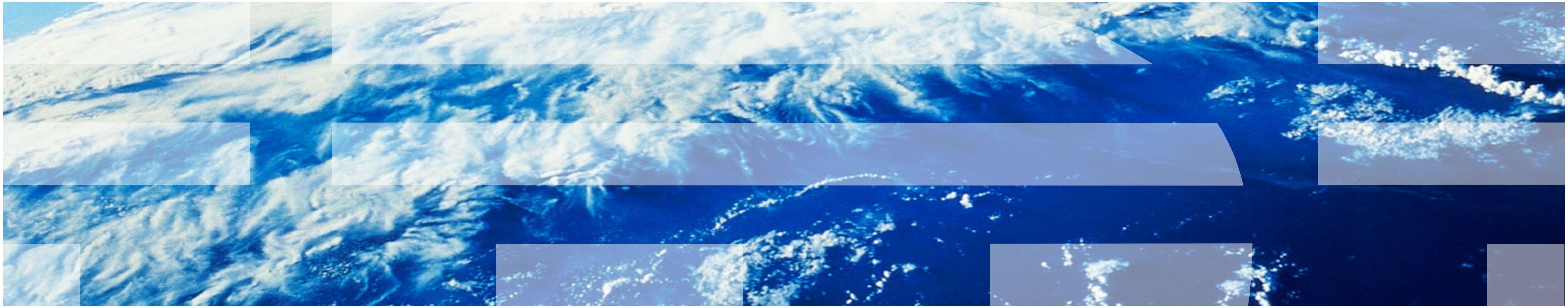
Konuşma Tanıma: Geriye ne kaldı?
(Speech Recognition: What's Left?)

Dr. Michael Picheny
Senior Manager, Speech Technologies
IBM Research AI
IBM TJ Watson Research Center



Bu konuşmayı yapmak üzere beni davet ettiğiniz ve misafirperverliğiniz için teşekkür ederim!

Inspirations for this Talk

- My two thesis advisors at MIT, Nat Durlach (left, deceased) and Lou Braida (right) (1993)
- Both honored at the Acoustical Society of America in Boston (June 2017) with two special sessions
- Fundamental contributions in Psychoacoustics and Sensory Communication Aids
- Taught me how to scientifically assess aspects of human perception
- Learned how to do research from them – to be thorough and to question

IBM has a Long History of Innovations in AI

First working chess program

Bernstein (1957)

First demonstration of machine learning (checkers)

Samuel (1959, 1967)

First demonstration of neural network with reinforcement learning in complex domain (TD-gammon)

Tesauro (1995)

First computer to defeat world chess champion (Deep Blue)

Campbell, Hoane & Hsu (1997)

First computer to defeat best human Jeopardy! Players (Watson)

Ferrucci, et al. (2011)

Some AI challenges we are tackling today at IBM Research AI

Media

Create highlights of sports events

Compliance

Is my organization compliant with latest regulatory documents

Industrial

Guide me through fixing malfunctioning components

Visual Inspection

Find rust on electric towers, using drones

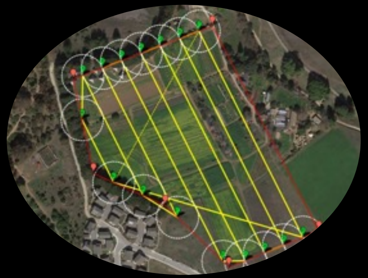
Customer Care

Bot that can guide a user through buying the right insurance policy

Marketing / Business

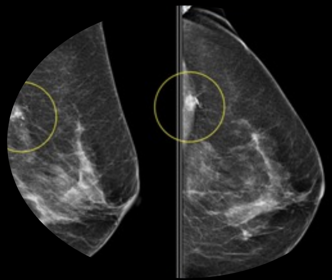
Summarize the strategic intent of a company based on recent news articles

IoT



Predict yield of field based on images and sensor data

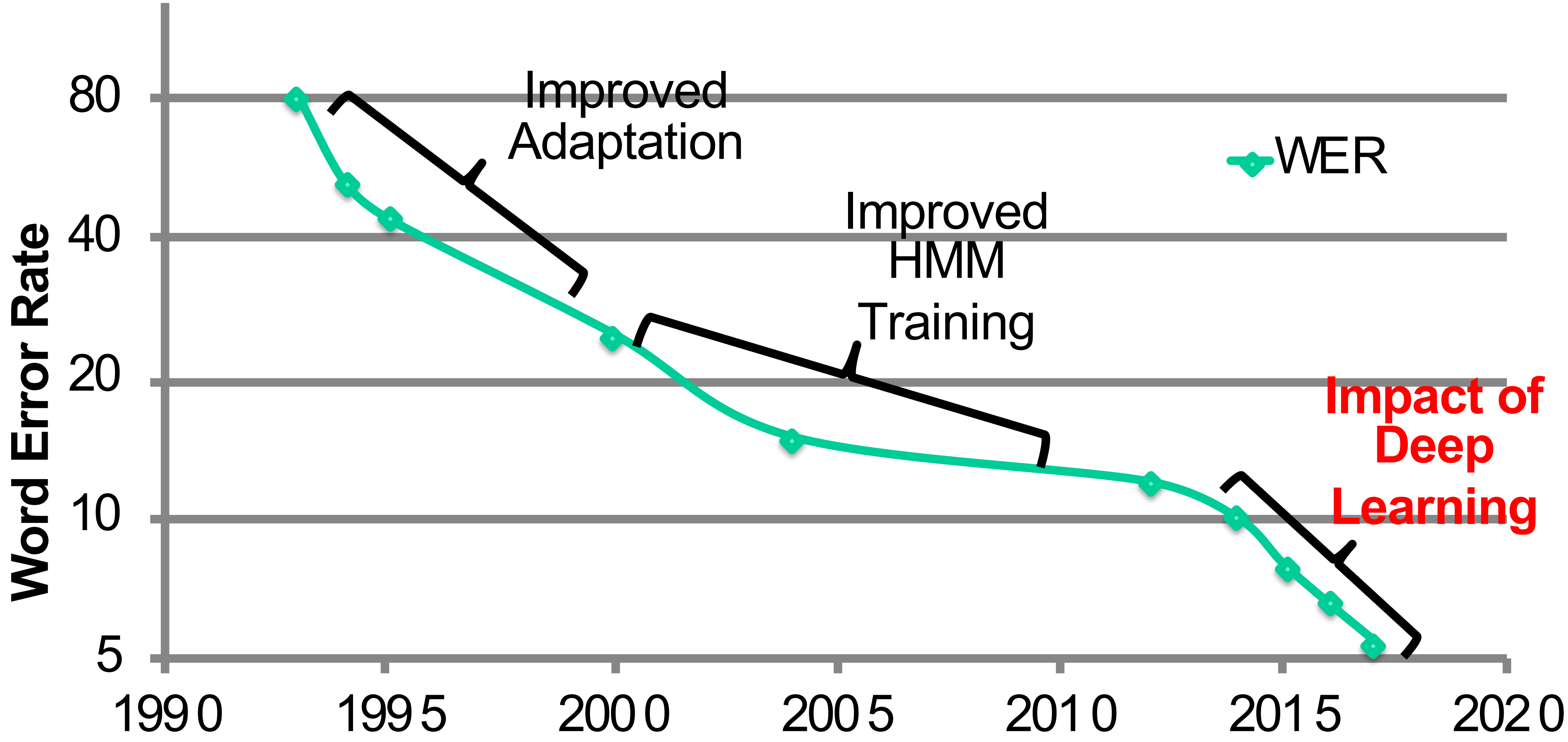
Healthcare



Improve the accuracy of breast cancer screening

Historical Performance in Speech Recognition

- Task is **transcription** of “SWITCHBOARD” – Human-Human Landline Telephone conversations on directed topics
- SWITCHBOARD is a popular public benchmark in the Speech Recognition Community
    - Difficult enough to present challenges but clearly understandable by humans

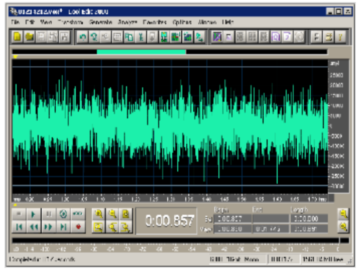


Why has Speech Recognition Proven so Difficult?

