An Efficient Algorithm for the Transcription of Spontaneous Speech

Ami Moyal
ACLP – Afeka Center for Language Processing
Afeka Academic College of Engineering

SpeechTEK Europe
May 27, 2010

ACLP
Afeka Center for Language Processing

- A research and instruction laboratory for spoken and written language processing
- Research activities in speech and text processing
- Joint R&D projects with industry
- Consulting services to the industry
- Specialized courses for industry professionals
- Project opportunities for final year students

Goal: To become a source of knowledge in the fields of spoken and written language processing and related applications.

Agenda

- Project Description
- LVCSR Systems
- Project Main Activities
- Voicemail Content
- Lexicon Construction
- Efficient Implementation
- Preliminary Results
- Summary and Next Steps

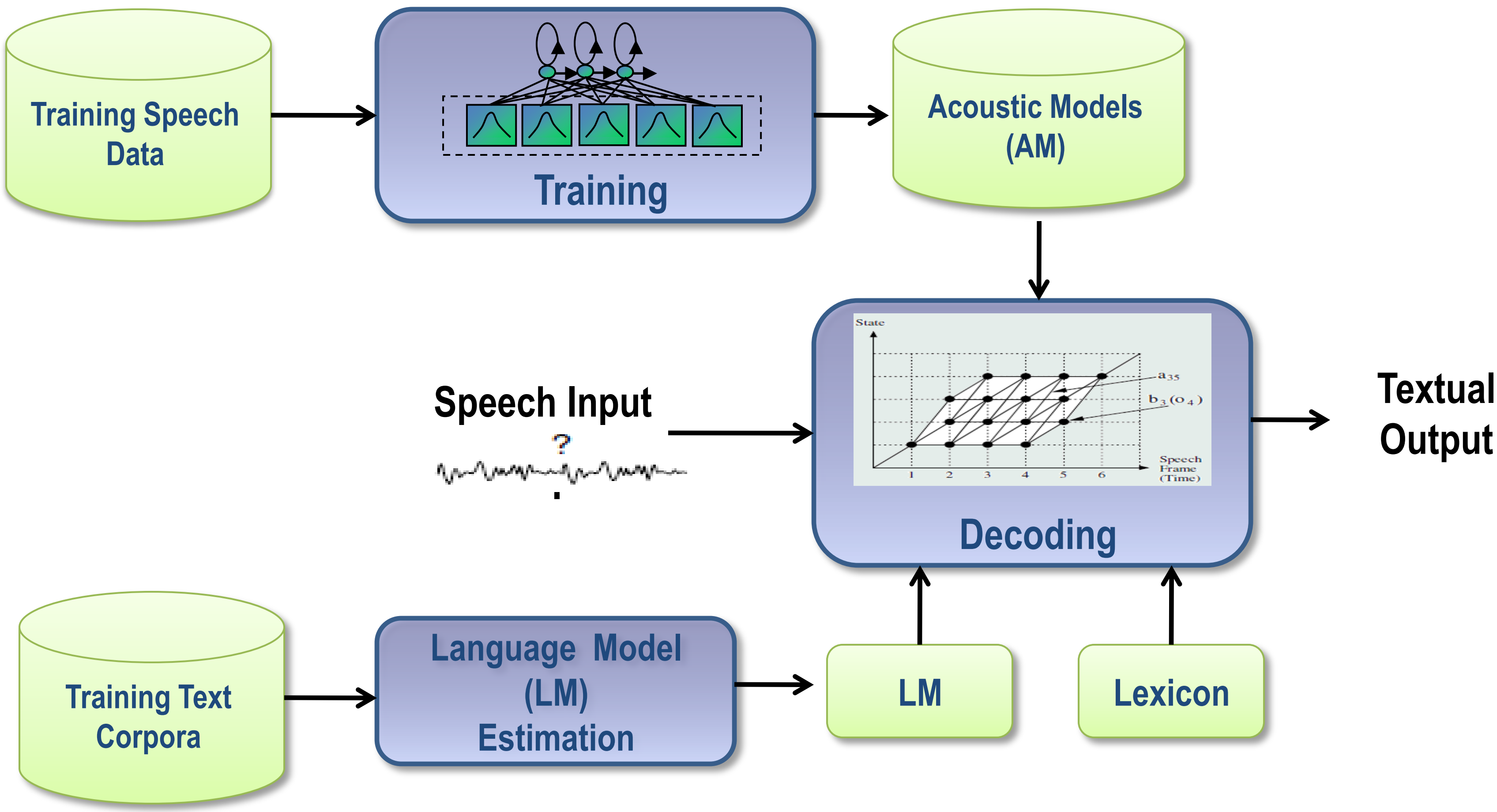
Project Description

- Research project funded by the Chief Scientist of the Israeli Ministry of Industry and Commerce
- Encourages the delivery of technology from academic institutes to the industry
- Joint research between Afeka (academic institute) and SpeechModules LTD (industry company)
- 11 researchers
- Two Years

Project Focus

- Spontaneous speech transcription
- Focus on messaging transcription
- Requirement for various applications:
  - VM to SMS
  - SMS speech Input
  - IM speech input
- Testing on real VM database
- Very Large Vocabulary – 250K words
- Computational complexity vs. recognition performance in various LVCSR structures
- Use results in mass market messaging transcription application

