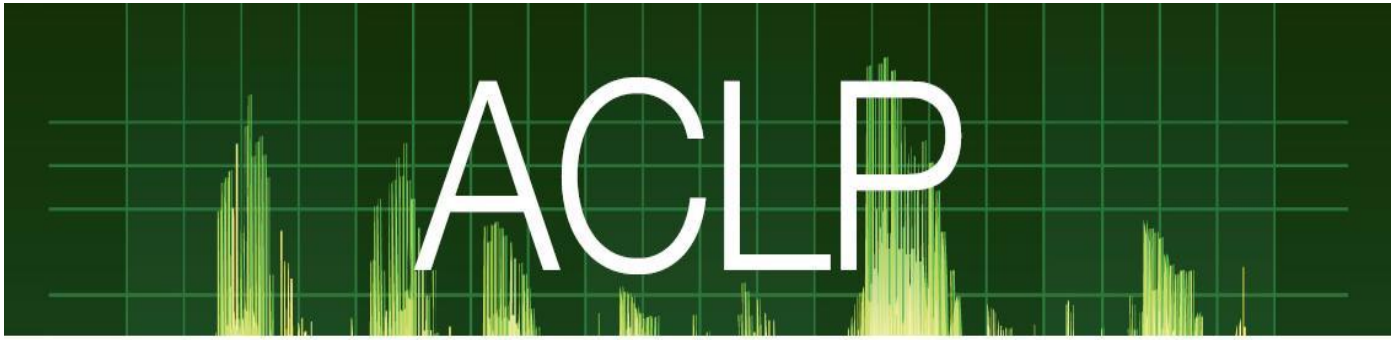
Approaches in Very Large Vocabulary Continues Speech Recognition



Afeka Center for Language Processing (ACLP)

