DEEP LEARNING FOR SPEECH RECOGNITION

Anantharaman Palacode Narayana Iyer

JNResearch

ananth@jnresearch.com

15 April 2016

REFERENCES

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Deep Neural Networks for Acoustic Modeling in Speech Recognition

The shared views of four research groups

CS 224S / LINGUIST 285
Spoken Language Processing

Dan Jurafsky
Stanford University

Lecture 1: Introduction, ARPAbet, Articulatory Phonetics

Chapter 7

Connectionist Temporal Classification

Towards End-to-End Speech Recognition with Recurrent Neural Networks

Alex Graves GRAVES@CS.TORONTO.EDU
Google DeepMind, London, United Kingdom

Navdeep Jaitly NDJAITLY@CS.TORONTO.EDU
Department of Computer Science, University of Toronto, Canada

CS224D: Deep Learning for Natural Language Processing

Andrew Maas
Stanford University
Spring 2015

Neural Networks in Speech Recognition

AGENDA

- Types of Speech Recognition and applications
- Traditional implementation pipeline
- Deep Learning for Speech Recognition
- Future directions

SPEECH APPLICATIONS

- Speech recognition:
  - Hands-free in a car
  - Commands for Personal assistants – e.g Siri
  - Gaming
- Conversational agents
  - E.g. agent for flight schedule enquiry, bookings etc
- Speaker identification
  - E.g Forensics
- Extracting emotions and social meanings
- Text to speech

TYPES OF RECOGNITION TASKS

- Isolated word recognition
- Connected words recognition
- Continuous speech recognition (LVCSR)
- The above can be realized as:
  - Speaker independent implementation
  - Speaker dependent implementation