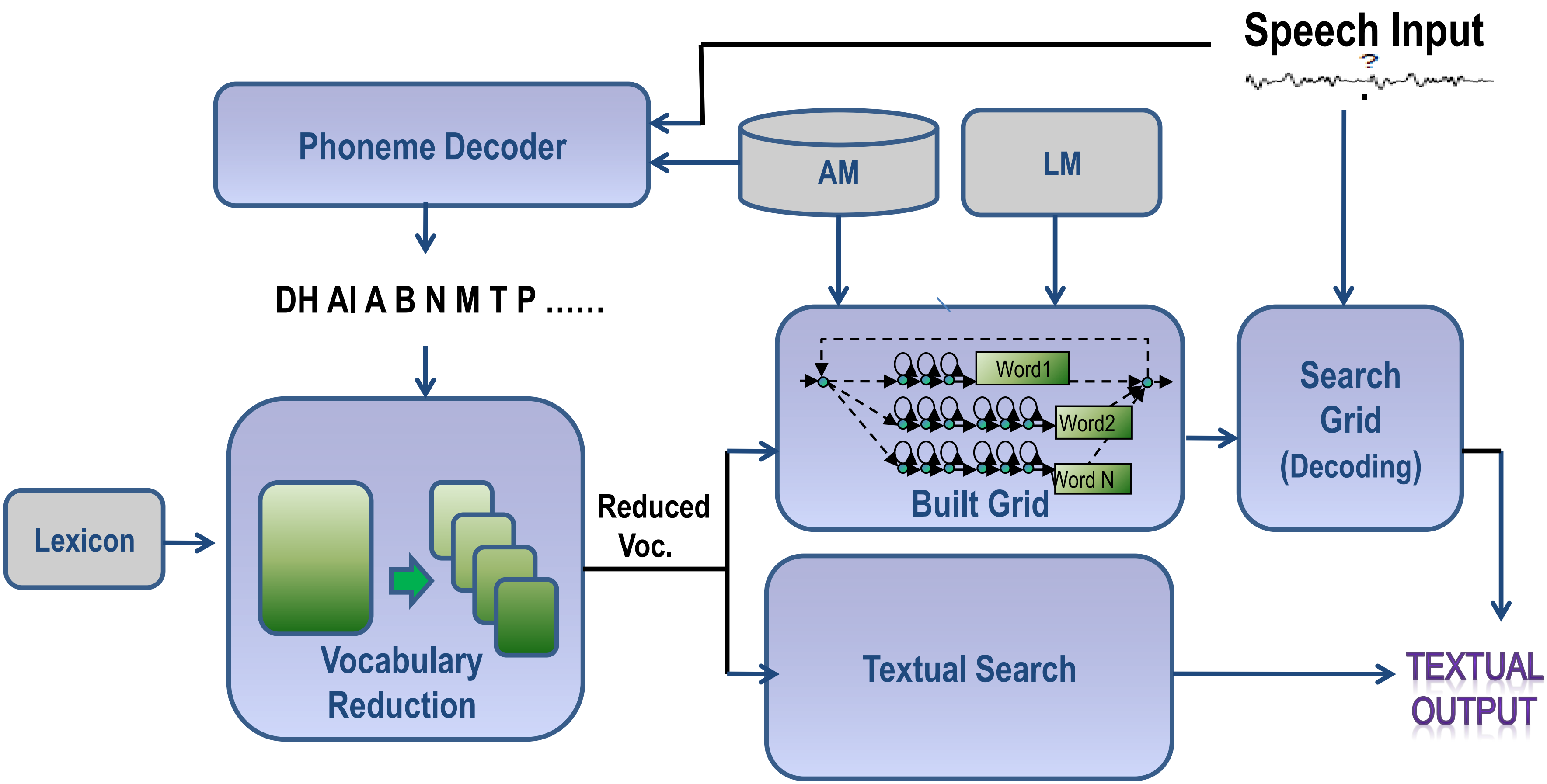
LVCSR System Overview



Main Challenges

- Performance with very large vocabularies
- Spontaneous speech
- Computational complexity
- RT operation in large scale deployments

Multi-Stage LVCSR



Multi-Stage LVCSR
Potential Advantages

- Better performance – reduced vocabulary at final recognition stage
- Less sensitivity to vocabulary size
- Reduced computational complexity
  - Efficient Additional stages vs. smaller search space
- One engine using a huge lexicon for various domains