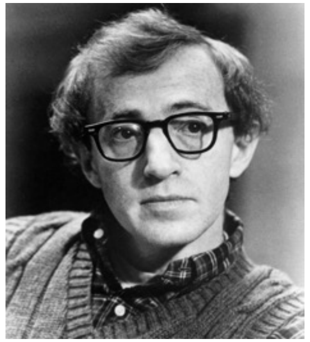
Speaker Variation

Channel Variation

Background Noise

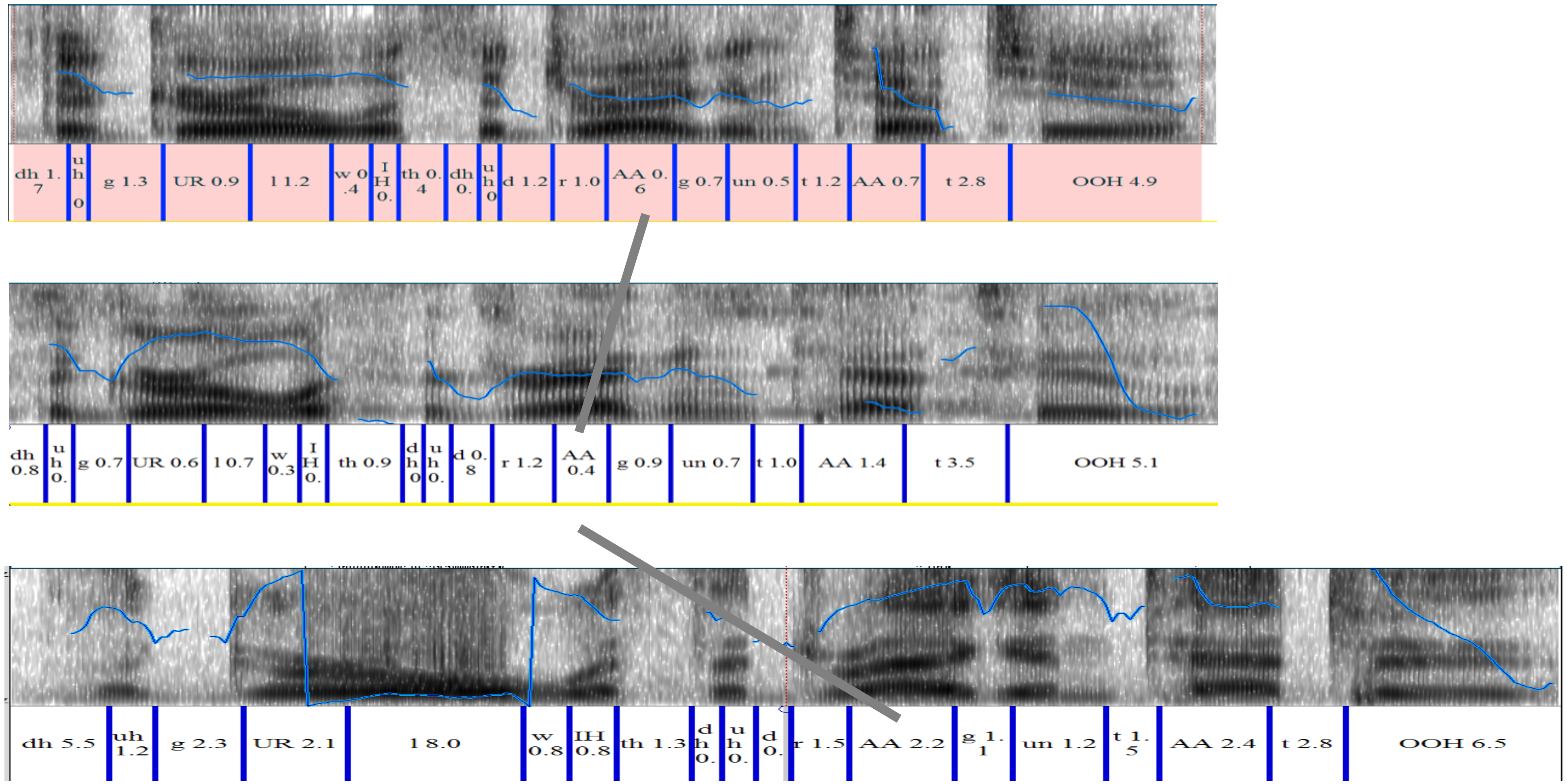


Accent



Speaking Style

Huge Acoustic Variability for Same Underlying Text



"The Girl with the Dragon Tattoo"

Inherent variability of Speech biggest challenge

Basic Formulation of Speech Transcription Problem

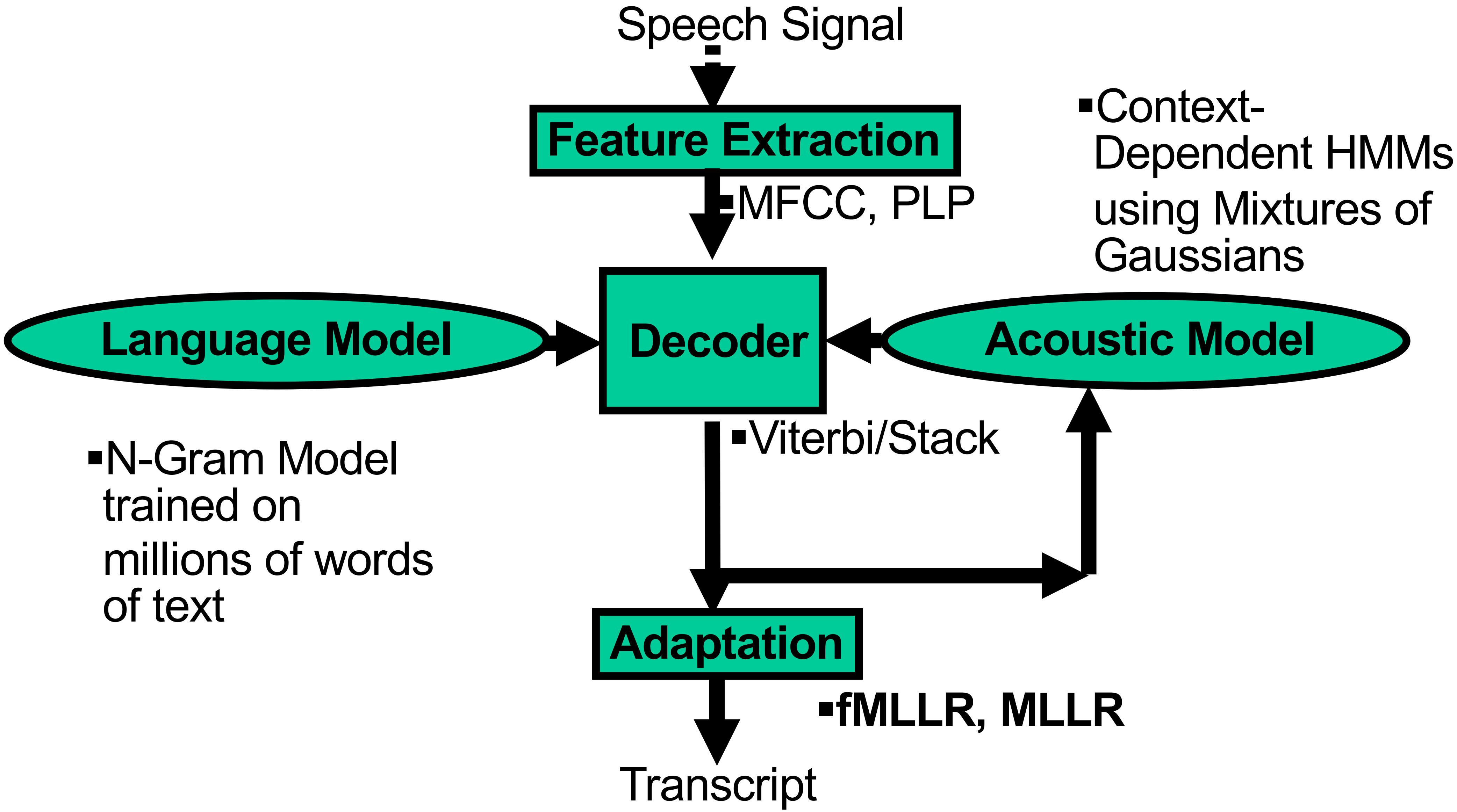
Choose W to maximize:

$$P(W|X) = \frac{P(X|W)\,P(W)}{\cancel{P(X)}}$$

W = vocabulary
X = extracted features from the speech signal
P(X|W) = Acoustic Model
P(W) = Language Model

Hypothesis Search

Traditional Speech Recognition System (pre-2011)



How do we build a transcription system? (vocabulary)

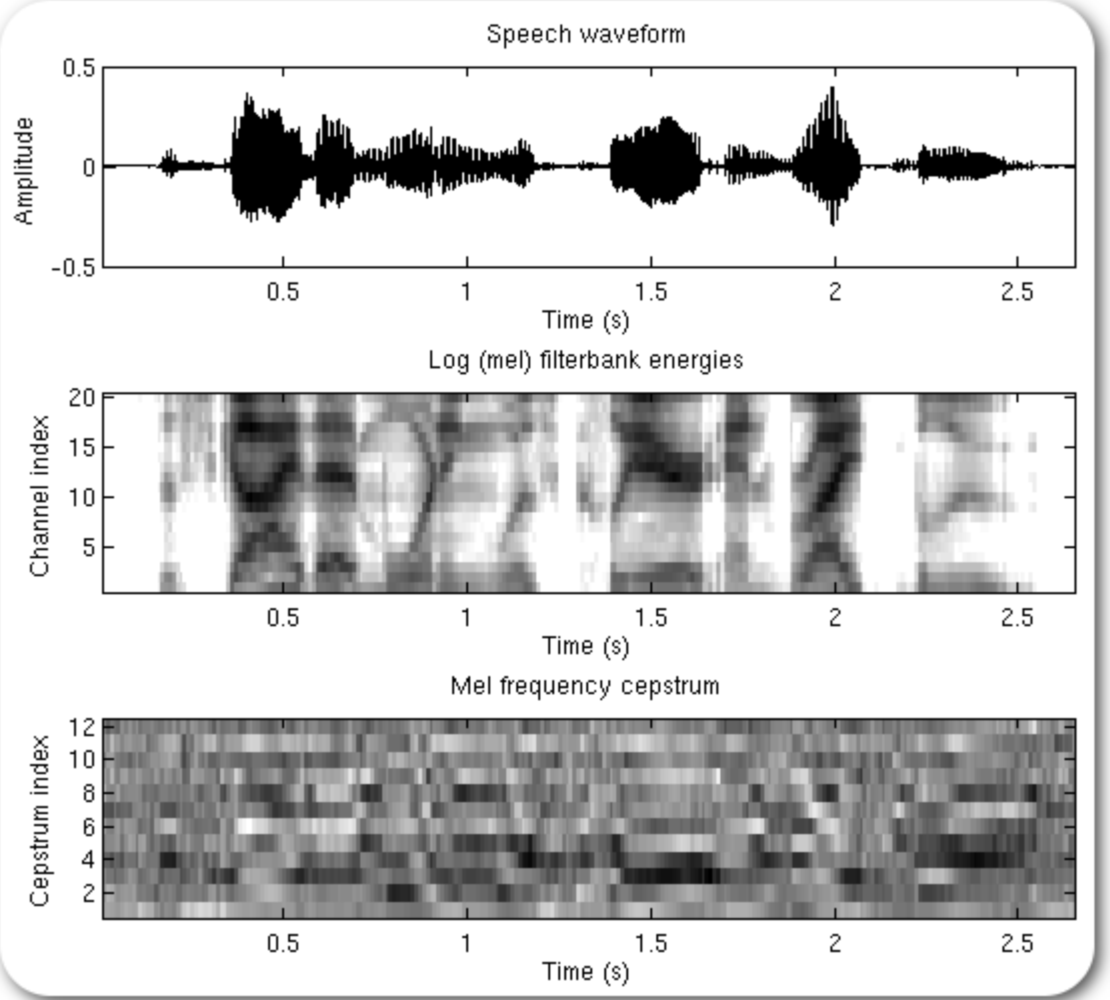
- Choosing a vocabulary
  - Take a lot of text, count number of words, take most frequent
  - Can also look at intersection of frequently occurring words in diverse corpora (e.g., news stories vs conversations)
- Lexicon (Mapping from word spellings to pronunciations) issues
  - Words may have multiple pronunciations – Tomayto vs Tomahto
  - Pronunciation hard to predict from orthography – e.g. “through”
  - Text may have misspelings (err….mispellings ☺)