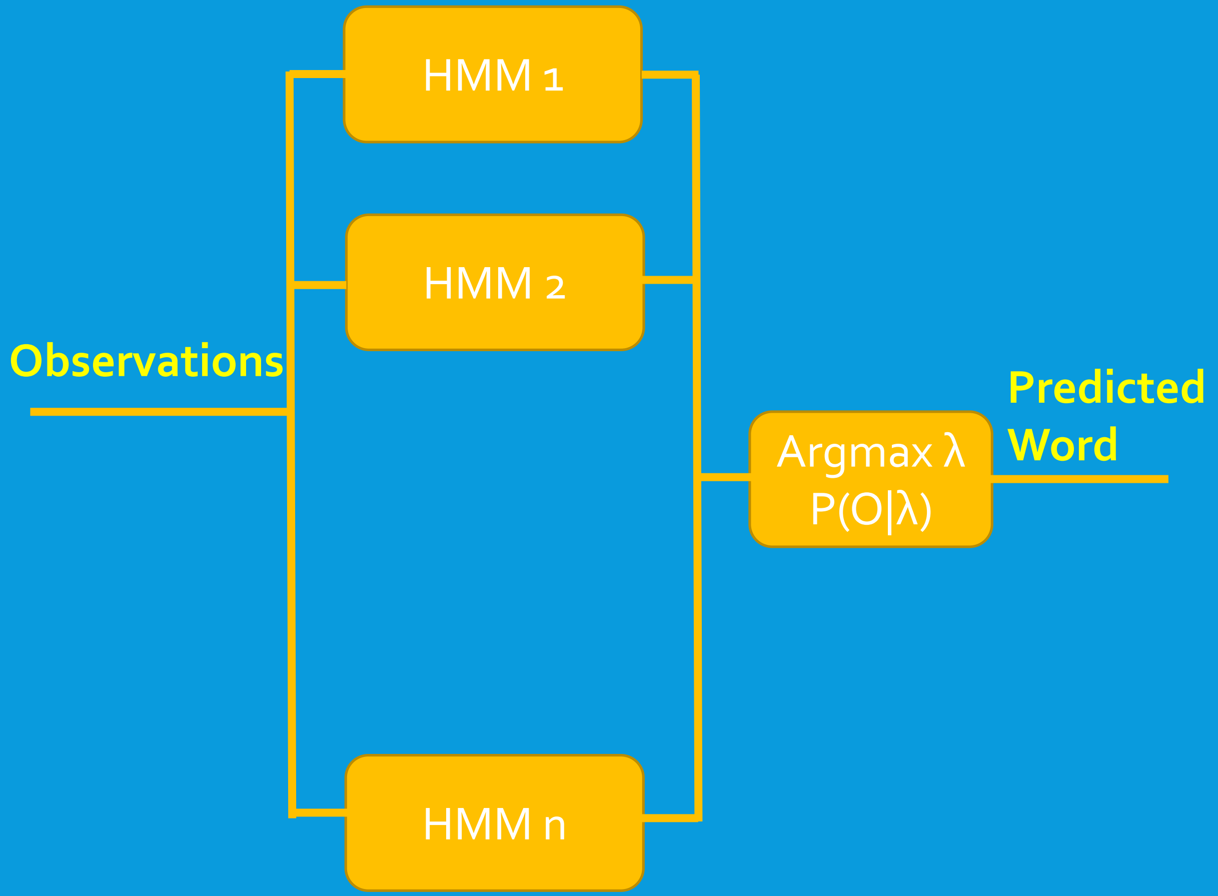
SPEECH RECOGNITION IS PROBABILISTIC

Steps:

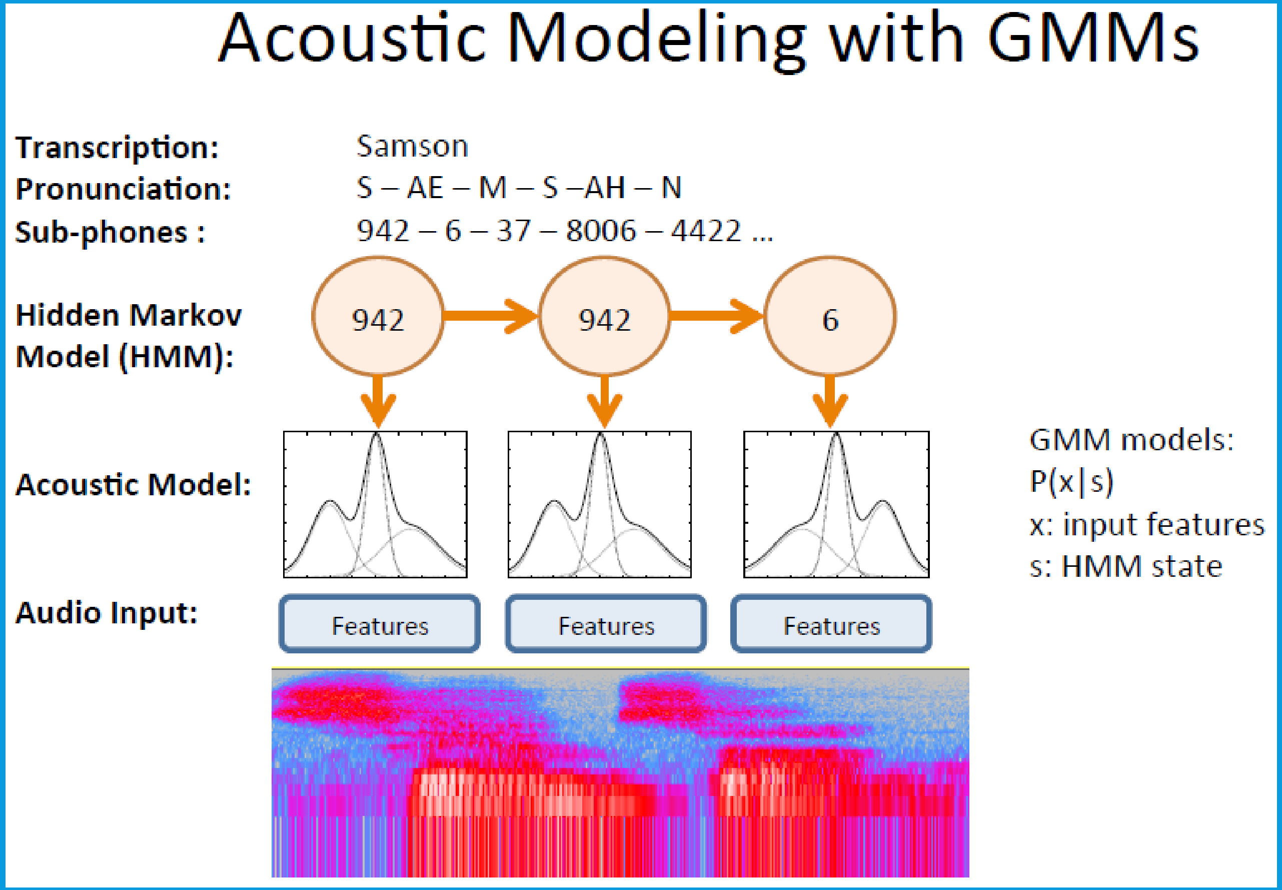
- Train the system
- Cross validate, finetune
- Test
- Deploy