Dikla Shchnike

Speech Recognition Day – 08.06.10

AFEKA TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING

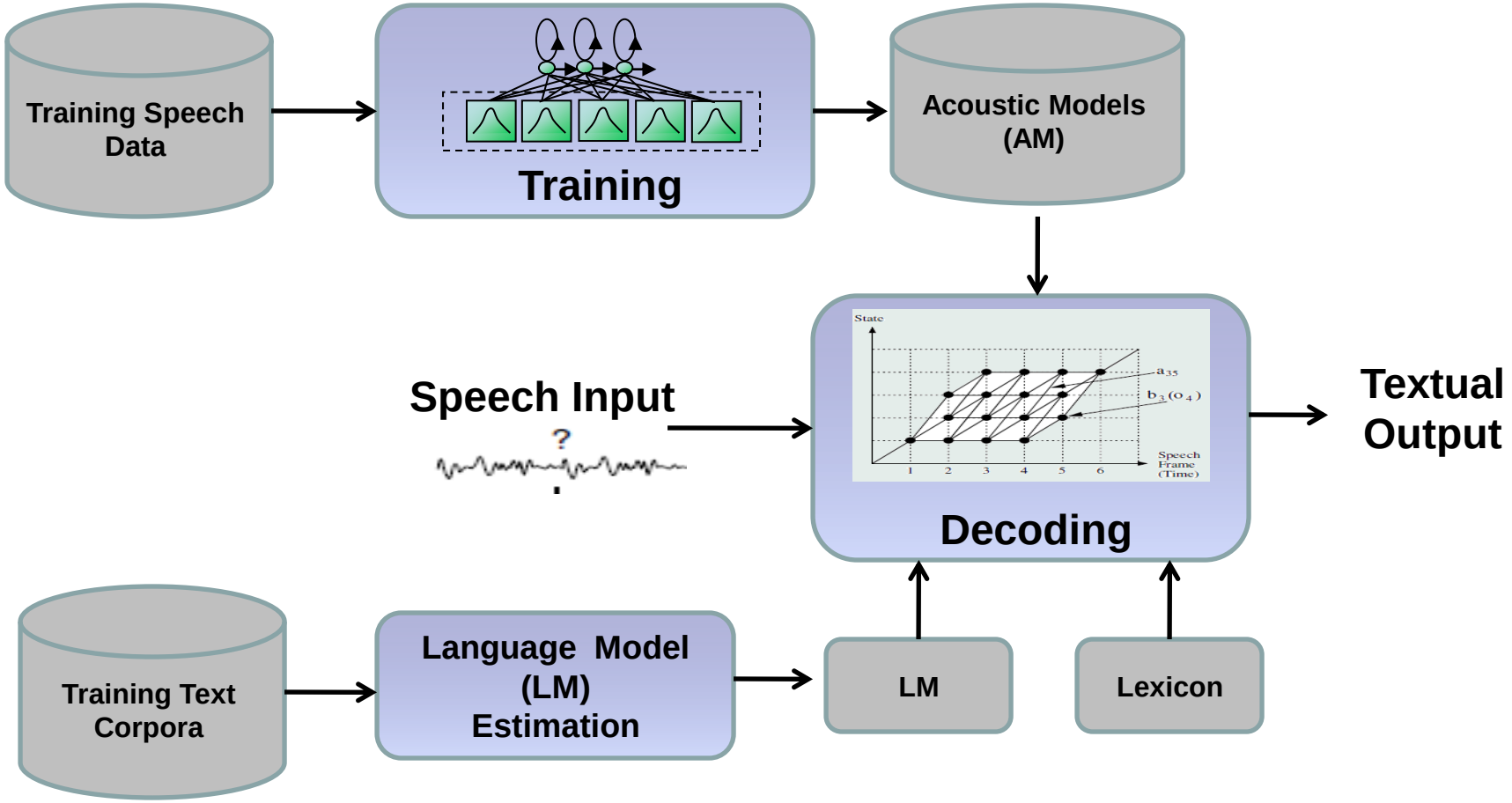
Motivation

- The Need:
  - LVCSR engine target market applications:
    - ✓ Dictation
    - ✓ VM Transcription
    - ✓ Navigation
    - ✓ Call Centers…
- Main Challenges:
  - Performance
  - Computational complexity
  - RT in large scale deployment

