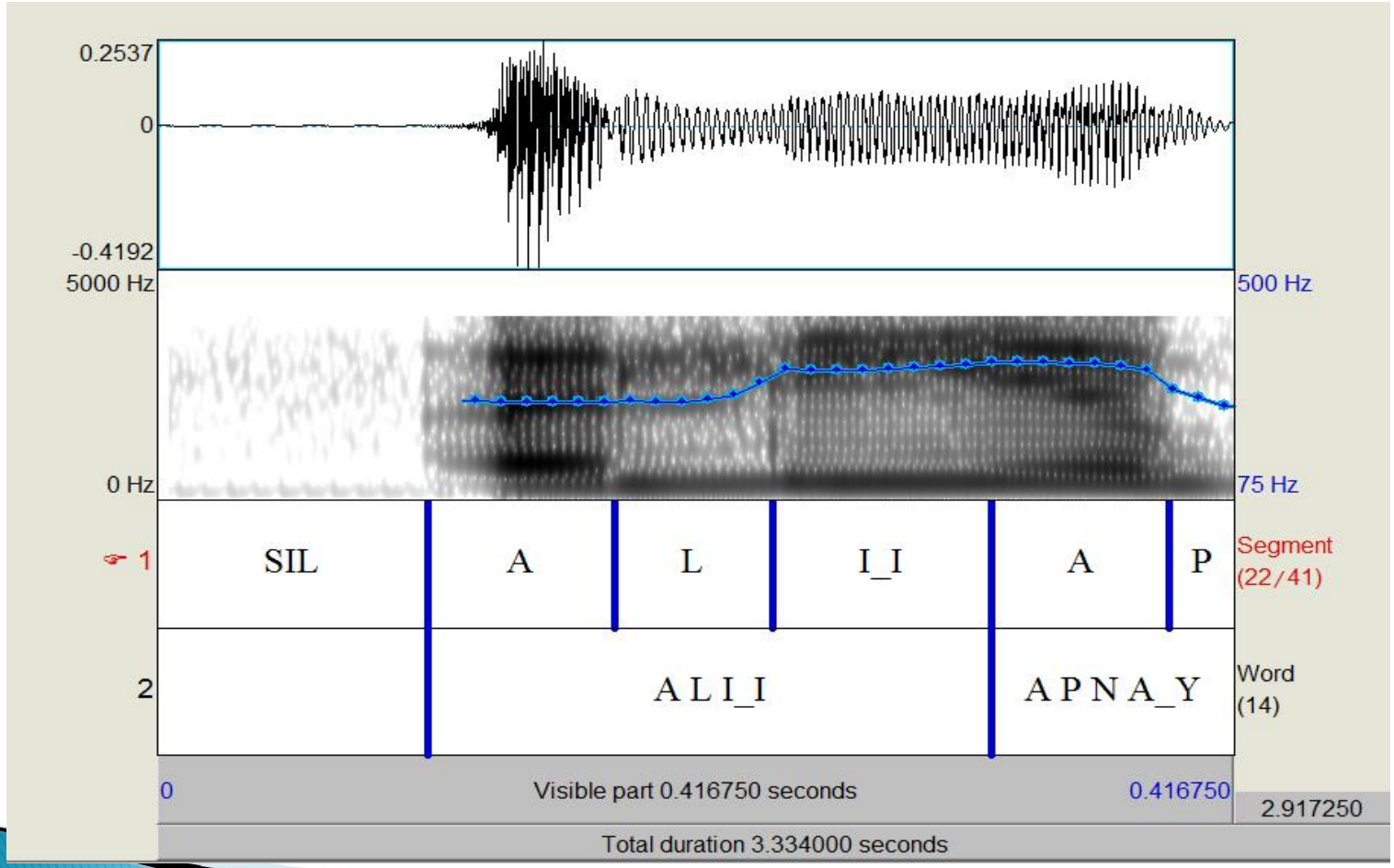


Segmental	Supra-Segmental
• Current Phoneme • Previous two Phonemes • Next two Phonemes	• Syllable • Stress • Word • Phrase • POS