How do we build a transcription system? (vocabulary, “W”)

- Language issues
  - Arabic written w/o vowels
    - صباح الخير
    - “Good Morning”
  - Chinese written w/o white spaces between words
    - 中国圈养大熊猫今年将迎来历史上最好的繁殖期，目前已经有三十只大熊猫成功配对。
    - "This year will usher in the best breeding season in history for giant pandas in captivity, so far 30 giant pandas have been successfully paired."
- Recognizer cannot produce words outside vocabulary
  - Depending on task, vocabulary sizes from 5000-500000 words common
  - Computation does not grow linearly because many words share parts of other words
    - “house” vs “houseboat”

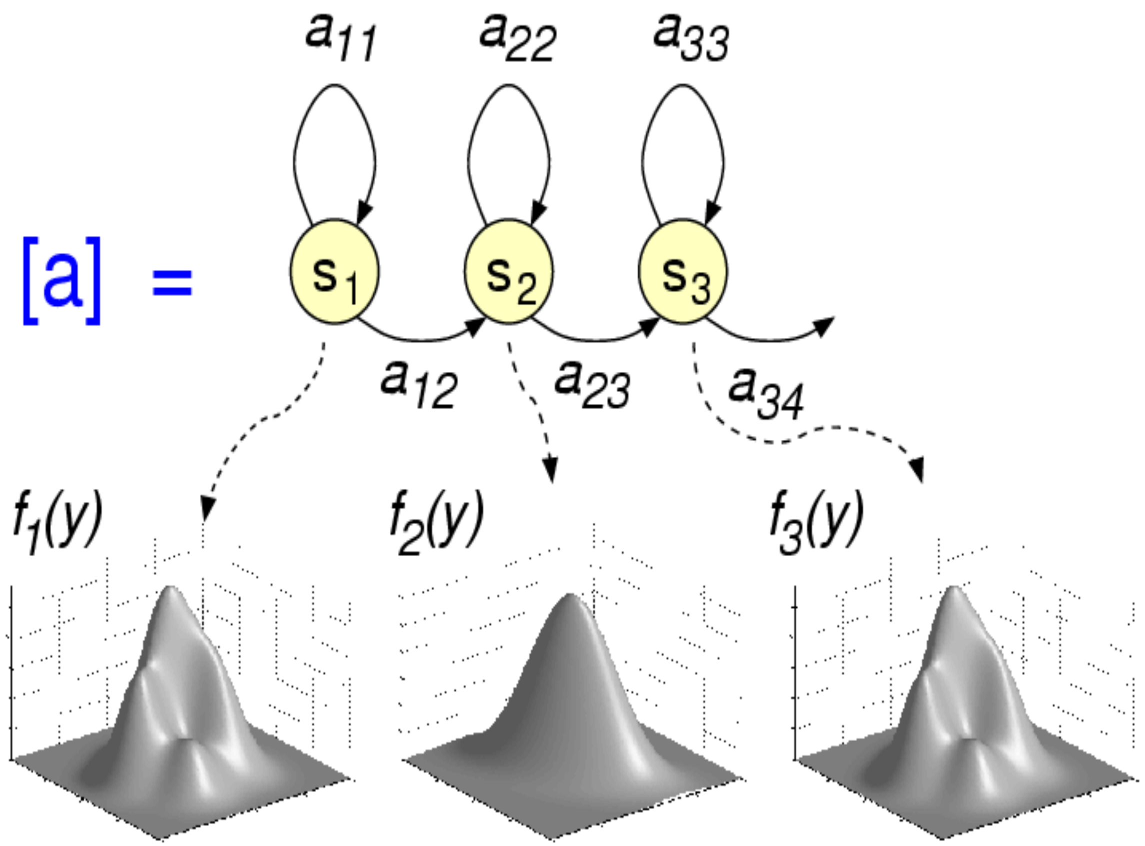
How do we build a transcription system? (features, “X”)

- Extract time-frequency features that are related to processing in human auditory system (“Mel Frequency Representation”, “Perceptual Linear Prediction”, etc.)
- De-Correlate the features to allow for easier modeling (“MFCCs”,)
  - Usually augmented with time derivatives of features



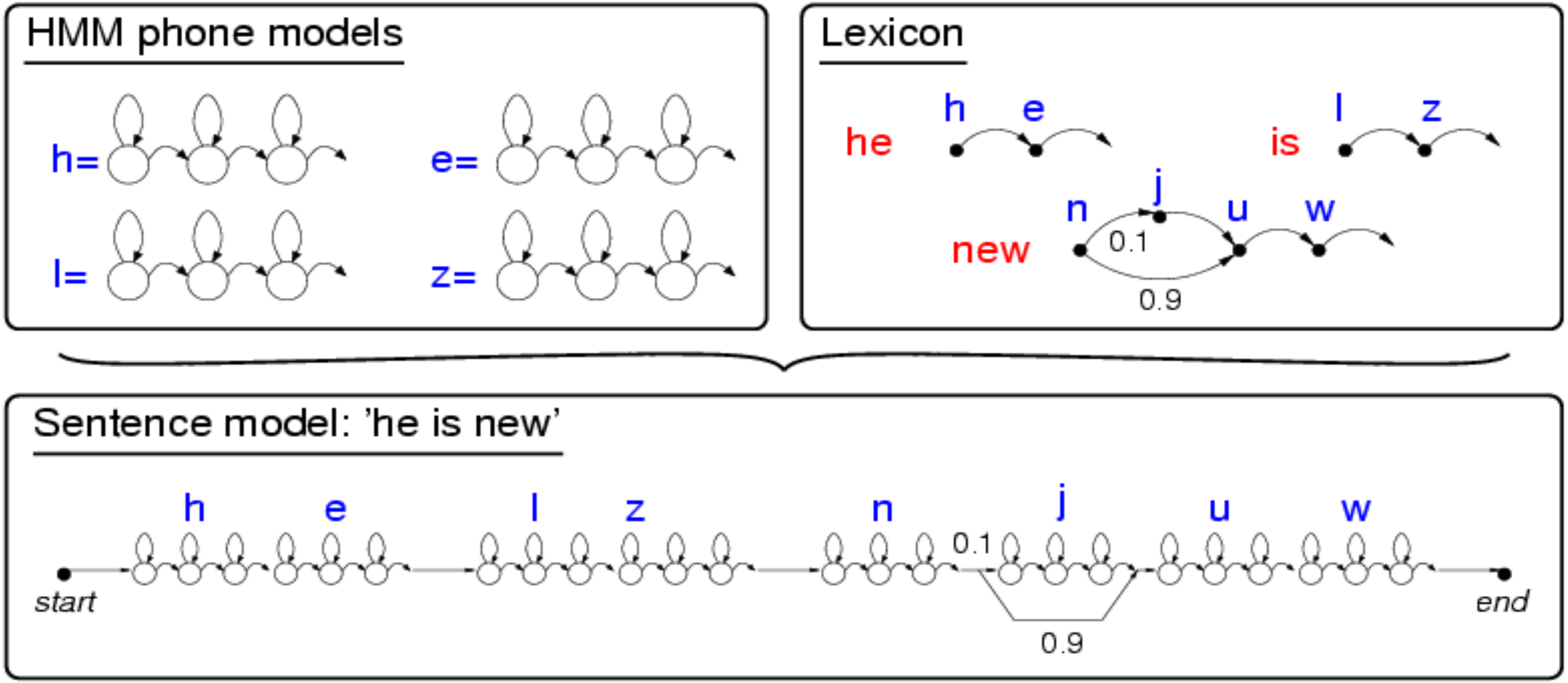
How do we build a transcription system? (Acoustic Model "P(X|W)")

Hidden Markov Models



Every time you take a transition, you output a feature vector x

How do we build a transcription system? (Acoustic Model)



How do we build a transcription system? (Acoustic Model)

- Build models for different sounds in different contexts
- Efficient algorithms exist to train the models from a set of transcripts and data
- Push-button toolkits exist that enable easy creation of such models.
- Additional enhancements include training algorithms targeted at improving discrimination power across words and phones rather than just increasing the likelihood of the training data.

