Discriminative Training

Lecture 22

CS 753

Instructor: Preethi Jyothi

Recall: MLE for HMMs

Maximum likelihood estimation (MLE) sets HMM parameters so as to maximise the objective function:

$$\mathcal{L} = \sum_{i=1}^{N} \log P_\lambda(X_i | W_i)$$

where
$X_1, \ldots, X_i, \ldots X_N$ are training utterances
(Assume M_i is the HMM corresponding to the word sequence W_i of X_i and λ corresponds to the HMM parameters)

What are some conceptual problems with this approach?

Discriminative Learning

- Discriminative models directly model the class posterior probability or learn the parameters of a joint probability model discriminatively so that classification errors are minimised
  - As opposed to generative models that attempt to learn a probability model of the data distribution
- [Vapnik] "***one should solve the (classification/recognition) problem directly and never solve a more general problem as an intermediate step***"

[Vapnik]: V. Vapnik, Statistical Learning Theory, 1998

Discriminative Learning

- Two central issues in developing discriminative learning methods:

    1. Constructing suitable objective functions for optimisation

    2. Developing optimization techniques for these objective functions

Estimating acoustic model parameters

- If A: speech utterance and O_A: acoustic features corresponding to the utterance A,

$$W^* = \arg\max_W P_\lambda(O_A|W)P_\beta(W)$$

- ASR decoding: Return the word sequence that jointly assigns the highest probability to O_A

- How do we estimate λ in $P_\lambda(O_A|W)$?
    - MLE estimation
    - MMI estimation
    - MPE/MWE estimation

Estimating acoustic model parameters

- If A: speech utterance and O_A: acoustic features corresponding to the utterance A,

$$W^* = \arg\max_W P_\lambda(O_A|W)P_\beta(W)$$

- ASR decoding: Return the word sequence that jointly assigns the highest probability to O_A

- How do we estimate λ in $P_\lambda(O_A|W)$?
    - MLE estimation
    - MMI estimation
    - MPE/MWE estimation

Covered in this class

Maximum mutual information (MMI) estimation: Discriminative Training

- MMI aims to directly maximise the posterior probability (criterion also referred to as conditional maximum likelihood)

$$\mathcal{F}_{\mathrm{MMI}} = \sum_{i=1}^{N} \log P_\lambda(W_i|X_i)$$

$$= \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|W_i)P(W_i)}{\sum_{W_j} P_\lambda(X_i|W_j)P(W_j)}$$

- P(W) is the language model probability

Why is it called MMI?

- Mutual information $I(X, W)$ between acoustic data X and word labels W is defined as:

$$
\begin{aligned}
I(X,W) &= \sum_{X,W} \Pr(X,W) \log \frac{\Pr(X,W)}{\Pr(X)\Pr(W)} \\
&= \sum_{X,W} \Pr(X,W) \log \frac{\Pr(W|X)}{\Pr(W)} \\
&= H(W) - H(W|X)
\end{aligned}
$$

where H(W) is the entropy of W and H(W|X) is the conditional entropy

Why is it called MMI?

- Assume H(W) is given via the language model. Then, maximizing mutual information becomes equivalent to minimising conditional entropy

$$H(W|X) = -\frac{1}{N}\sum_{i=1}^{N} \log \Pr(W_i|X_i)$$
$$= -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\Pr(X_i|W_i)\Pr(W_i)}{\sum_{W'} \Pr(X_i|W')\Pr(W')}$$

- Thus, MMI is equivalent to maximizing:

$$\mathcal{F}_{\mathrm{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|W_i)P(W_i)}{\sum_{W_j} P_\lambda(X_i|W_j)P(W_j)}$$

MMI estimation

How do we compute this?

$$\mathcal{F}_{\mathrm{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|W_i)P(W_i)}{\sum_{W_j} P_\lambda(X_i|W_j)P(W_j)}$$

- Numerator: Likelihood of data given correct word sequence
- Denominator: Total likelihood of the data given all possible word sequences

Recall: Word Lattices

- A word lattice is a pruned version of the decoding graph for an utterance
- Acyclic directed graph with arc costs computed from acoustic model and language model scores
- Lattice nodes implicitly capture information about time within the utterance