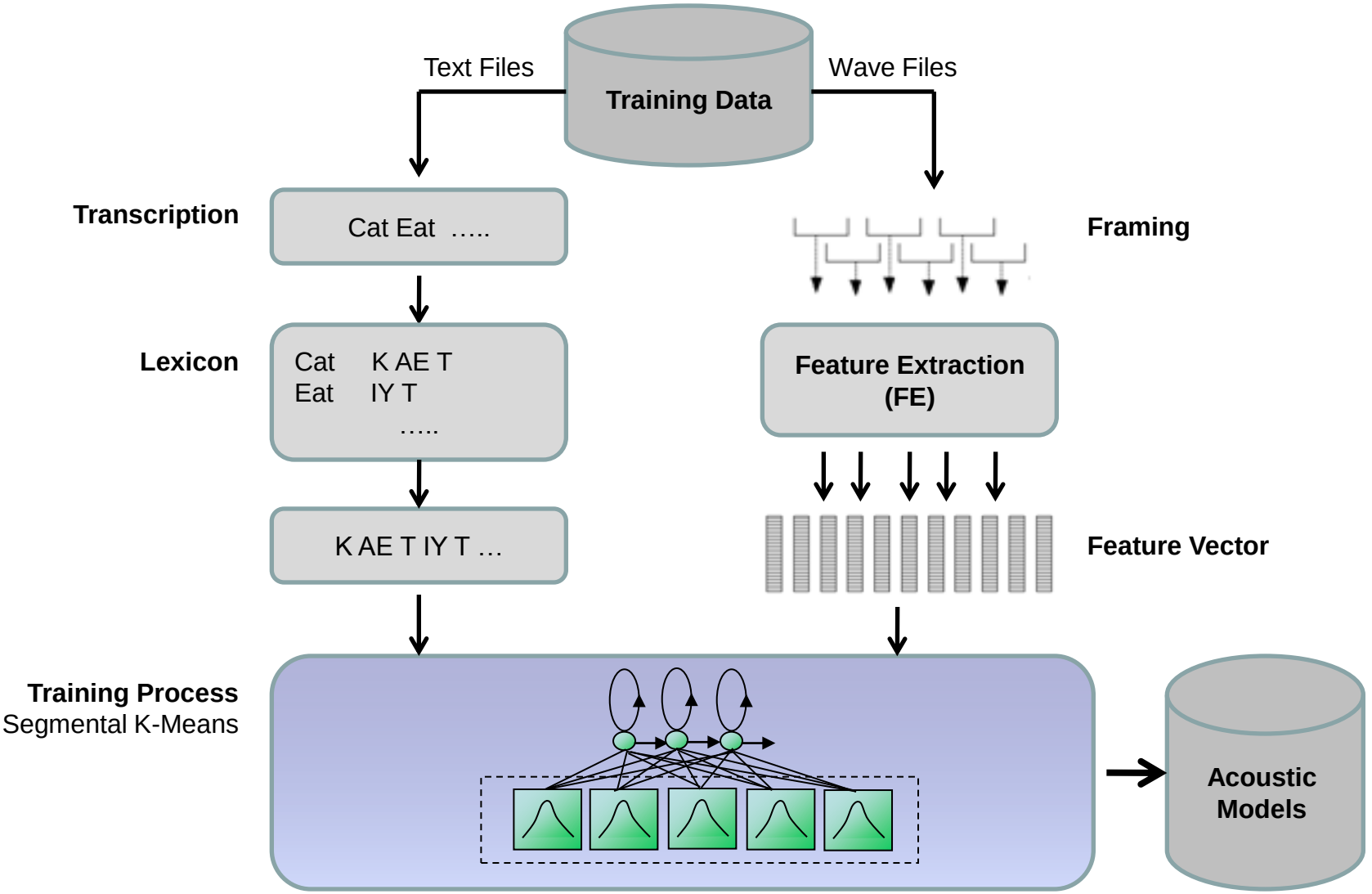
Research Focus

Performance vs. Computational Complexity
using various approaches

Classic LVCSR System Overview

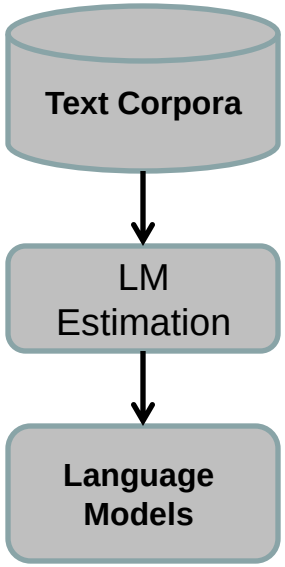


The Training Phase: *AM*

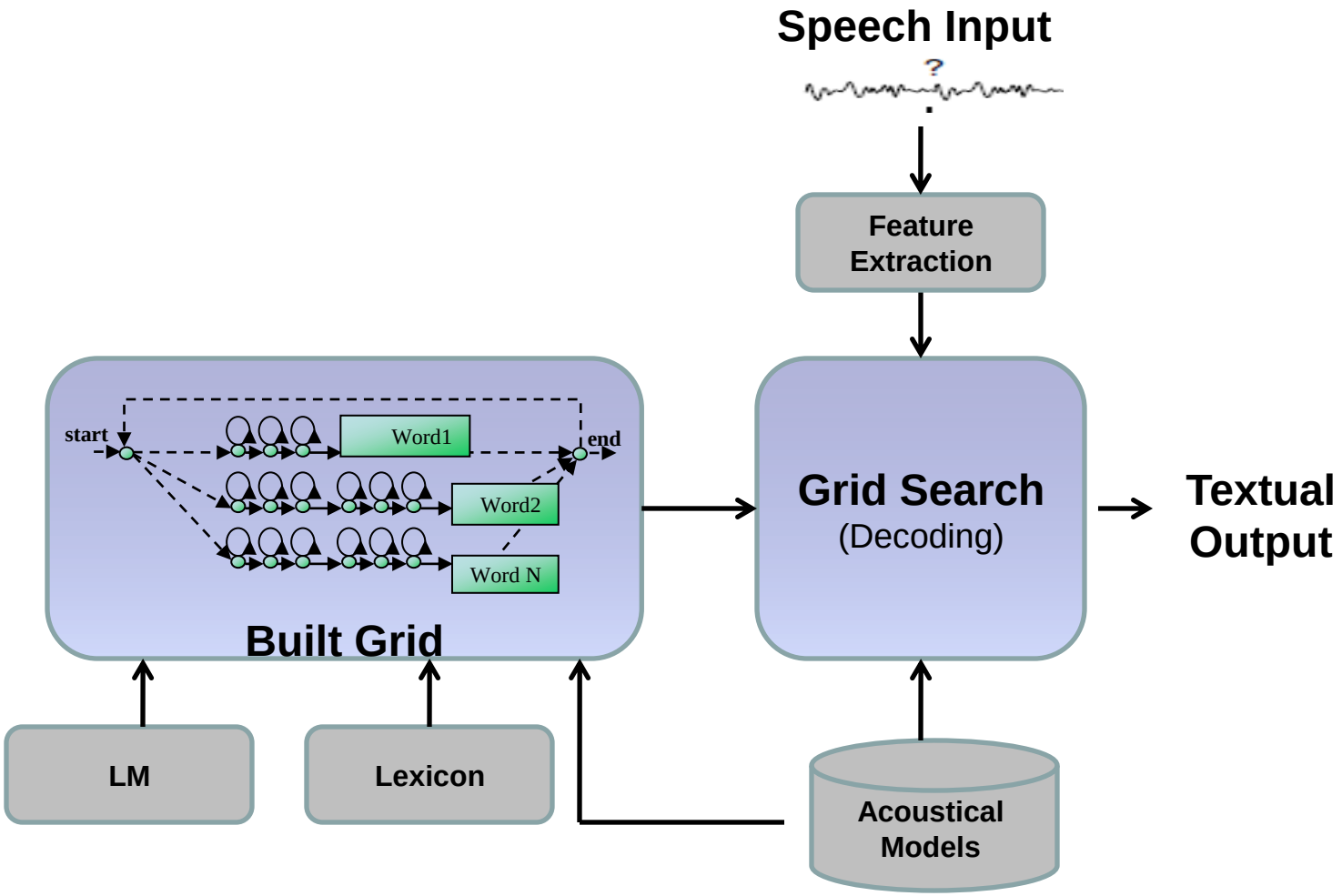


The Training Phase: *LM*

- Large textual corpora
- Variety of domains
- N-Gram:
  - Uni-gram
  - Bi-gram
  - Tri-gram
  - ...

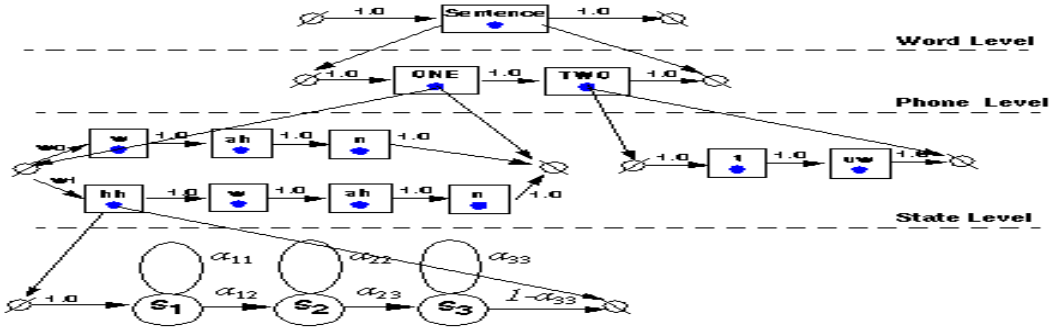


The Decoding Phase



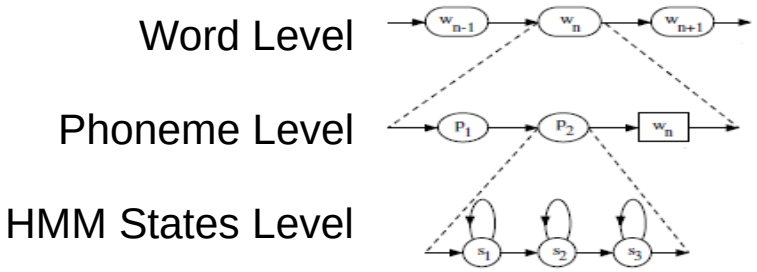
Classic LVCSR: *Main Advantages*

- One-time grid construction
- One stage of decoding
- Focused search



Classic LVCSR: *Disadvantages*

- High computational complexity



- Huge grid size
- Response time
- Mediocre recognition performance
- Every word must be included in the LM

Experimental Environment

- Language: American English
- Database: Macrophone
  - Channel: Telephone
  - No. of speakers: 4505
  - No. of recordings per speaker: 44
  - 1,037,924 total words (tokens)
  - 12,092 unique words (types)
- Read Speech

Experimental Environment : *Train*

- Database: Macrophone Training (4005 speakers)
- Features: MFCC 13 + Δ + $\Delta\Delta$
- Acoustic Model Topology:
  - HMM 3 state left to right
  - Tied state triphones
  - 16 Mix
- LM Topology:
  - Bi-Gram