Project Main Activities

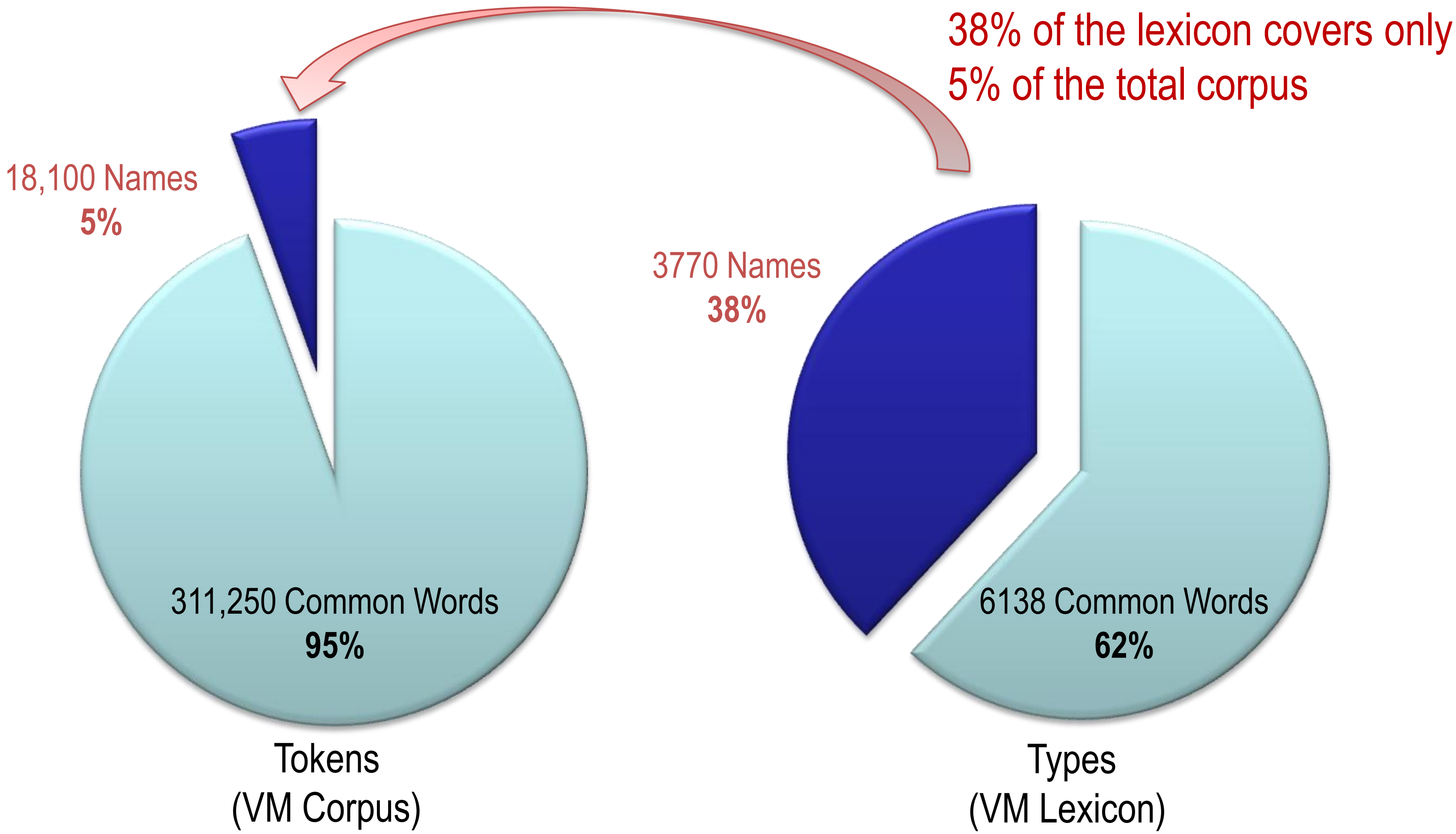
- Lexicon design and construction
- Word hypothesis creation
- Efficient phoneme sequence distance measure
- Testing on real VM DB

Voicemail Content

- Spontaneous speech
- Very large vocabulary
- Hesitations and other disfluencies
- Unstructured speech
- Large number of names