ISOLATED WORD RECOGNITION

- From the audio signal generate features. MFCC or Filter banks are quite common
- Perform any additional pre-processing
- Using a code book of a given size, convert these features in to discrete symbols. This is the vector quantization procedure that can be implemented with k-means clustering
- Train HMM's using Baum Welch algorithm
  - For each word in the vocabulary, instantiate a HMM
  - Intuitively choose the number of states
  - The set of symbols are all valid values of the code book
- Use the HMM to predict unseen input

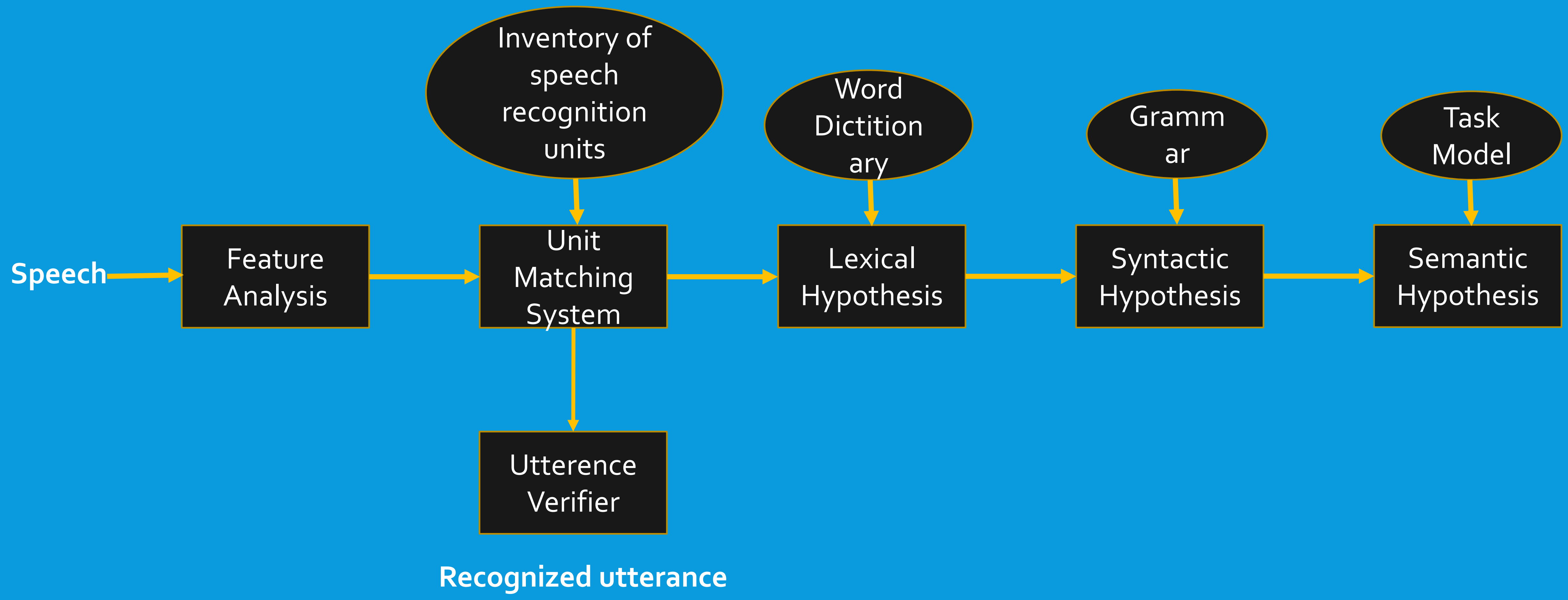


CONTINUOUS SPEECH RECOGNITION



- ASR for continuous speech is traditionally built using Gaussian Mixture Models (GMM)
- The emission probability table that we used for discrete symbols is now replaced by GMM
- The parameters of this model are learnt as a part of the training using Baum Welch procedure