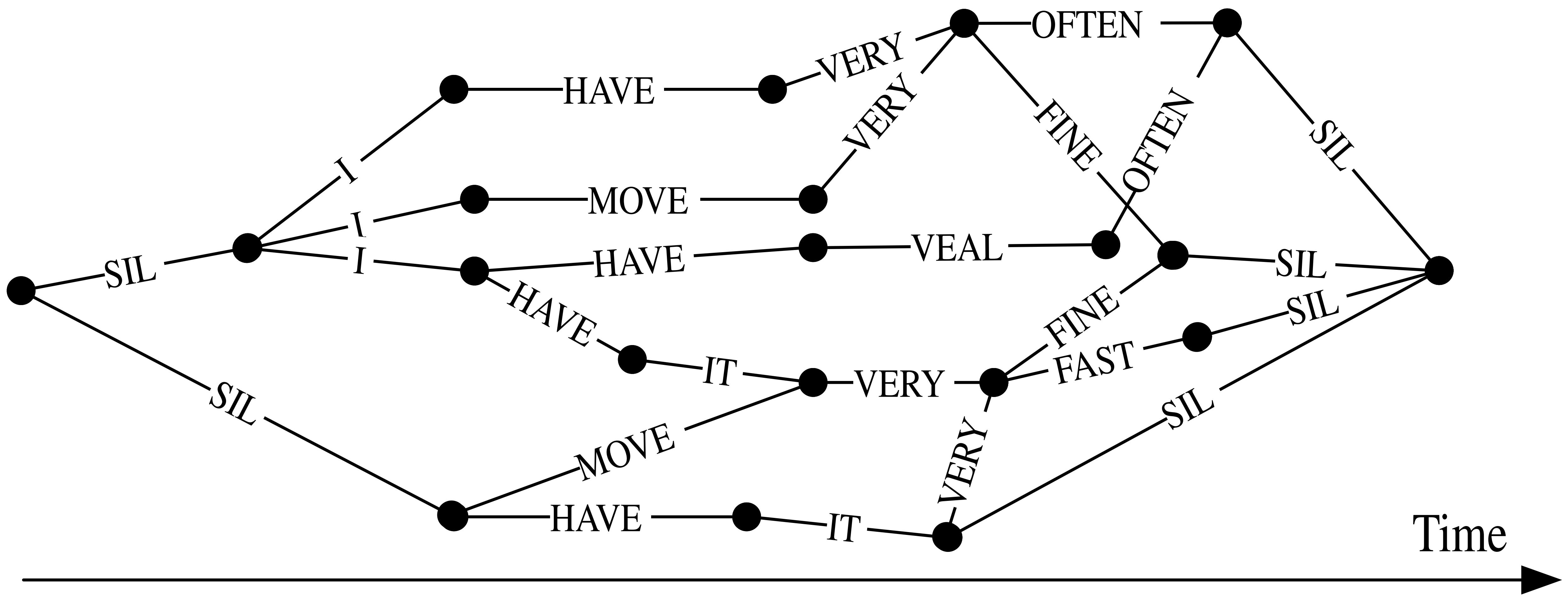


Image from [GY08]: Gales & Young, Application of HMMs in speech recognition, NOW book, 2008

MMI estimation

How do we compute this?

$$\mathcal{F}_{\text{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|W_i)P(W_i)}{\boxed{\sum_{W_j} P_\lambda(X_i|W_j)P(W_j)}}$$

- Numerator: Likelihood of data given correct word sequence
- Denominator: Total likelihood of the data given all possible word sequences
    - Estimate by generating lattices, and summing over all the word sequences in the lattice

MMI Training and Lattices

- Computing the denominator: Estimate by generating lattices, and summing over all the words in the lattice
- Numerator lattices: Restrict G to a linear chain acceptor representing the words in the correct word sequence. Lattices are usually only computed once for MMI training.
- HMM parameter estimation for MMI uses the extended Baum-Welch algorithm [V96,WP00]
- Like HMMs, can DNNs also be trained with an MMI-type objective function? Yes!

[V96]:Valtchev et al., Lattice-based discriminative training for large vocabulary speech recognition, 1996

[WP00]: Woodland and Povey, Large scale discriminative training for speech recognition, 2000

Sequence-discriminative (MMI) Training of DNNs

- In a hybrid system, DNNs are typically trained to optimise the cross-entropy objective function using SGD
- We could maximise MMI instead, that is maximise the mutual information between the distributions of the observation and word sequences
- Compute gradients of the MMI objective function with respect to the activations at the output layer

[V et al.]:Vesely et al., Sequence discriminative training of DNNs, Interspeech 2013

MMI results on Switchboard

- Switchboard results on two eval sets (SWB, CHE). Trained on 300 hours of speech. Comparing maximum likelihood (ML) against discriminatively trained GMM systems and MMI-trained DNNs.