Experimental Environment: *Test*

- DB: Macrophone DEVTEST (500 Speakers)
- Sentences only
- 3139 Utterances
- Lexicon size: 8.7K

Performance Measures

For phoneme, word and character sequences

- Correct Rate

$$CR = \frac{Hit}{TotalWords} \cdot 100$$

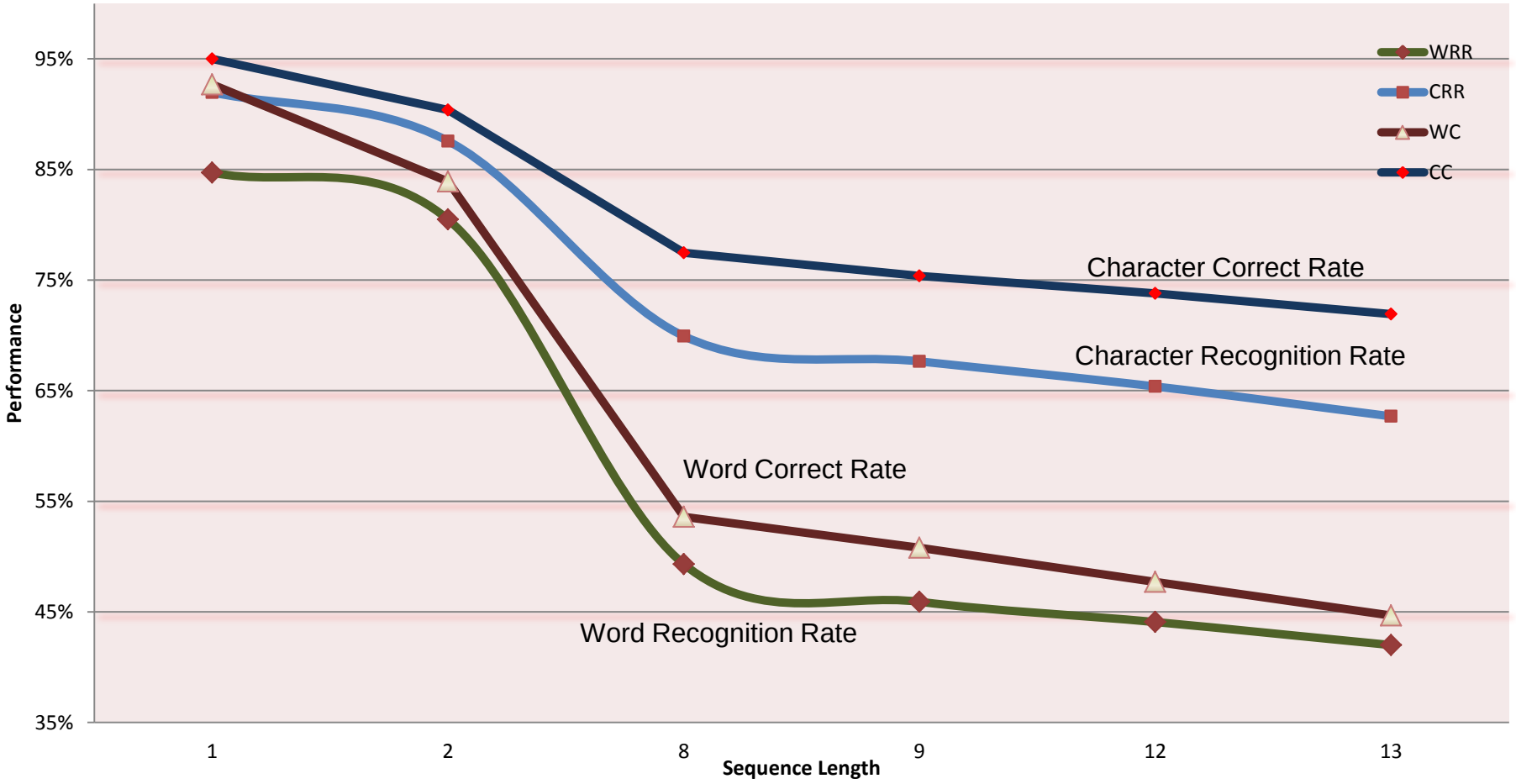
- Error Rate

$$ER = \frac{Deletion + Insertion + Subsition}{TotalWords} \cdot 100$$