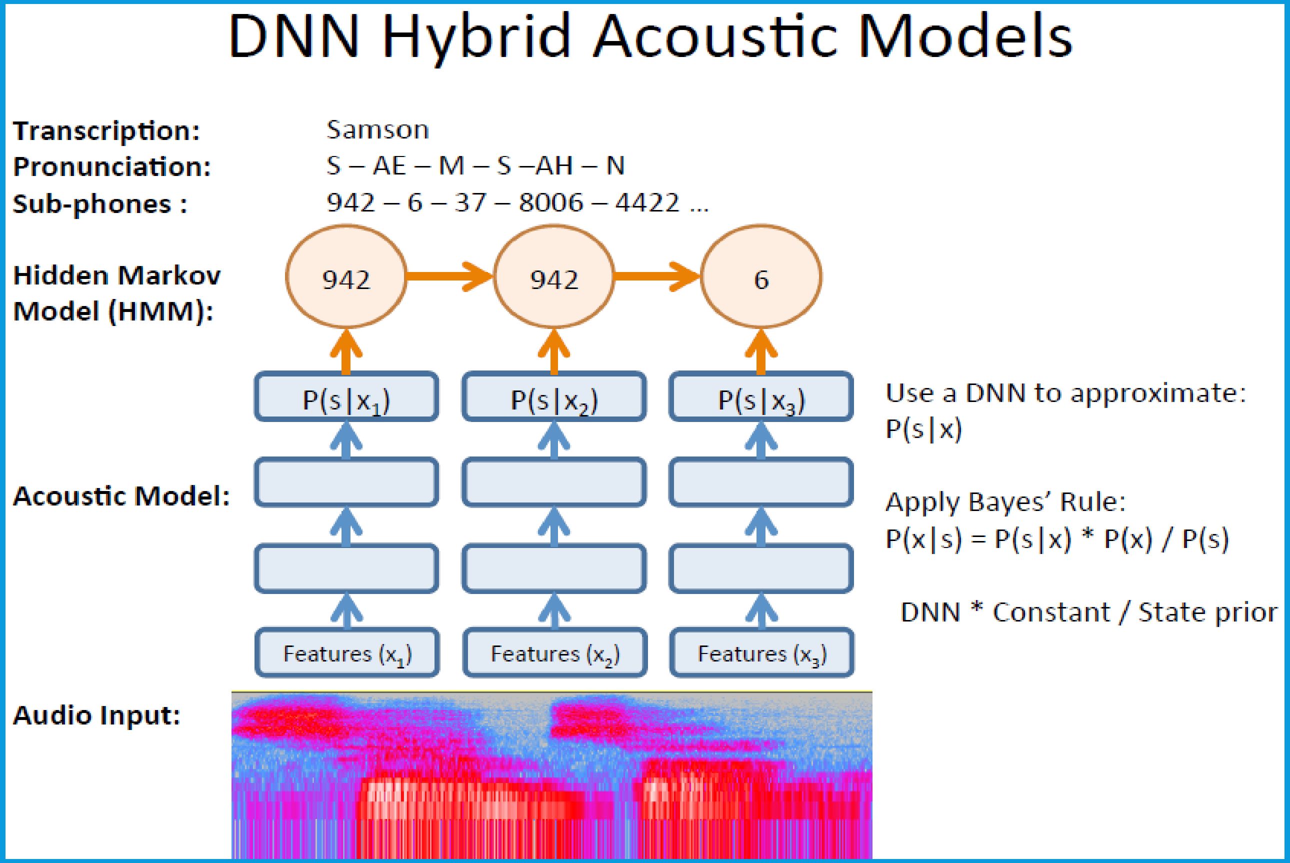
KNOWLEDGE INTEGRATION FOR SPEECH RECOGNITION

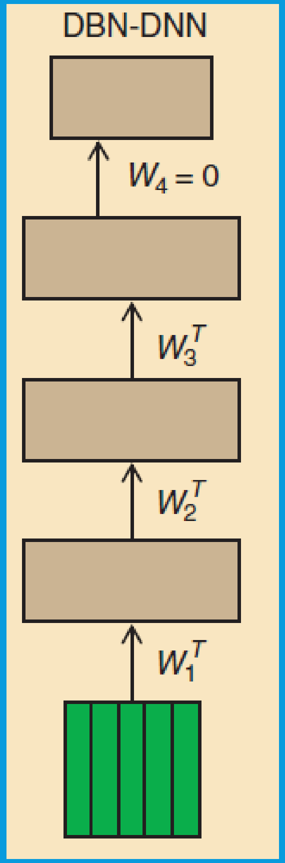


SOME CHALLENGES

- We don’t know the number of words
- We don’t know the boundaries
- They are fuzzy and non unique
- For V word reference patterns and L positions there are exponential combinatorial possibilities

USING DEEP NETWORKS FOR ASR





- Replace the GMM with a Deep Neural Networks that directly provides the likelihood estimates
- Interface the DNN with a HMM decoder
- Issues:
  - We still need the HMM with its underlying assumptions for tractable computation

EMERGING TRENDS

- HMM-free ASRs
  - Avoids phoneme prediction and hence the need to have a phoneme database
  - Active area of research
- Current state of the art adopted by the industry uses DNN-HMM
- Future ASRs are likely to be fully neural networks based