	SWB	CHE	Total
GMM ML	21.2	36.4	28.8
GMM MMI	18.6	33.0	25.8
DNN CE	14.2	25.7	20.0
DNN MMI	12.9	24.6	18.8

[V et al.]:Vesely et al., Sequence discriminative training of DNNs, Interspeech 2013

Another Discriminative Training Objective: Minimum Phone/Word Error (MPE/MWE)

- MMI is an optimisation criterion at the sentence-level. Change the criterion so that it is directly related to sub-sentence (i.e. word or phone) error rate.

- MPE/MWE objective function is defined as:

$$\mathcal{F}_{\mathrm{MPE/MWE}} = \sum_{i=1}^{N} \frac{\sum_{W} P_\lambda(X_i|W_i)P(W_i)A(W,W_i)}{\sum_{W'} P_\lambda(X_i|W')P(W')}$$

where $A(W, W_i)$ is phone/word accuracy of the sentence W given the reference sentence W_i i.e. the total phone count in W_i minus the sum of insertion/deletion/substitution errors of W

MPE/MWE training

$$\mathcal{F}_{\mathrm{MPE/MWE}} = \sum_{i=1}^{N} \log \frac{\sum_{W} P_\lambda(X_i|W_i)P(W_i)A(W, W_i)}{\sum_{W'} P_\lambda(X_i|W')P(W')}$$

- The MPE/MWE criterion is a weighted average of the phone/word accuracy over all the training instances
- A(W, W_i) can be computed either at the phone or word level for the MPE or MWE criterion, respectively
- The weighting given by MPE/MWE depends on the number of incorrect phones/words in the string while MMI looks at whether the entire sentence is correct or not

MPE results on Switchboard (GMMs)

- Switchboard results on eval set SWB. Trained on 68 hours of speech. Comparing maximum likelihood (MLE) against discriminatively trained (MMI/MPE/MWE) GMM systems

	SWB	%WER redn
GMM MLE	46.6	-
GMM MMI	44.3	2.3
GMM MPE	43.1	3.5
GMM MWE	43.3	3.3

[V96]:Valtchev et al., Lattice-based discriminative training for large vocabulary speech recognition, 1996

[WP00]: Woodland and Povey, Large scale discriminative training for speech recognition, 2000

Sequence-discriminative training results on Switchboard (DNNs)