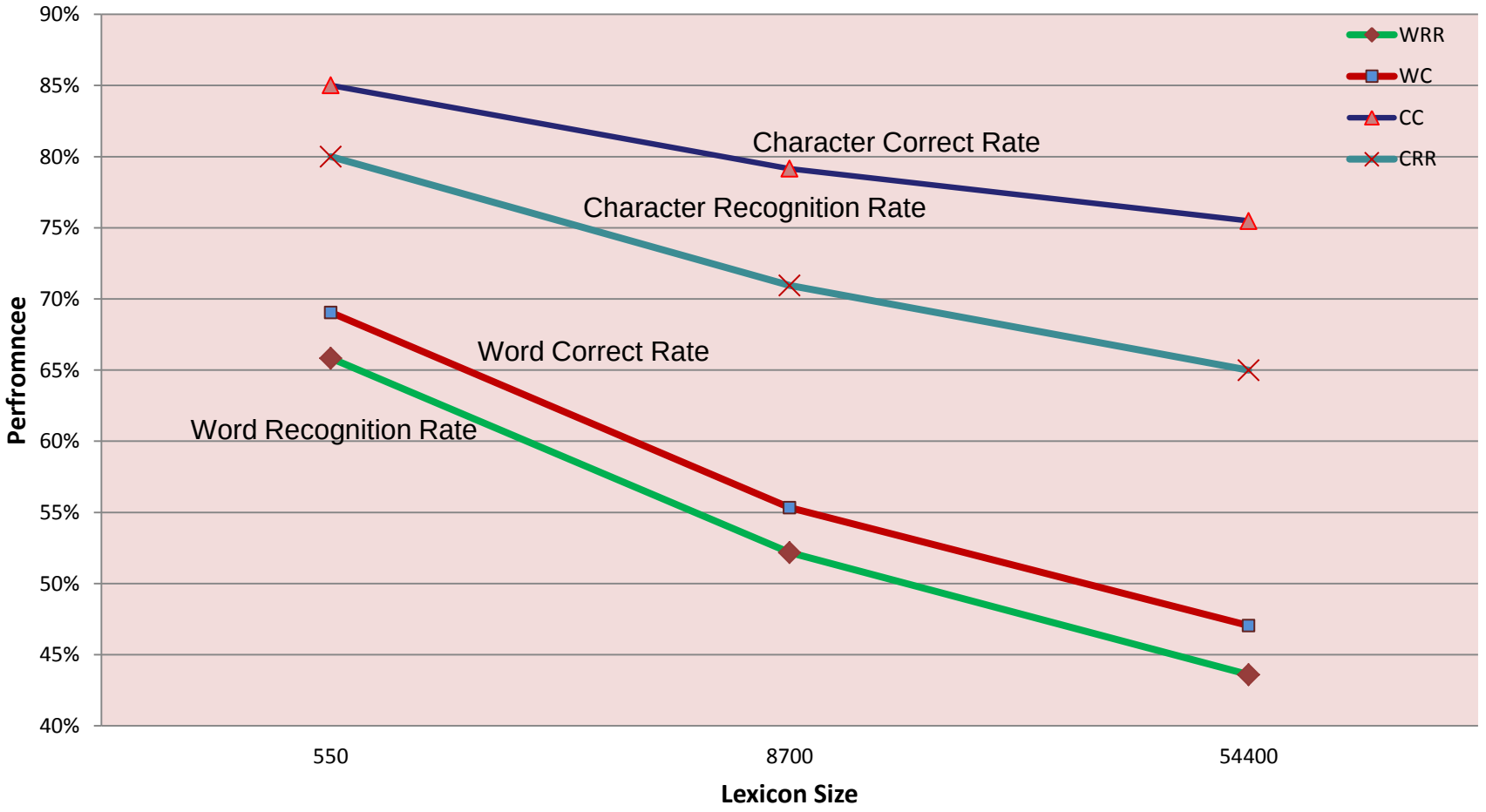
- Recognition Rate

$$RR = 100 - ER$$

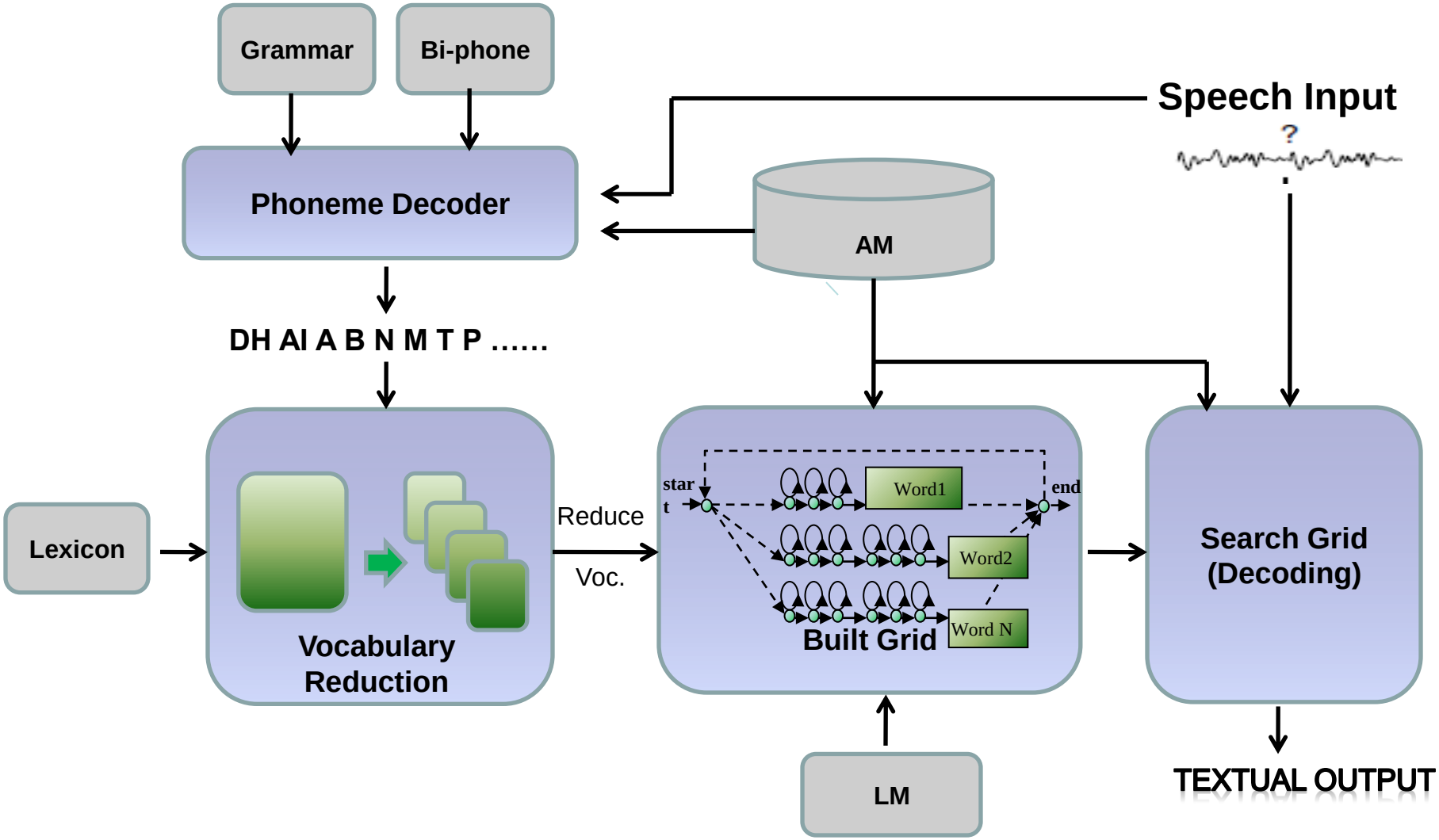
Performance vs. Sequence Length



Performance vs. Vocabulary Size – *10 word Utterance*



Multi-Stage LVCSR



Multi-Stage LVCSR: *Potential Advantages*

- Increased performance
  - The LVCSR works with a reduced vocabulary
- Reduced computational complexity
  - Additional stage/s vs. smaller LVCSR search space

Multi Stage LVCSR: *Disadvantages*

- Phoneme decoder recognition rate.
- Highly dependent on second stage coverage rate
- Sequential – potential increase in overall error rate

Classic vs. Multi stage

First Stage - Phoneme Decoder			
PCR [%]	PER [%]	PRR [%]	Processing Time [sec]
59.97	56.87	43.13	3.3