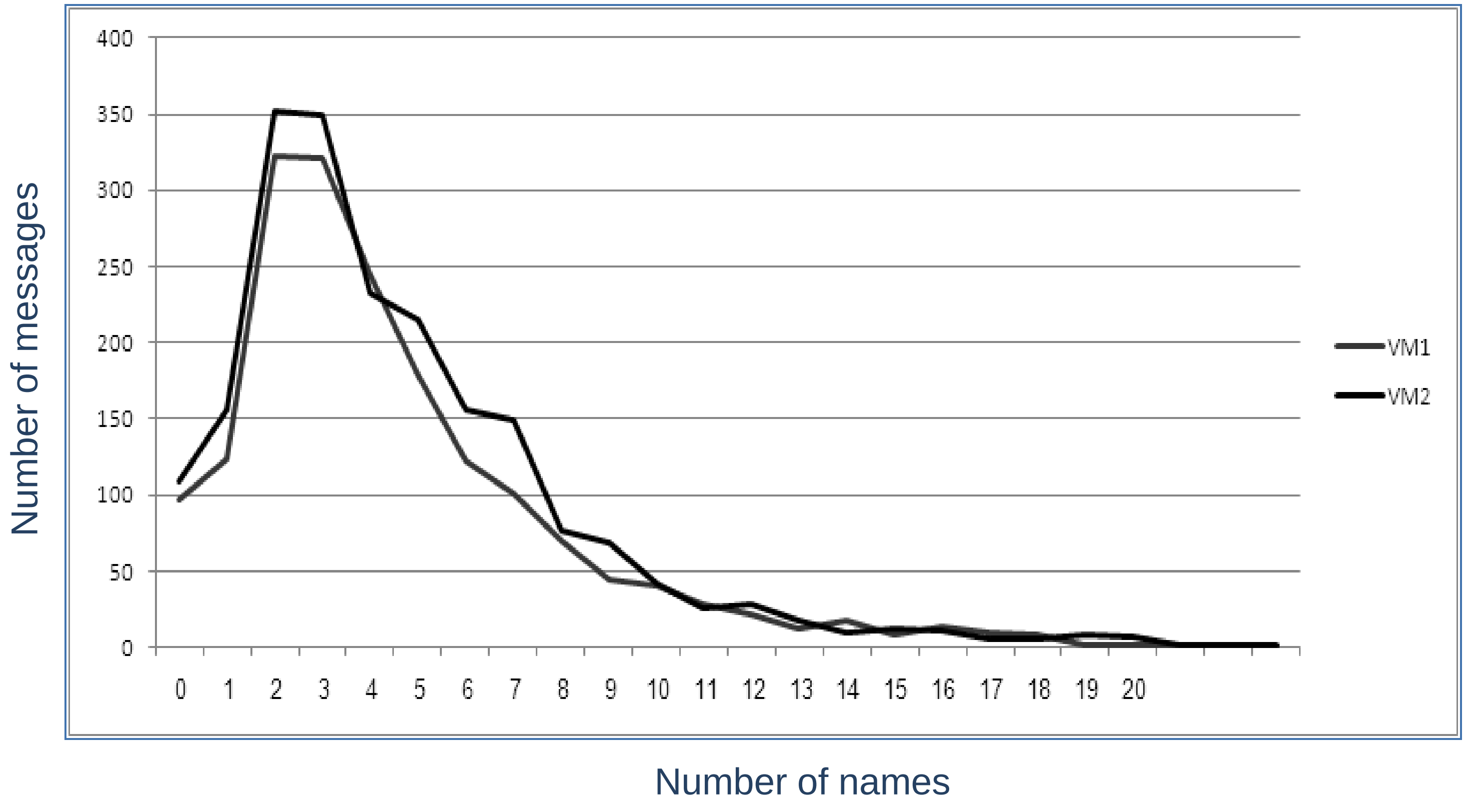
- text

Names in VM DB



Names in Voice Messages

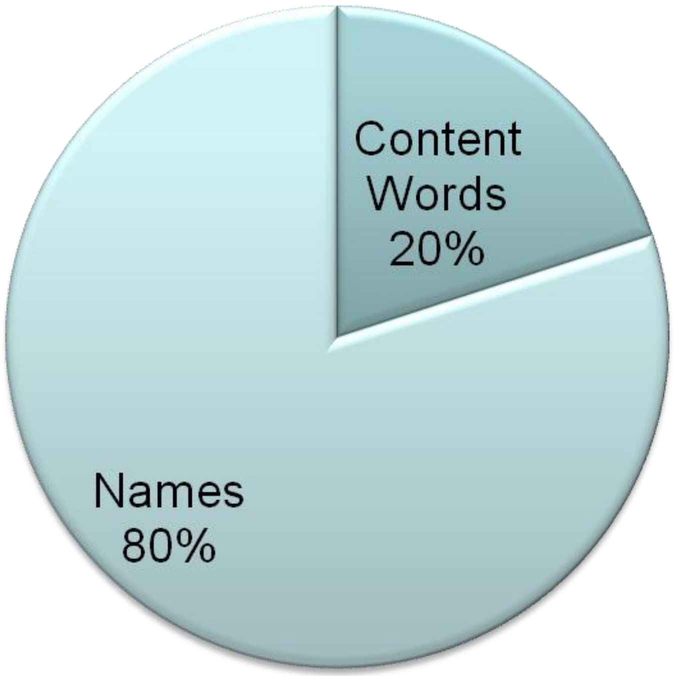
Peak at 2-3 names per voice message



Voicemail Lexicon Design

For 90% per domain coverage – 250K

- Content words (50K)
- Names (200K)
  - 150K people
  - 50K places, organizations