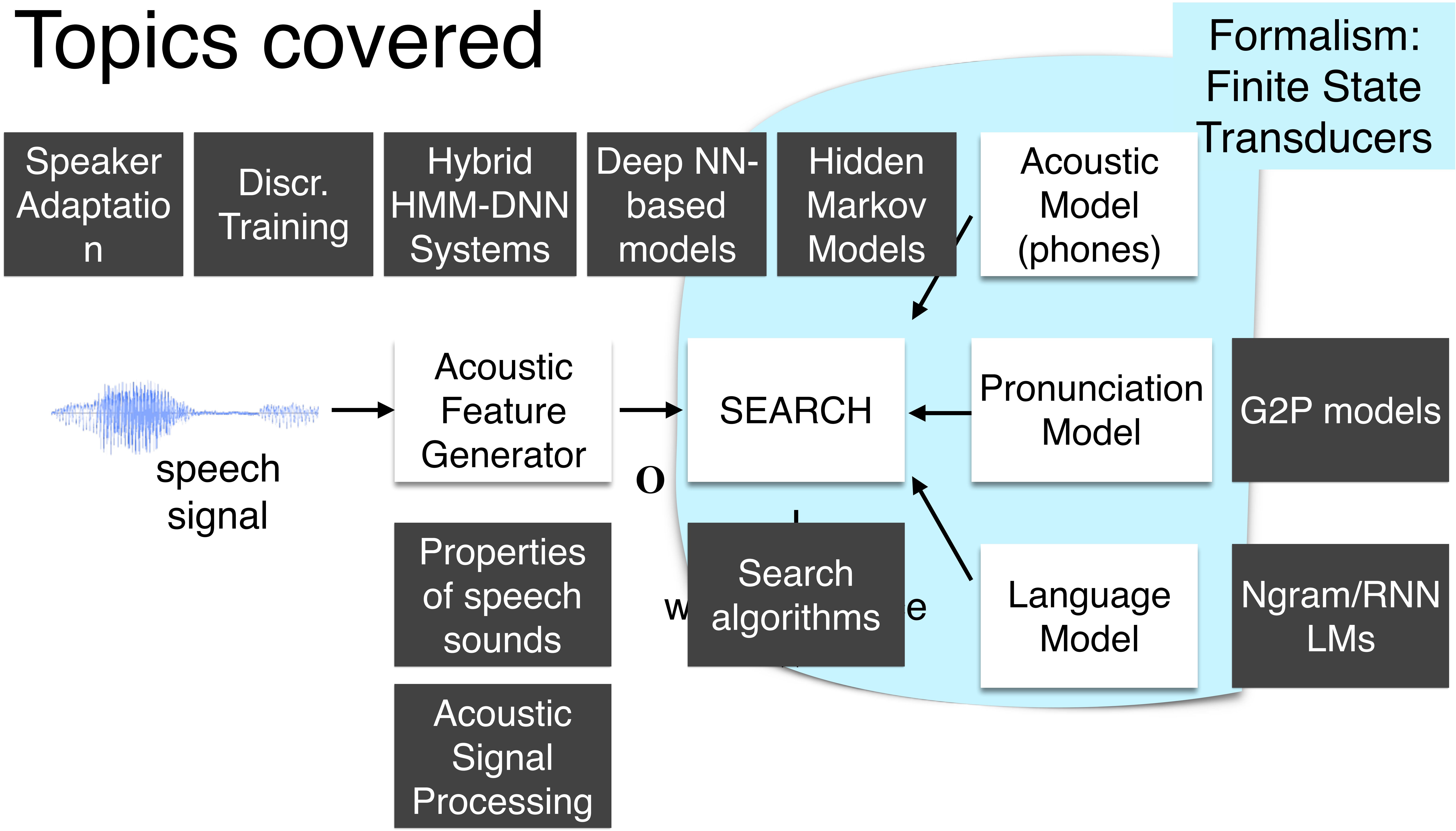
- Switchboard results from DNNs trained on the full 300 hour training set, using different optimization criteria

	SWB	CHE	Total
GMM MMI	18.6	33.0	25.8
DNN CE	14.2	25.7	20.0
DNN MMI	12.9	24.6	18.8
DNN sMBR	12.6	24.1	18.4
DNN MPE	12.9	24.1	18.5