Second Stage – Vocabulary Reduction			
Lexicon Size	Reduced Voc.	Coverage[%]	Processing Time [sec]
8.7K	4.4K	93.38%	0.32
54.4K	27K	96.08%	**2.25**

Lexicon Size	Classic LVCSR				Third Stage - Multi stage LVCSR			
	CCR [%]	CER [%]	CRR [%]	Processing Time [sec]	CCR [%]	CER [%]	CRR [%]	Processing Time [sec]
8.7K	75.29	33.35	66.65	13.0	73.64	35.44	64.56	7.0
54.4K	75.48	35.00	65.00	**65**	74.58	36.01	63.99	**31**

Next Steps

- A more extensive set of experiments should be conducted
  - Test on spontaneous speech
  - Increased lexicon size – 100k, 250k
- Test LVCSR with Pruning
- Lattice – in phoneme decoder stage
- Replace grid search with textual algorithm

The End

Acoustic Model Topology

- 39 monophones
- 6 noises
- #Triphones before DCT: 7865
- #Triphones after DCT: 3377
- #States After DCT: 1476
- Global covariance
- 16 mixtures for triphones
- 32 mixtures for noise.

Test Set Information

- Sentences from : Timit, ATIS, WSJ

Item	#Utt.	Average Phoneme in Sentence	Average Word in Sentence	Sentence Duration [sec]
TIMIT	1334	36.19	7.9	3.31
ATIS	925	37.49	9.38	3.48
WSJ	880	39.06	9.08	3.61

Phoneme Decoder Performance

Item	#Utt.	Correct [%]	Del [%]	Sub [%]	Ins [%]	PER [%]	PRR [%]
Timit	1334	58.19	7.64	34.16	17.11	58.92	41.08
WSJ	880	58.31	7.78	33.92	16.37	58.06	41.94
ATIS	925	64.12	6.88	29.00	16.93	52.81	47.19
VM1	1341	46.29	14.40	39.31	23.69	77.39	22.61
VM2	1144	45.56	14.39	40.04	26.32	80.76	19.24

Processing Time Information
Full Analysis of 3 Utt.

- Feature Extraction – 0 sec
- Built Grid

Phoneme			Bigram		
47 nodes 1669 arcs	4403 nodes 29803 arcs	8759 nodes 60537 arcs	27228 nodes 78027 arcs	54451 nodes 151922 arcs	8756 nodes 6426195 arcs
0 [sec]	0 [sec]	1 [sec]	1 [sec]	2 [sec]	37 [sec]

- Decoding

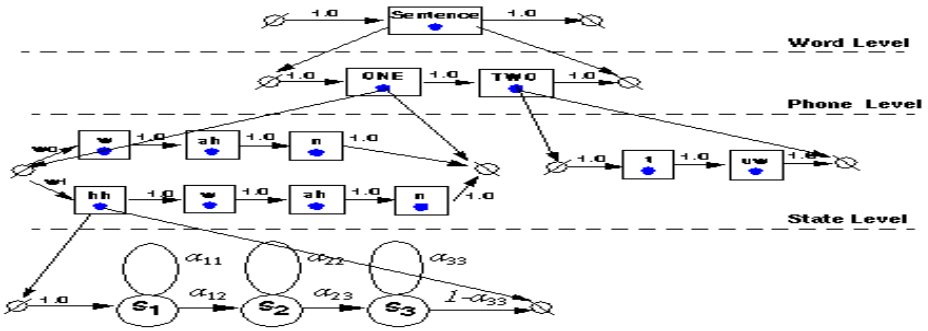
Phoneme			Bigram		
47 nodes 1669 arcs	4403 nodes 29803 arcs	8759 nodes 60537 arcs	27228 nodes 78027 arcs	54451 nodes 151922 arcs	8756 nodes 6426195 arcs
3.3 [sec]	7 [sec]	13 [sec]	31 [sec]	65 [sec]	83 [sec]

Computer Parameters

- Intel Dual Core CPU
- Processor Clock 2.5GHz
- 3GB of RAM


boy
house
pizza
tree
milk
zoo
ring
baby
fish
car
apple
ear


three
red
five
green
ten
white
right
orange
purple
nice


Sentence
1.0
Word Level
ONE
TWO
Phone Level
w
ah
n
hh
uw
State Level
a_{11}
a_{22}
a_{33}
a_{12}
a_{23}
$1-a_{33}$
S_1
S_2
S_3