Lexicon Status

- A word list for a lexicon of 250K was compiled
- An initial version of a 100K Lexicon was completed
- The 100K-word lexicon is currently being used for testing

Word Hypothesis Creation

Only anchor-based hypotheses will be evaluated

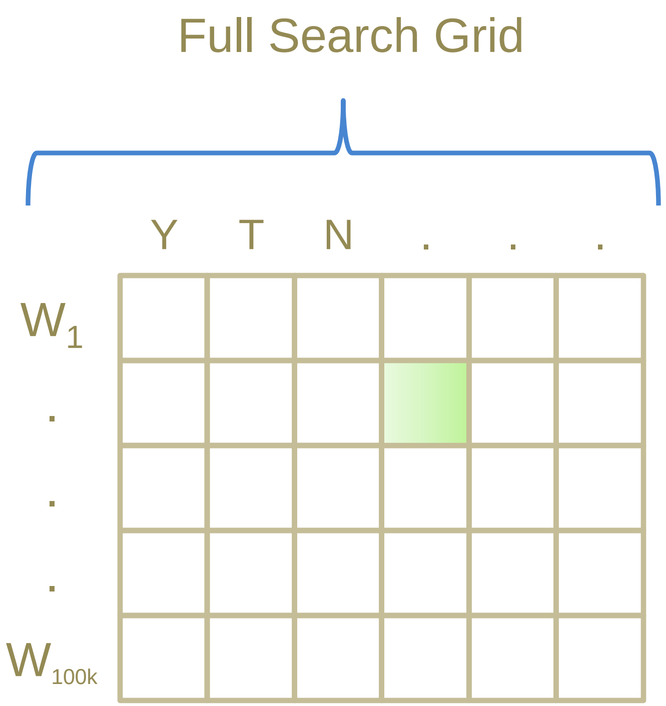
Recognized Phoneme Sequence

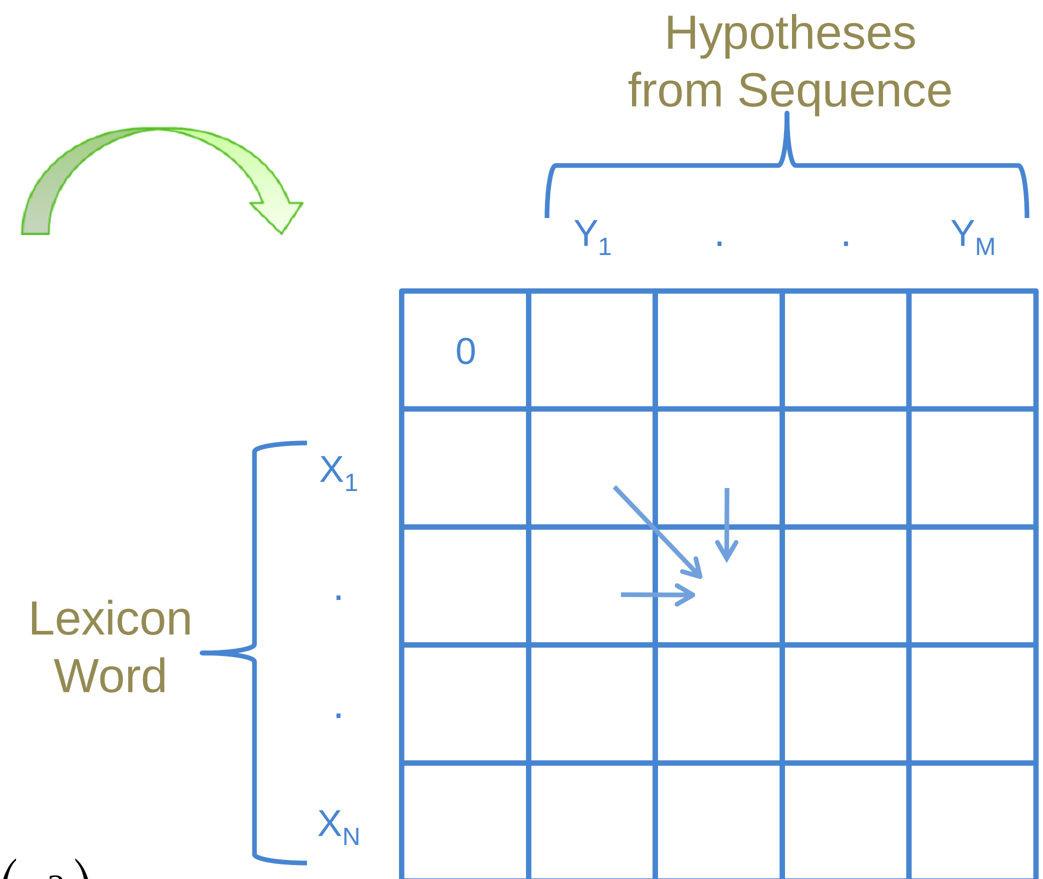
Lexicon	$\overbrace{\text{SH} \quad \text{OW}}^{\text{Hyp 3}}$		$\overbrace{\text{M} \quad \text{IY} \quad \text{DH}}^{\text{Hyp 2}}$			$\overbrace{\text{AH} \quad \text{F} \quad \text{E} \quad \text{R}}^{\text{Hyp 1}}$				
	SH	***OW***	***M***	***IY***	***DH***	***AH***	***F***	***E***	***R***	...
about AH B AW T	0	0	0	0	0	D_1	0	0	0	
keep K IY P	0	0	D_2	0	0	0	0	0	0	
show SH OW	D_3	0	0	0	0	0	0	0	0	
⋮										
Word 100,000										

Word Hypothesis Creation Current Status

- Various anchor types and methods have been tested
- Currently per word hypothesis the best anchor is chosen
- 50% decrease in time, loss of 0.5% in word coverage

Phoneme Distance Measure





$$Number\ of\ \ D\left(W_i, Hyp_j\right):$$

$$Grid\ Size \times O\left(n^2\right)$$

Phoneme Distance Measure Current Status

- Various Methods have been tested
- Diagonal distance measure - 90% decrease in time, loss of 2% in coverage