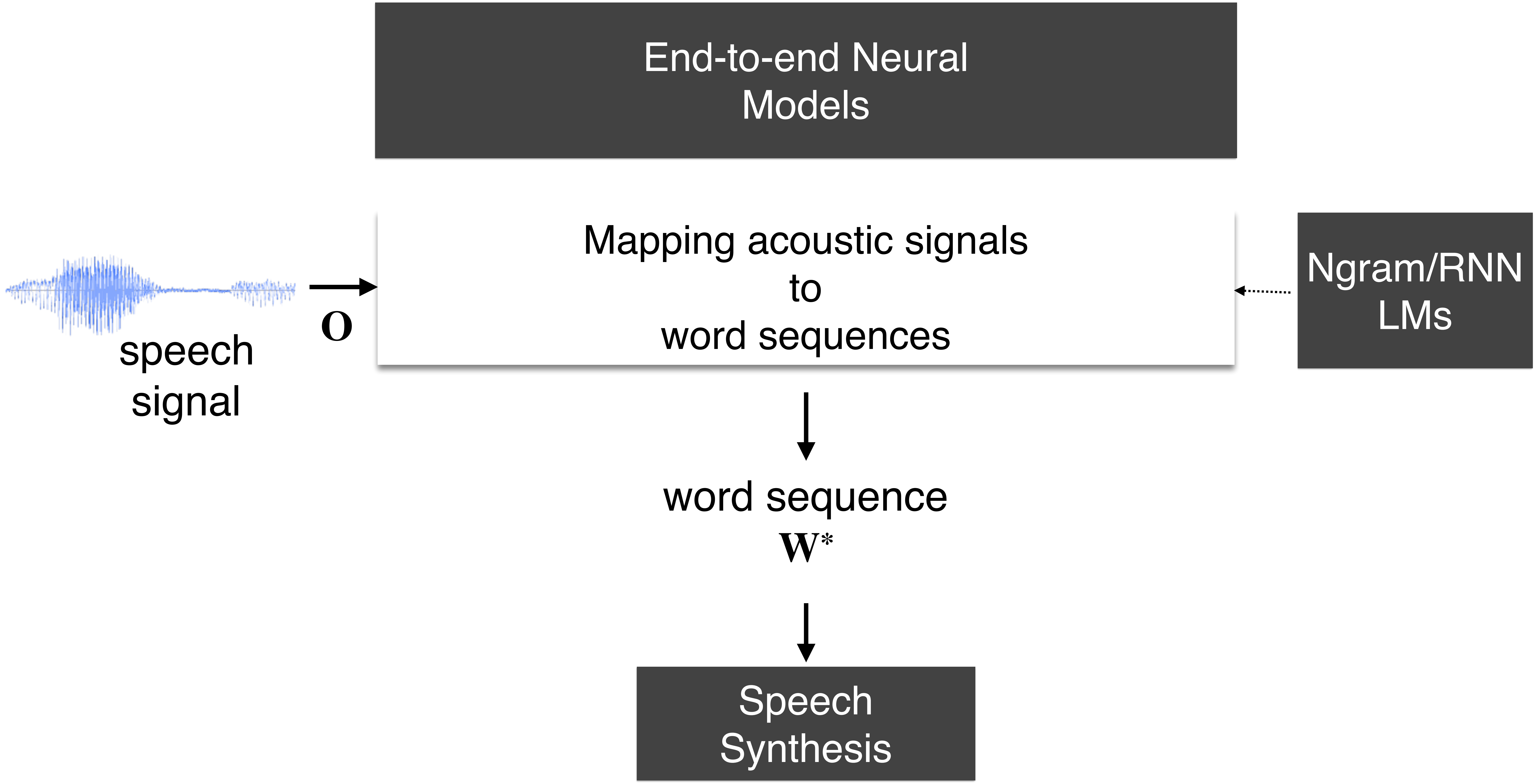
[V et al.]:Vesely et al., Sequence discriminative training of DNNs, Interspeech 2013