How do we build a transcription system? (Language Model “P(W)”)

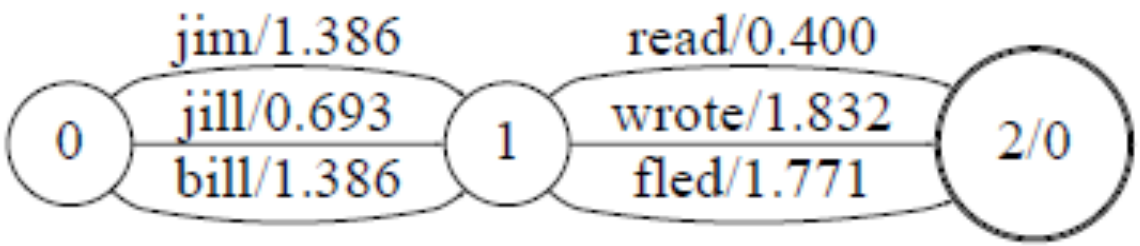
$$
\begin{aligned}
P(\omega = w_1 \cdots w_l) & \\
= \; & P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \cdots P(w_l|w_1 \cdots w_{l-1}) \\
= \; & \prod_{i=1}^{l} P(w_i|w_1 \cdots w_{i-1})
\end{aligned}
$$

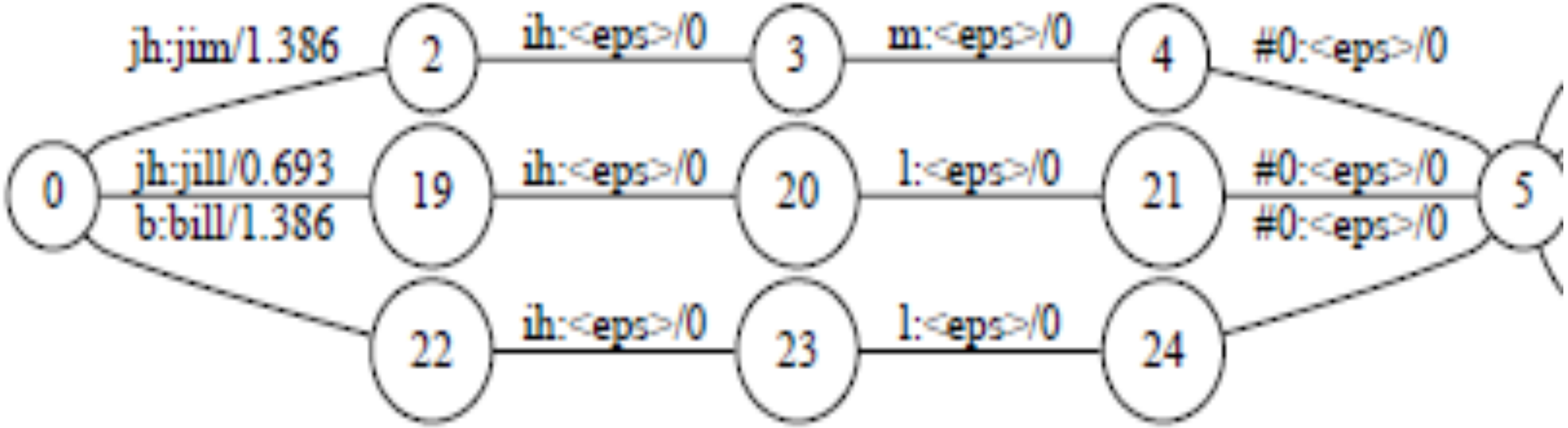
- Markov assumption: identity of next word depends only on last $n-1$ words, say n=3

$$P(w_i|w_1 \cdots w_{i-1}) \approx P(w_i|w_{i-2} w_{i-1})$$

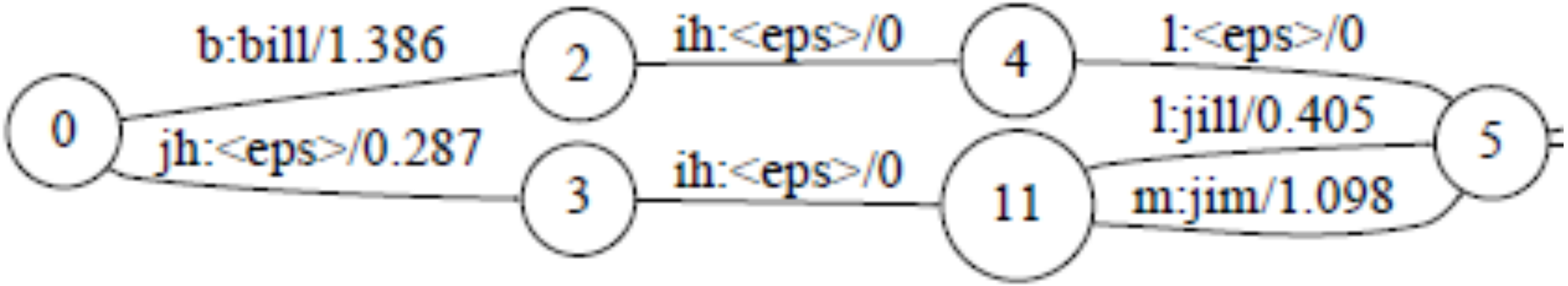
- Count the number of times a word occurs after a series of n words. Each separate context is called an “N-Gram”
- Typically built from millions or even billions of words of text. Small – 4M N-Grams
- Interesting to note that a single acoustic model suffices for a wide variety of applications but different LMs are needed for different situations

How do we build a transcription system? (Hypothesis Search)





How do we build a transcription system? (Hypothesis Search)



- Compile all knowledge sources into large graph, and simplify
- Efficient algorithms exists to search the graph.
- Some systems make multiple passes over the data with progressively more sophisticated models to reduce the overall computation
- Performance improvements can result by combining results of multiple systems together

Performance Metrics: How do we know if we are doing well?

- Obvious Success Metric – Word Error Rate (WER):
  – 100 x (Substitutions + Deletions + Insertions) / (Total Words in Reference transcripts)

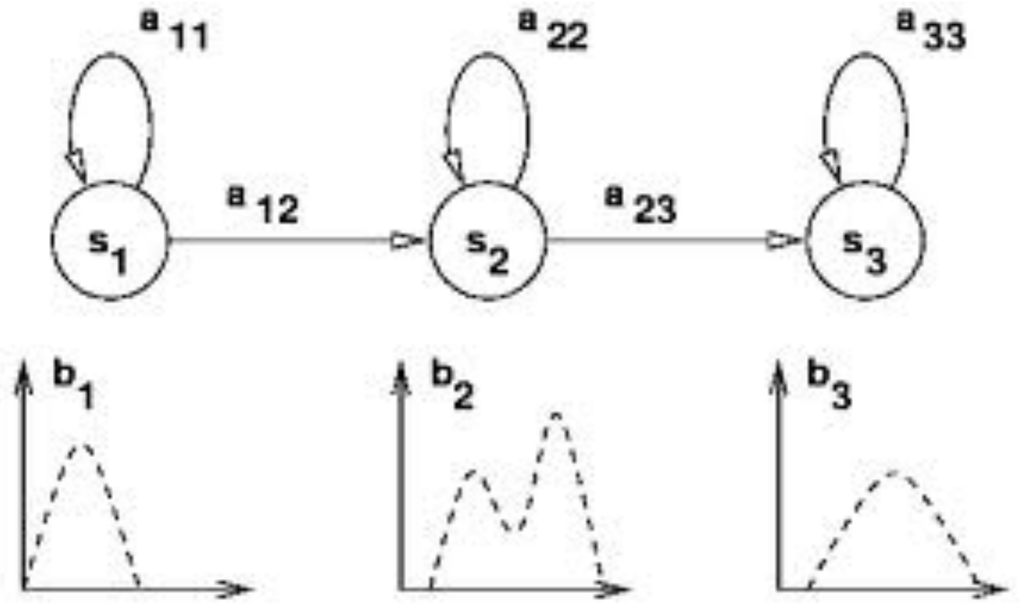
Ref:	THE	CAT	IN		THE		HAT
Hyp:		CAT	IS	ON	THE	GREEN	HAT
	Del		Sub	Ins		Ins	

Error rate = 100 x (1 S + 1 D + 2 I) / 5 = 80%

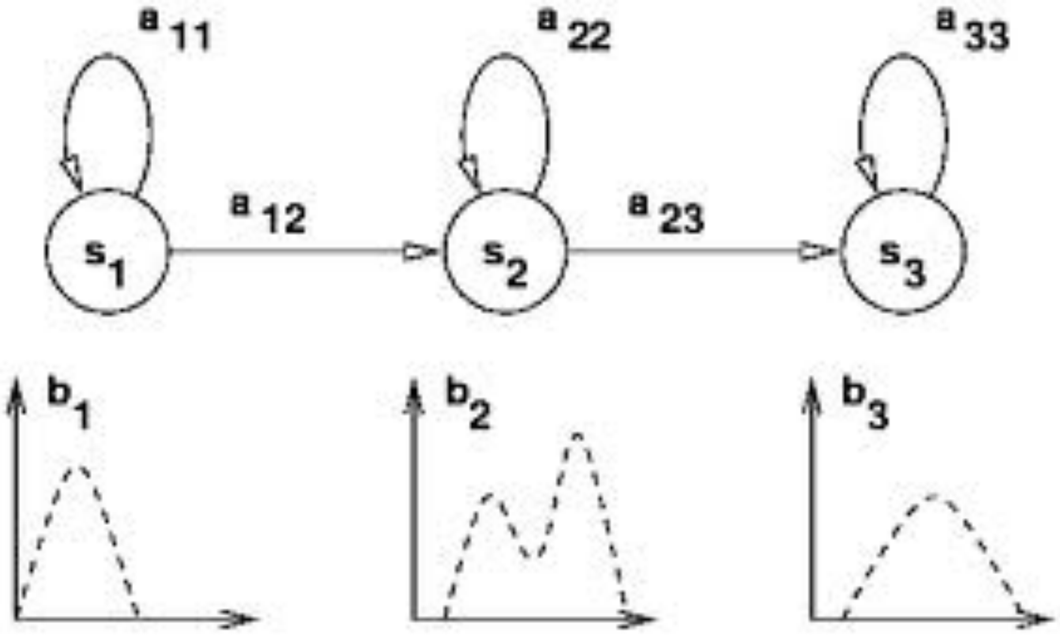
Neural Networks: Most Recent Driver of Improvements in Speech Recognition

Review:

- The acoustic model in speech recognition predicts p(**x**|**w**), the probability that a word **w** produces a sequence of observed feature vectors **x**
- A word is modeled as a sequence of phones using 3-state Hidden Markov Models; each HMM state corresponds to a context-dependent subphone unit c_i.
- Traditionally, the output distribution in each state has been modeled by a Gaussian Mixture Model (GMM) trained to maximize likelihood or discriminability.