- Adding weighted measure based on phoneme recognition performance – increase in original coverage by 5%

Experimental Environment
Macrophone Database

- Language - American English
- Channel - Telephone
- Number of Speakers – 4505
  - ~4000 – training, ~500 - testing
- Above 1 million words
- ~12K unique words
- Read speech

Experimental Environment
Voicemail Database

- Language - American English
- Channel – Telephone
- Typical message:
  - 21 seconds
  - 75 words
  - 250 phonemes
- Above 300K words
- ~10K unique words
- Spontaneous speech – Authentic messages

Acoustic Model Training

- Train DB - Macrophone Training (4005 speakers)
- Phoneme set - 39
- Features - MFCC 13 + Δ + $\Delta\Delta$
- Acoustic Model Topology:
- HMM 3 state left to right
- Tied state triphones
- 16 Mix

Preliminary Results

- Telephony
- Read Speech
- Correct Character Recognition – 85%
- CER – Character Error Rate – 20%
  - Substitution - 8.1%
  - Deletion - 6.88%
  - Insertion - 5.02%

Summary and Next Steps

- Completed in the initial 8 months of the 24 month project :
  - Infrastructure established
  - Initial versions for all three activity tracks
  - Preliminary results – reduction in search space with increase/no loss in word coverage
- Performance of the overall system compared to LVCSR engine?