CS-753 Concluding Remarks

Topics covered

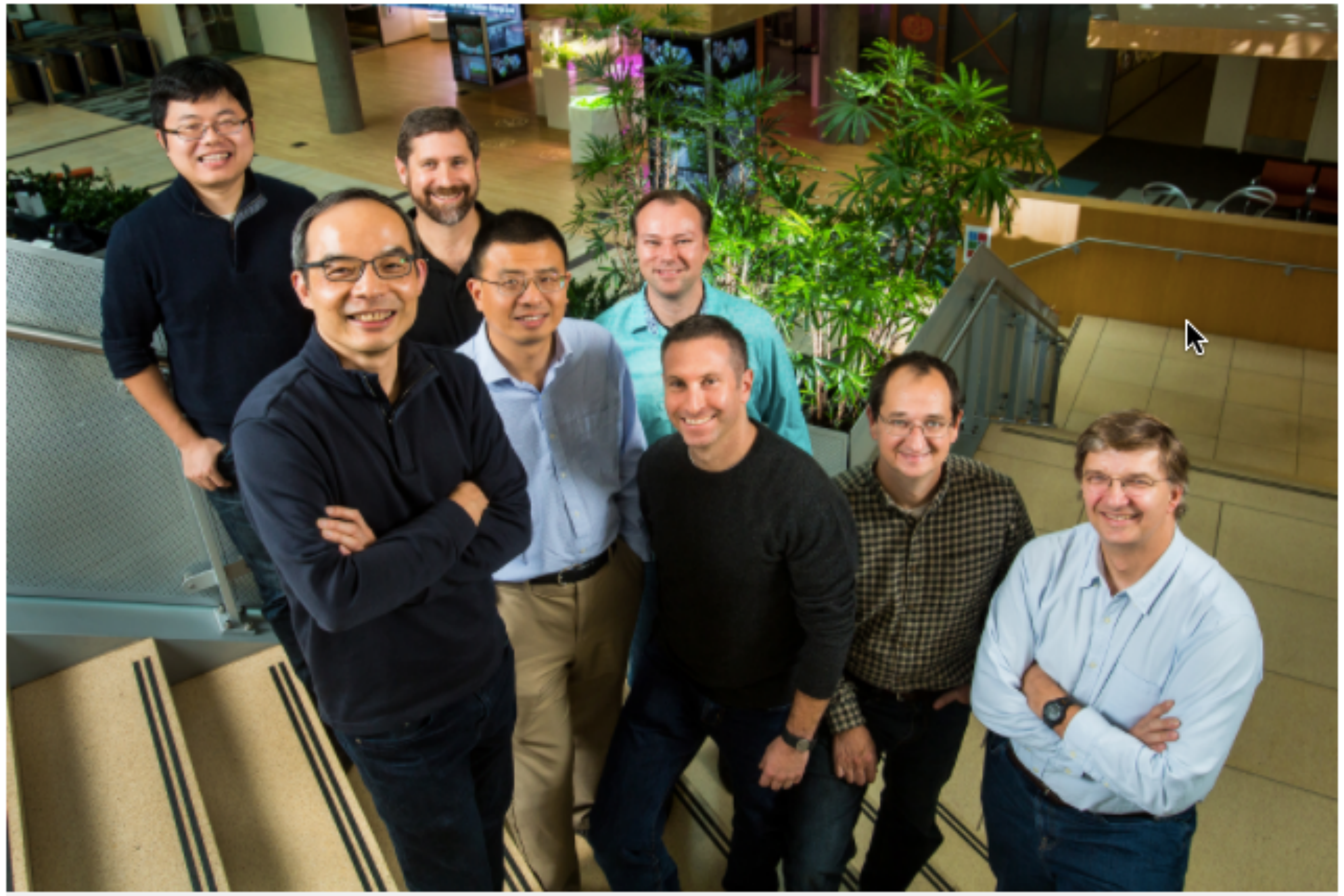


Topics covered



Exciting time to do speech research

Historic Achievement: Microsoft researchers reach human parity in conversational speech recognition



Google Cloud Platform

CLOUD SPEECH API BETA

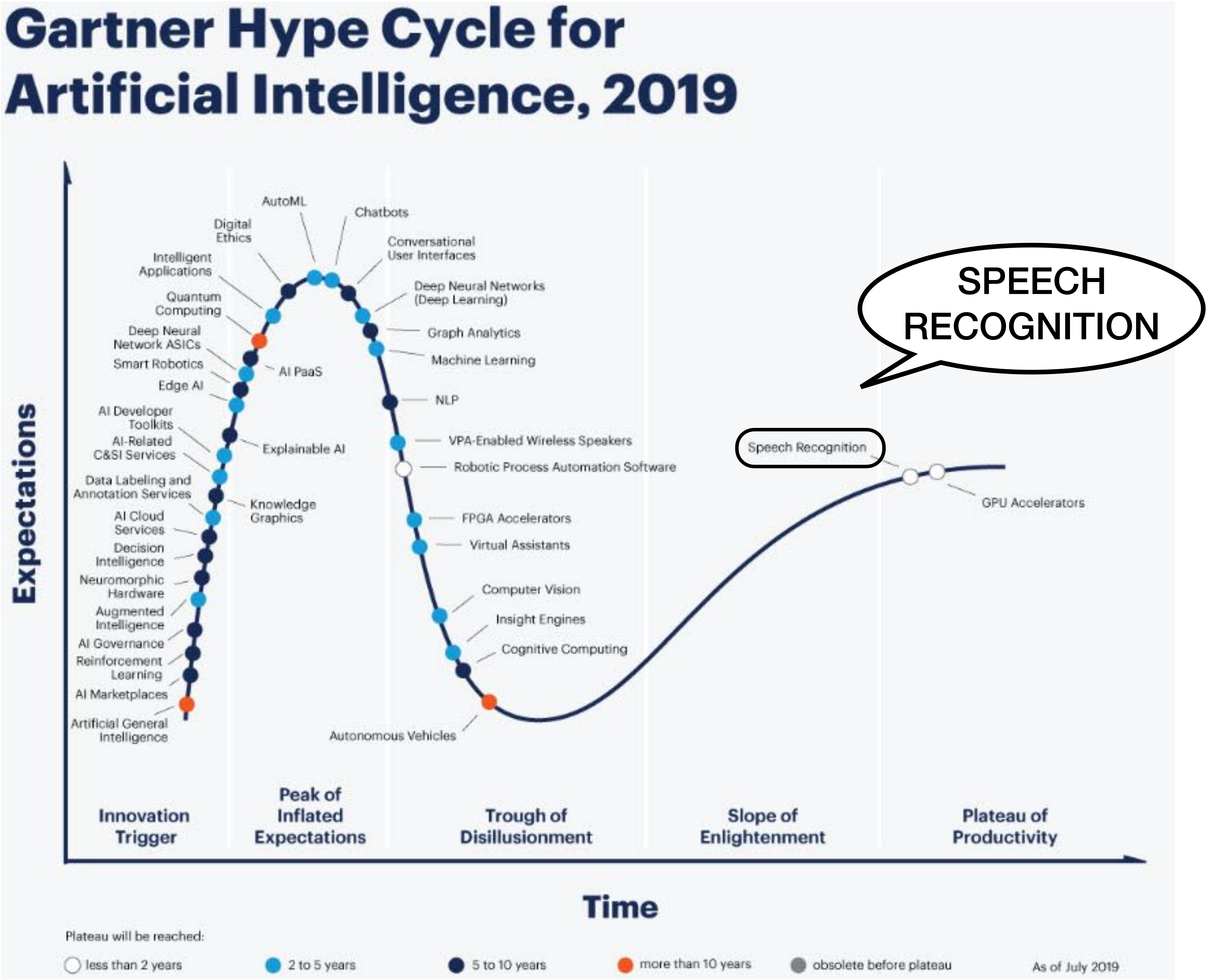
Speech to text conversion powered by machine learning