Neural Networks: Most Recent Driver of Improvements in Speech Recognition



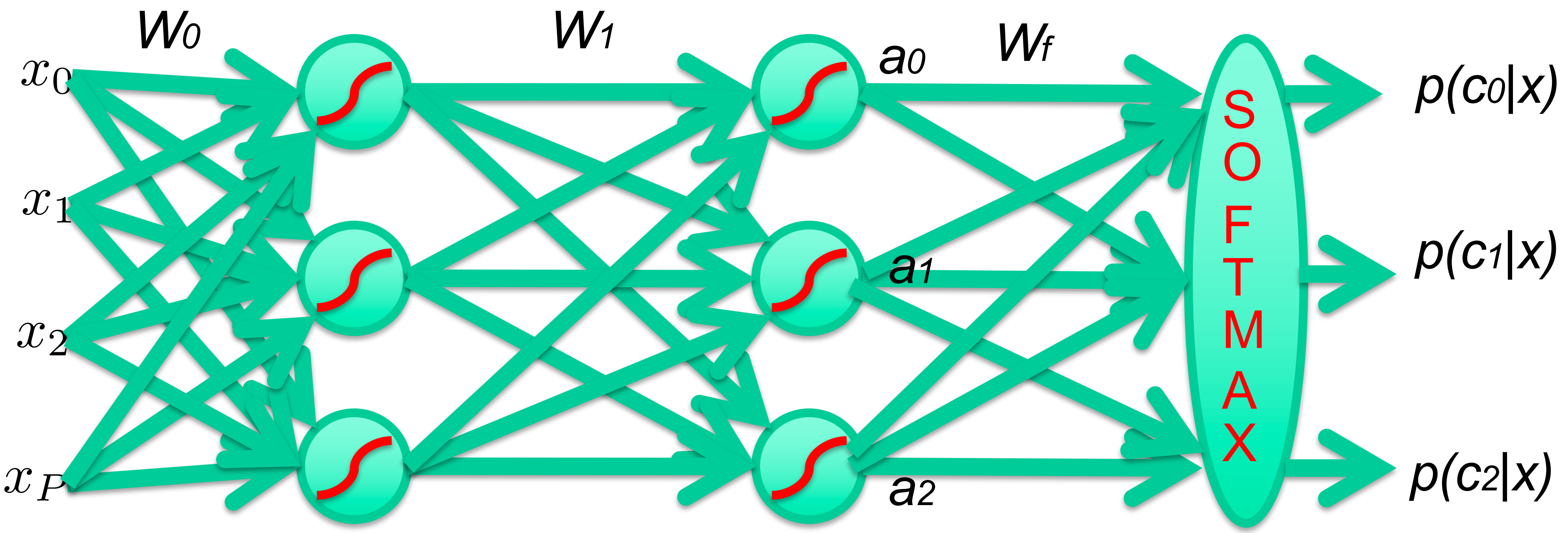
- Neural Networks can also be used for acoustic modeling instead of GMMs
  - Was originally tried in the early 1990s but until the onset of Deep Learning could not be made to perform as well as the GMMs

Multi-Layer Neural Network (aka "Feed-Forward" or "Deep Neural Network" – DNN)

- Neurons arranged in sequence of layers
- First layer inputs are the feature vector components (MFCC, PLP, etc)
- Final layer predicts the posterior probabilities of the sub-word classes c_i

$$p(c_i|x) = \frac{\exp\left(-(W_f\vec{a})_i\right)}{\sum_{j=1}^{n}\exp\left(-(W_f\vec{a})_j\right)}$$

"Softmax"



Training DNNs and Using them to Replace GMM Likelihoods

- Weights W in Neural networks are trained to minimize **Cross-Entropy** (CE) objective function
- $$L = -\sum_{i=1}^{N} y_{it}^{ref} \log p(c_i|x_t)$$
- $p(c_i \mid x_t)$ is the posterior probability that subphone c_i occurred at time t.
- y_t^{ref} is the target vector at time t.
  - "1" hot vector indicating occurrences of subphones over time.
  - Reference occurrences determined by alignment against set of existing models

	y				p			
c	1	0	0	0	.6	.1	.1	.1
a	0	1	0	0	.2	.7	.1	.1
t	0	0	1	0	.1	.1	.6	.3
s	0	0	0	1	.1	.1	.2	.5
					-.5	-.35	-.5	-.7

L = -2.05

Training DNNs and Using them to Replace GMM Likelihoods

- Training done using Stochastic Gradient Descent using back-propagation algorithm with computations migrated to GPUs for speed.
- NN gives posterior $p(c_i|x)$ so divide by class prior for subphone unit c_i to get likelihood

$$p(x|c_i) \sim \frac{\mathrm{p}(c_i|x)}{p(c_i)}$$

- NN likelihood can then replace the GMM likelihood as output distribution in the HMM (so-called **"Hybrid"** NN Acoustic Model)

Factors Affecting Neural Network Performance

Number of Predicted Subphone Units	WER
384	21.3
512	20.8
1024	19.4
2,220	**18.5**

Depth	WER
1	22.9
2	20.4
3	19.0
4	18.1
5	17.8
7	17.4

IBM Enhancement #1: Sequence Training

- Cross-entropy frame-based objective ignores that we are really interested in word/sentence discrimination, not frame discrimination
- **Idea:** Switch to a sequence criterion as an objective function

Frame based NN parameter Gradient update:

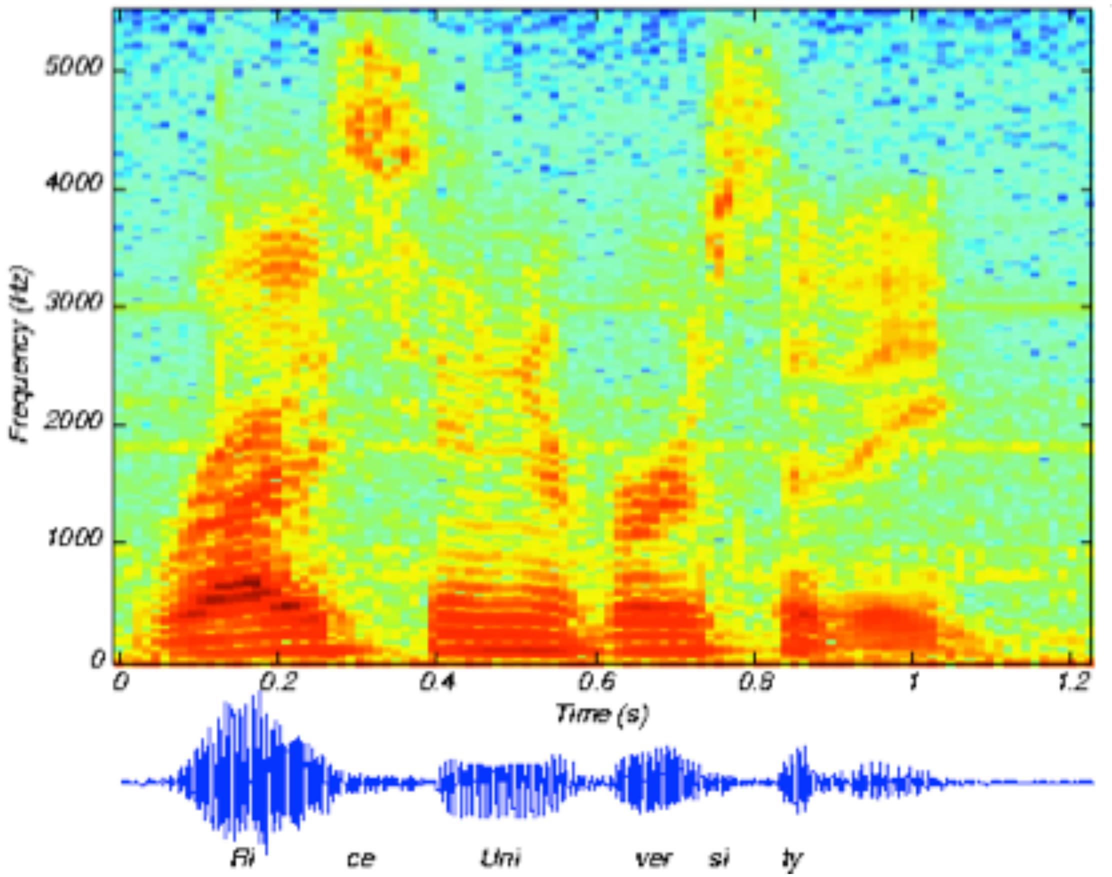
$$\frac{\partial L}{\partial a_t} \sim \sum_{i=1}^{N} p(c_{it}|x_t) - y_{it}^{ref}$$

Sequence based NN parameter Gradient update:

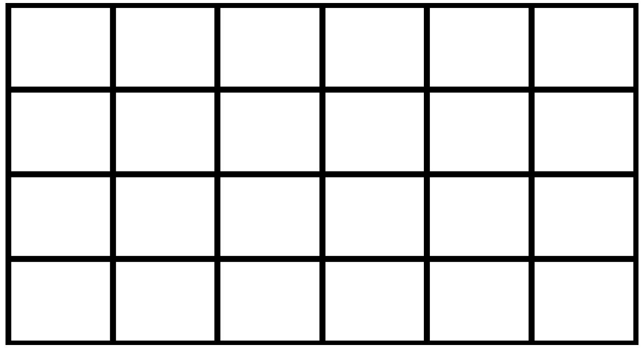
$$\frac{\partial L}{\partial a_t} \sim \sum_{i=1}^{N} p(c_{it}|x_1 \dots x_t \dots x_T) - p(c_{it}|x_1 \dots x_t \dots x_T, w_1 \dots w_L)$$

IBM Enhancement #2: Convolutional Neural Networks (CNNs)