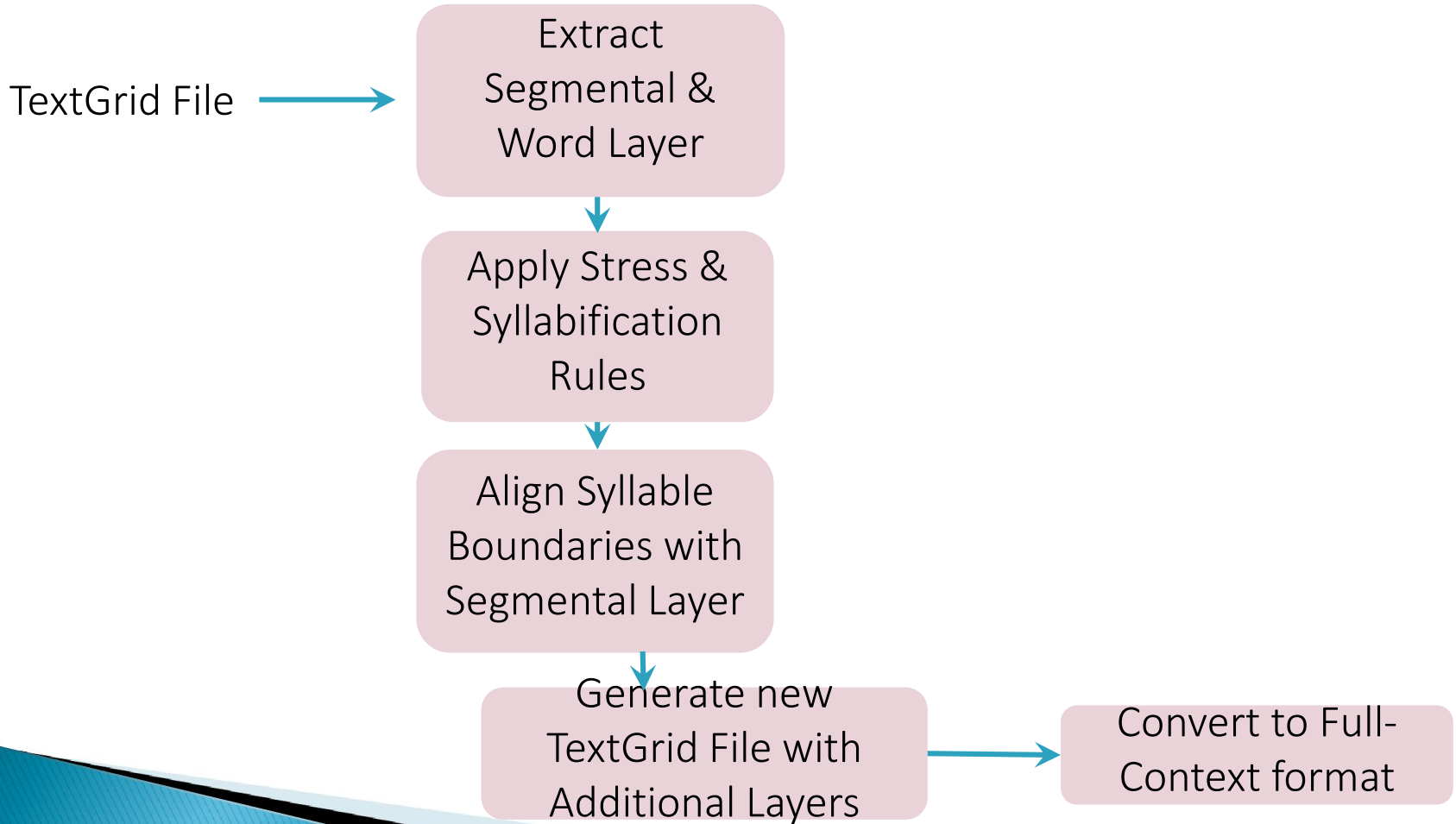
Steps to Generate Full-Context Labels:



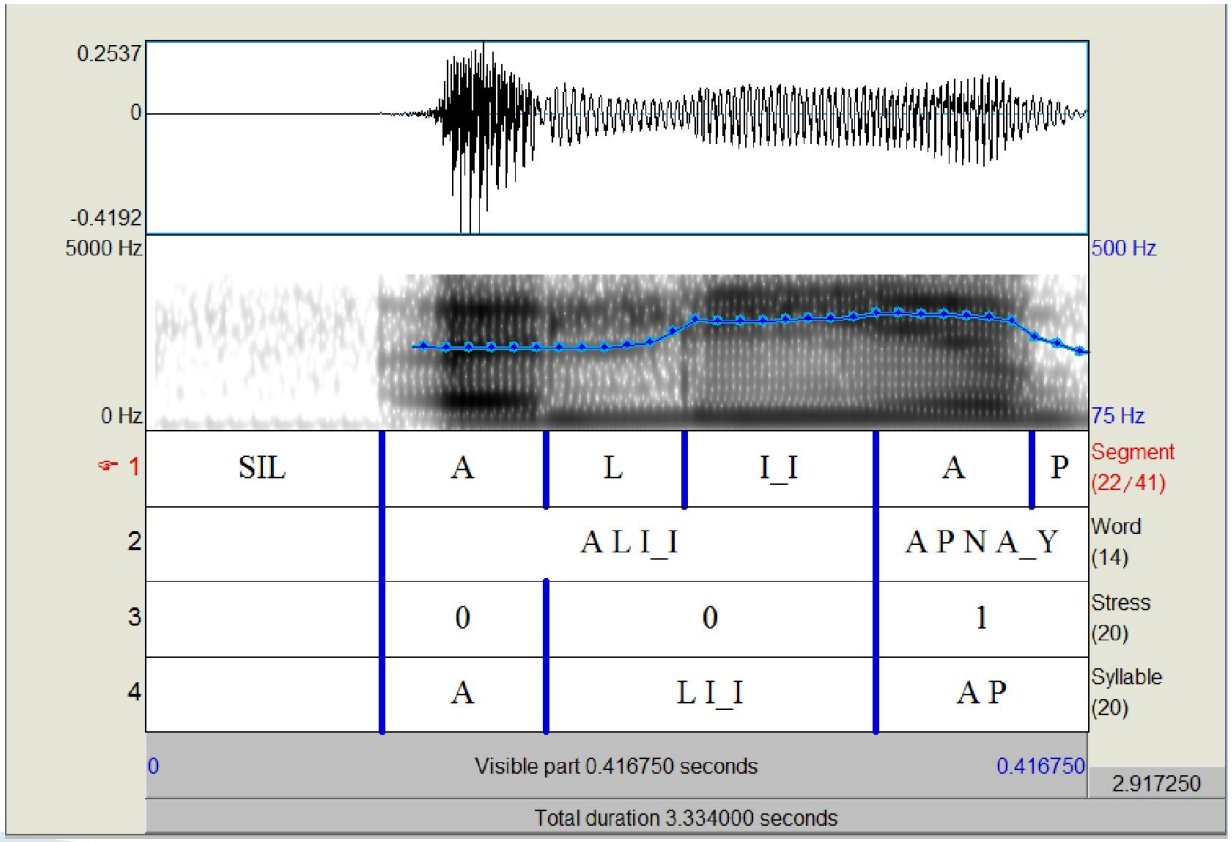
TextGrid Format:



Steps to Generate Full-Context Labels:



TextGrid Format with Additional Layers:



Context Clustering (Question Set) 1/2:

- Number of possible combinations are quite enormous with these 53 different contexts.
- With only Segmental Context 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

Context Clustering (Question Set) 2/2:

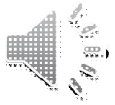
Phoneme

- {preceding, current, succeeding} phonemes

Stress/Syllable/Word/

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

Some Synthesized Examples:

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

Questions ?