TRY IT FREE   VIEW DOCUMENTATION

Powerful Speech Recognition

Google Cloud Speech API enables developers to **convert audio to text** by applying **powerful neural network models** in an easy to use API. The API **recognizes over 80 languages and variants,** to support your global user base. You can transcribe the text of users dictating to an application's microphone, enable command-and-control through voice, or transcribe audio files, among many other use cases. **Recognize audio uploaded in the request,** and integrate with your audio storage on Google Cloud Storage, by using the same technology Google uses to power its own products.

Called Hype Cycle for a reason…



What's next?

Need to do more…

- Robust to variations in age, accent and ability
- Handling noisy real-life settings with many speakers (e.g., meetings, parties)
- Handling pronunciation variability
- Handling new languages/ dialects

E.g.: ASR on accented speech

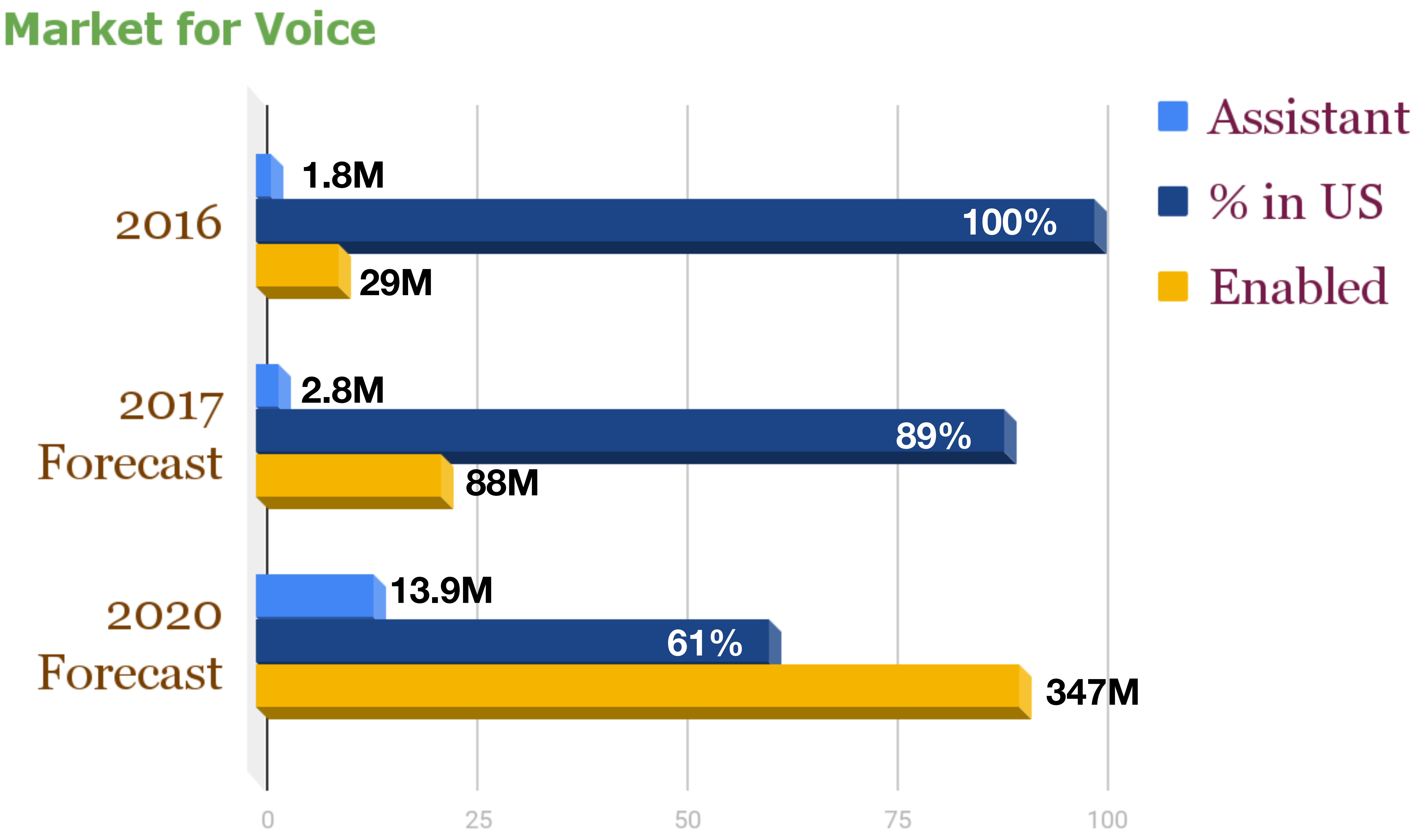
DESPITE THE JULY DECLINE DURABLE GOODS ORDERS REMAINS SEVEN POINT SEVEN PERCENT ABOVE THE YEAR EARLIER LEVEL