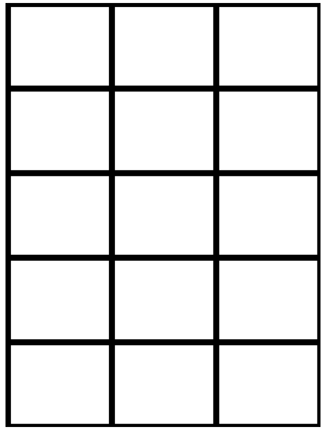
- Spectrographic representation clearly demonstrates speech is locally correlated in time and frequency.
- Idea: Try to construct a neural network that is designed to specifically capture these sorts of local correlations

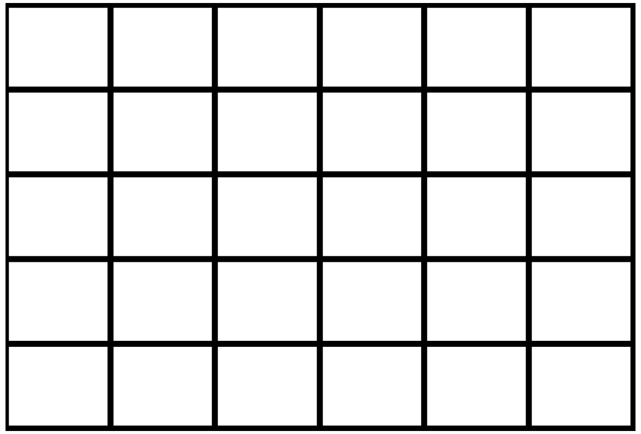


Review of DNN Weight Multiplication

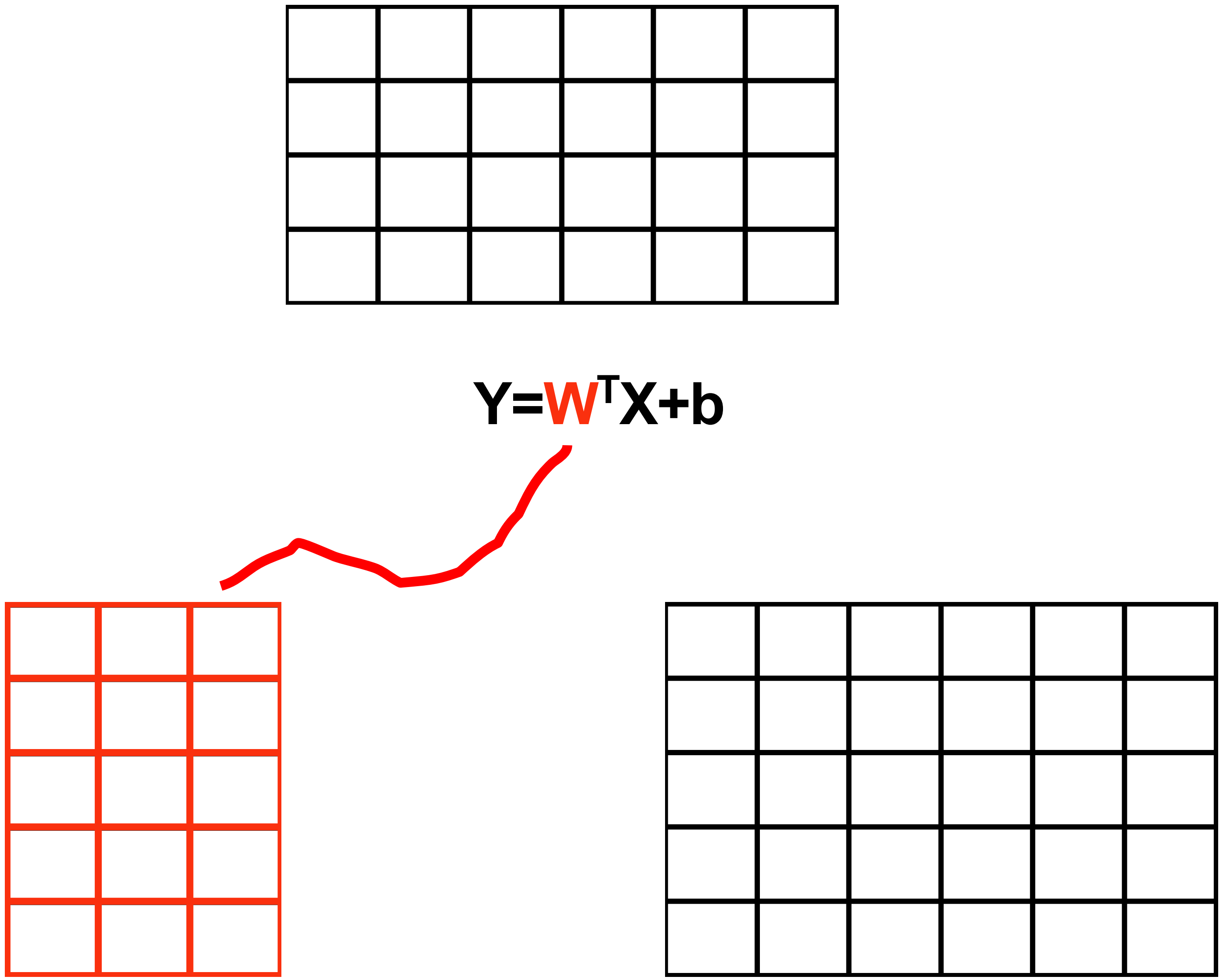


$$Y = W^T X + b$$



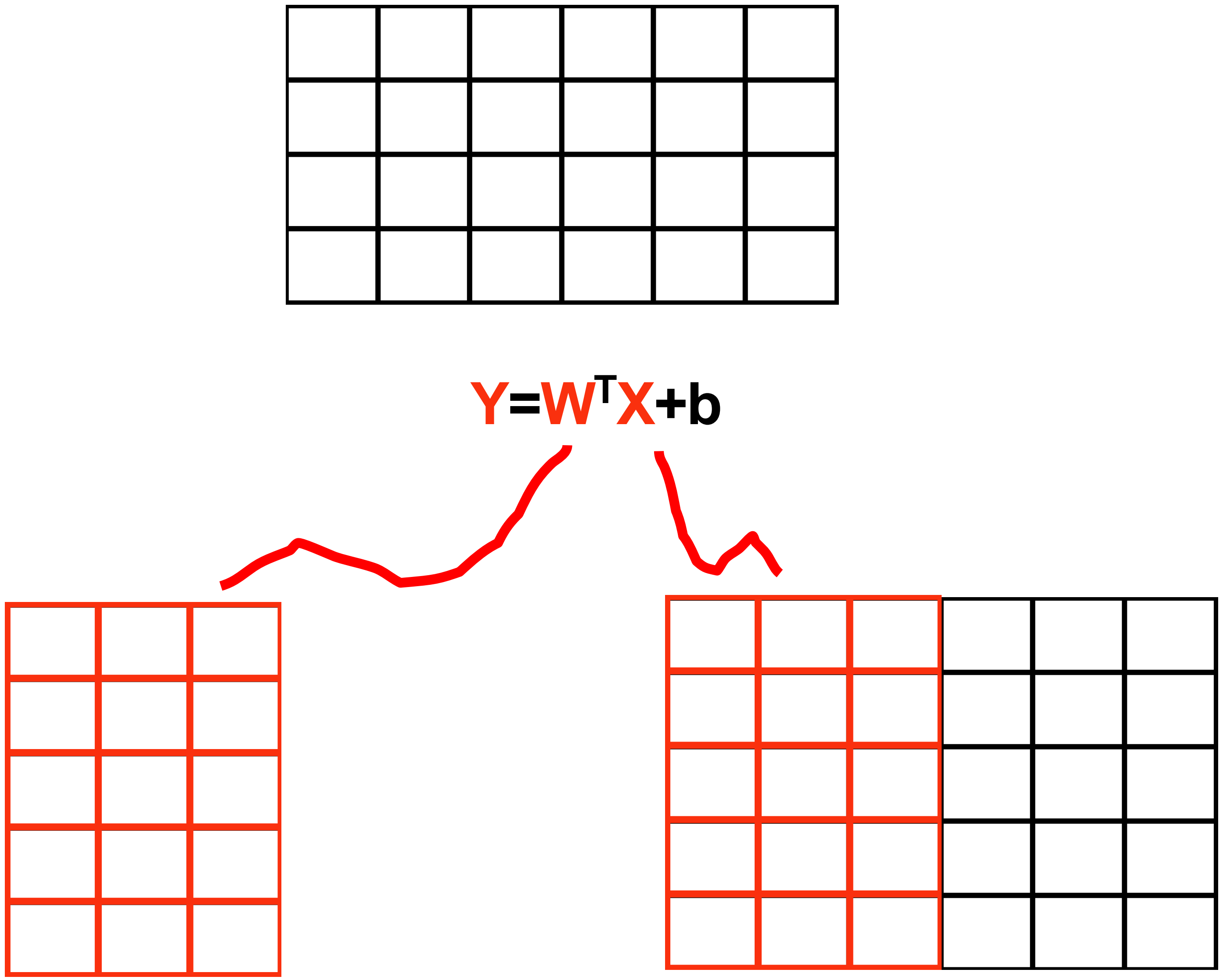


Review of DNN Weight Multiplication

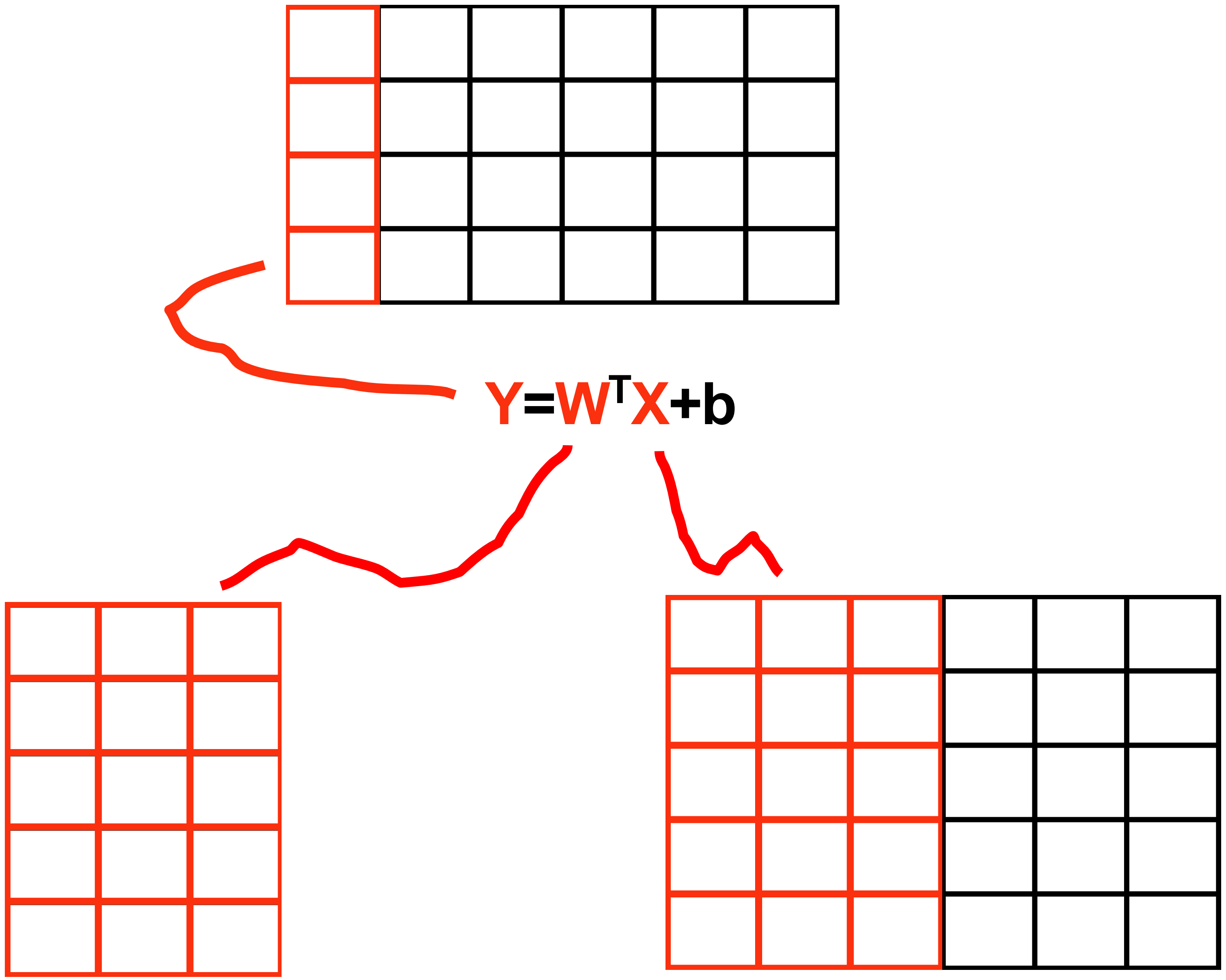


$$Y = W^T X + b$$

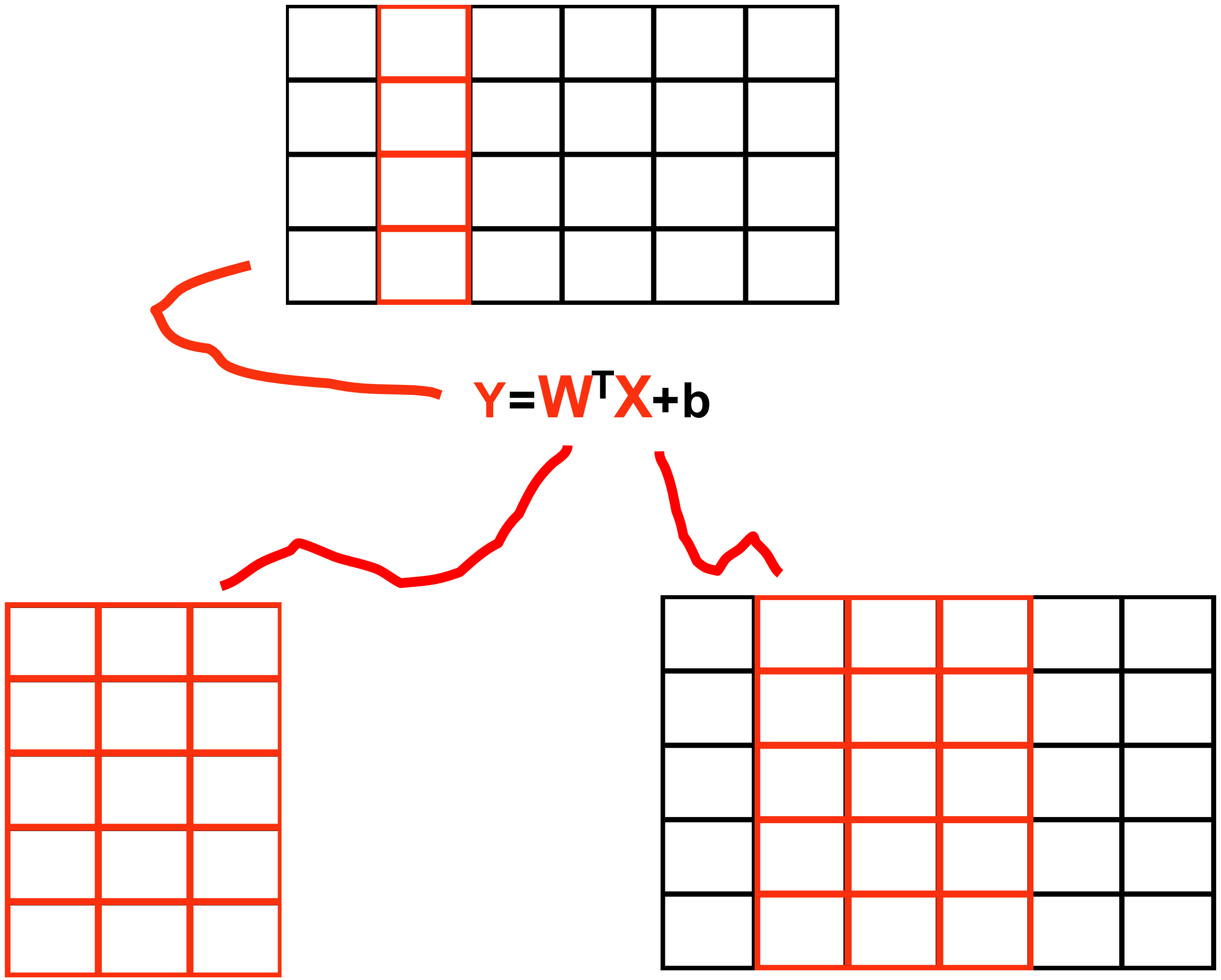
Review of DNN Weight Multiplication



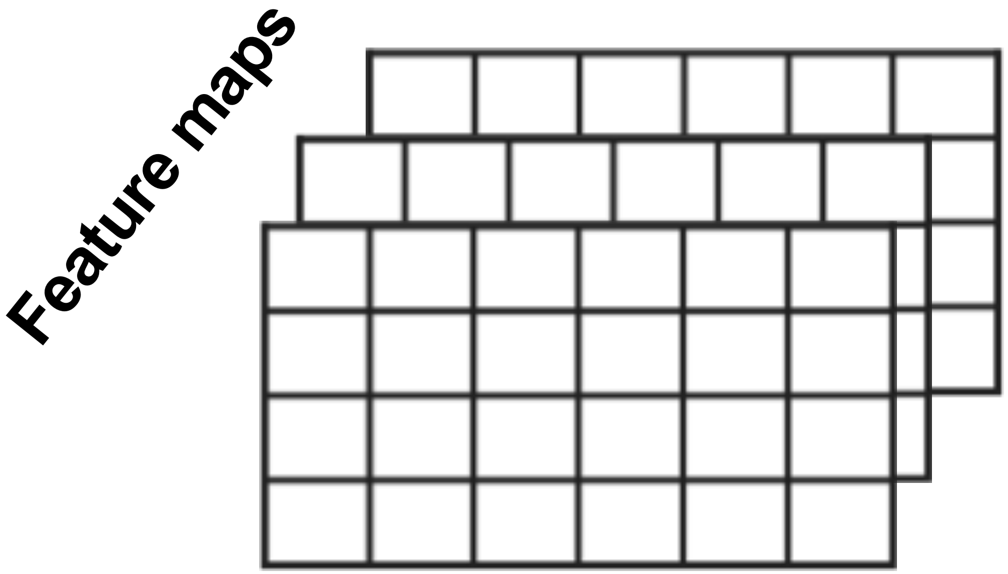
Review of DNN Weight Multiplication