WER 3%

DESPITE THE JULY DECLINE TO <UNK> ITS AUGUST REMAINED SEVEN POINT SEVEN OH CENT LEVEL THAT THE ABILITY OF THAT

WER 21%

Speech interfaces



What's next?

Need to do more…

- Robust to variations in age, accent and ability
- Handling noisy real-life settings with many speakers (e.g., meetings, parties)
- Handling pronunciation variability
- Handling new languages/dialects

… with less

- Fast (real-time) decoding using limited computational power/memory
- Faster training algorithms
- Reduce duplicated effort across domains/languages
- Reduce dependence on language-specific resources
- Train with less labeled data

Remaining Coursework

Participation Points

- Six in-class mini-quizzes
- Total points out of 20
  (Quiz 2 scaled to 4 points)
- $\geq$10 points gets full 5 participation points
- [8-10) — 4
  [6-8) — 3
  [4-6) — 2
  [2-4) — 1
  < 2 — 0

Quiz	Points	# of responses
1	3	96
2	10	79
3	4	99
4	4	76
5	2	68
6	3	53

Final Exam Syllabus

1. WFST algorithms/WFSTs used in ASR
2. HMM algorithms/EM/Tied state Triphone models
3. DNN-based acoustic models
4. N-gram/Smoothing/RNN language models
5. End-to-end ASR (CTC, LAS, RNN-T)
6. MFCC feature extraction
7. Search & Decoding
8. HMM-based speech synthesis models
9. Multilingual ASR
10. Speaker Adaptation
11. Discriminative training of HMMs

Questions can be asked on any of the 11 topics listed above. You will be allowed a single A-4 cheat sheet of **handwritten notes**; content on both sides permitted.

Final Project

<u>Deliverables</u>

- 4-5 page final report:
  - ✓ Task definition, Methodology, Prior work, Implementation Details, Experimental Setup, Experiments and Discussion, Error Analysis (if any), Summary
- Short talk summarizing the project:
  - ✓ Each team will get 8-10 minutes for their presentation and ≈5 minutes for Q/A
  - ✓ Clearly demarcate which team member worked on what part

Final Project Grading

- Break-up of 20 points:
  - 6 points for the report
  - 4 points for the presentation
  - 6 points for Q/A
  - 4 points for overall evaluation of the project

Final Project Schedule

- Presentations will be held on Nov 23rd and Nov 24th
- The final report in pdf format should be sent to `pjyothi@cse.iitb.ac.in` before Nov 24th
- The order of presentations will be decided on a lottery basis and shared via Moodle before Nov 9th