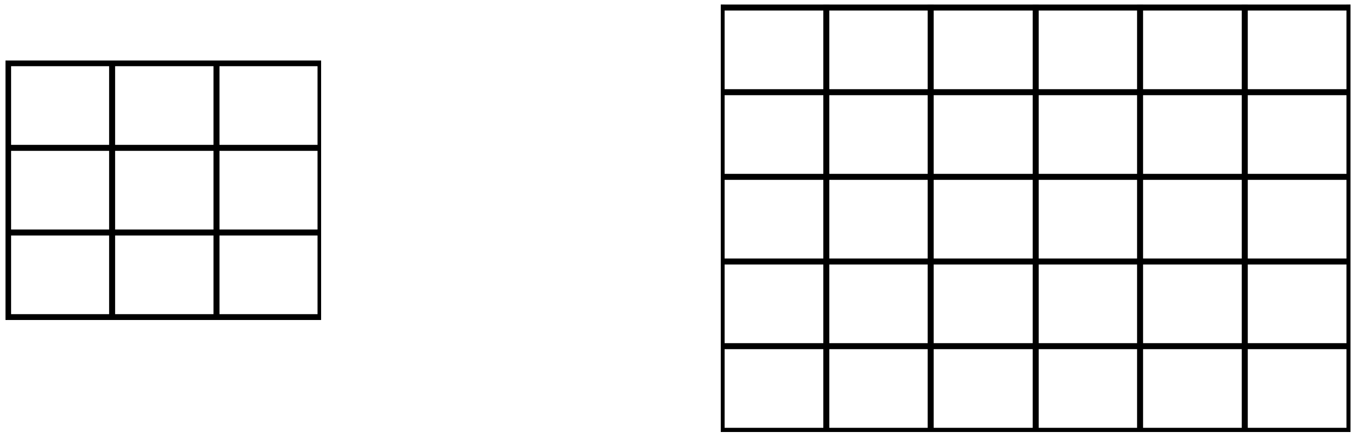
Review of DNN Weight Multiplication



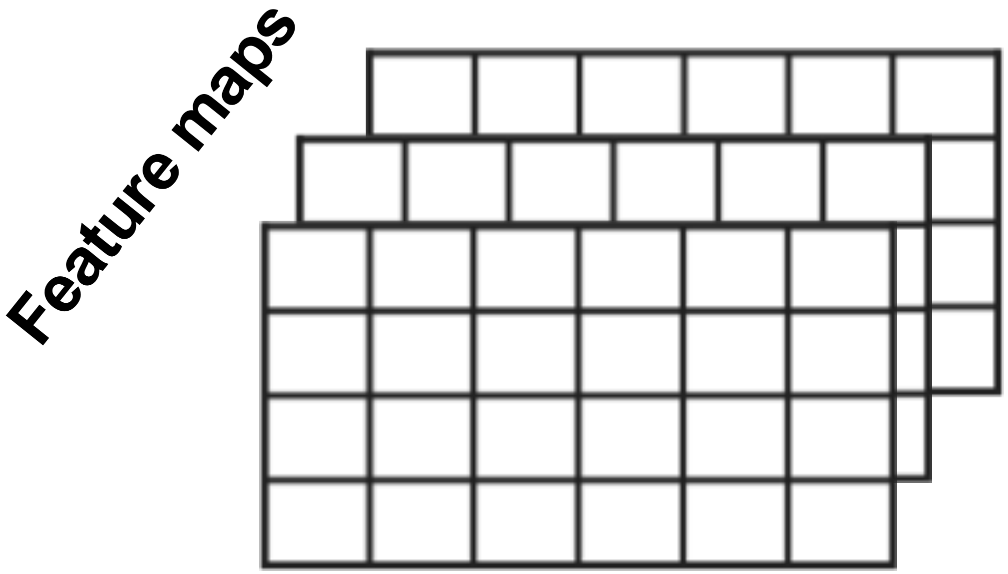
CNN Weight Multiplication



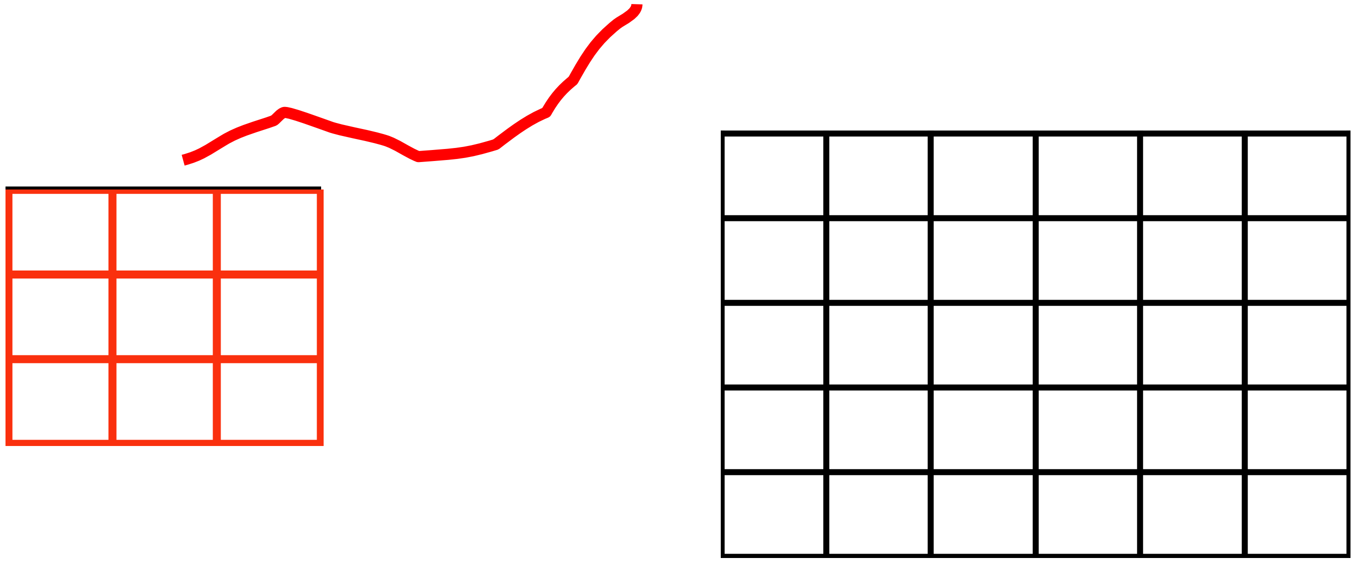
$$Y=W^TX+b$$



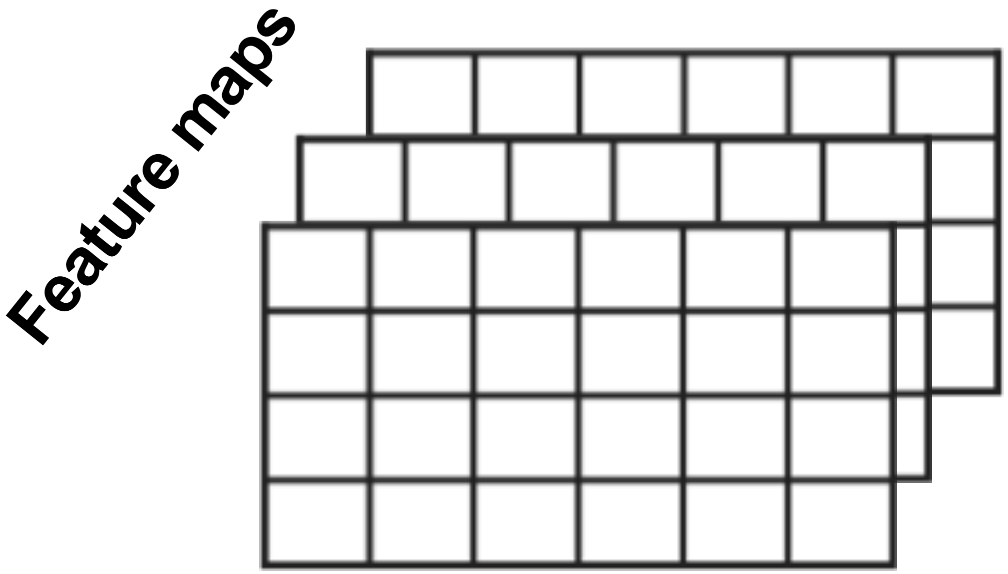
CNN Weight Multiplication



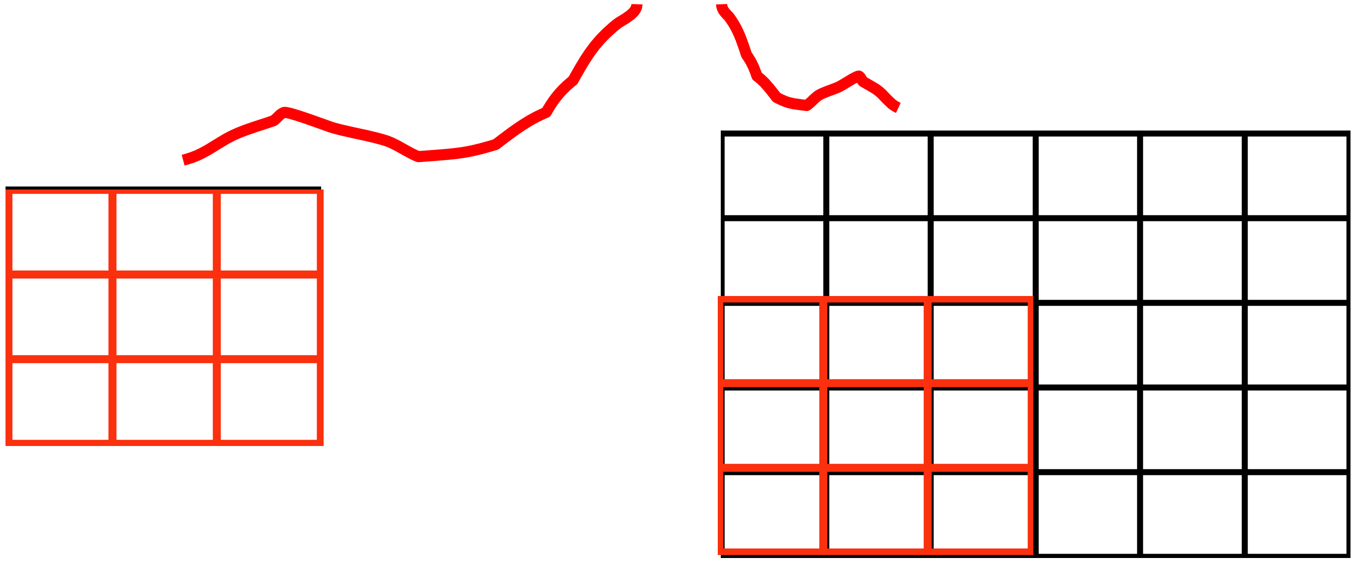
$$Y = W^T X + b$$



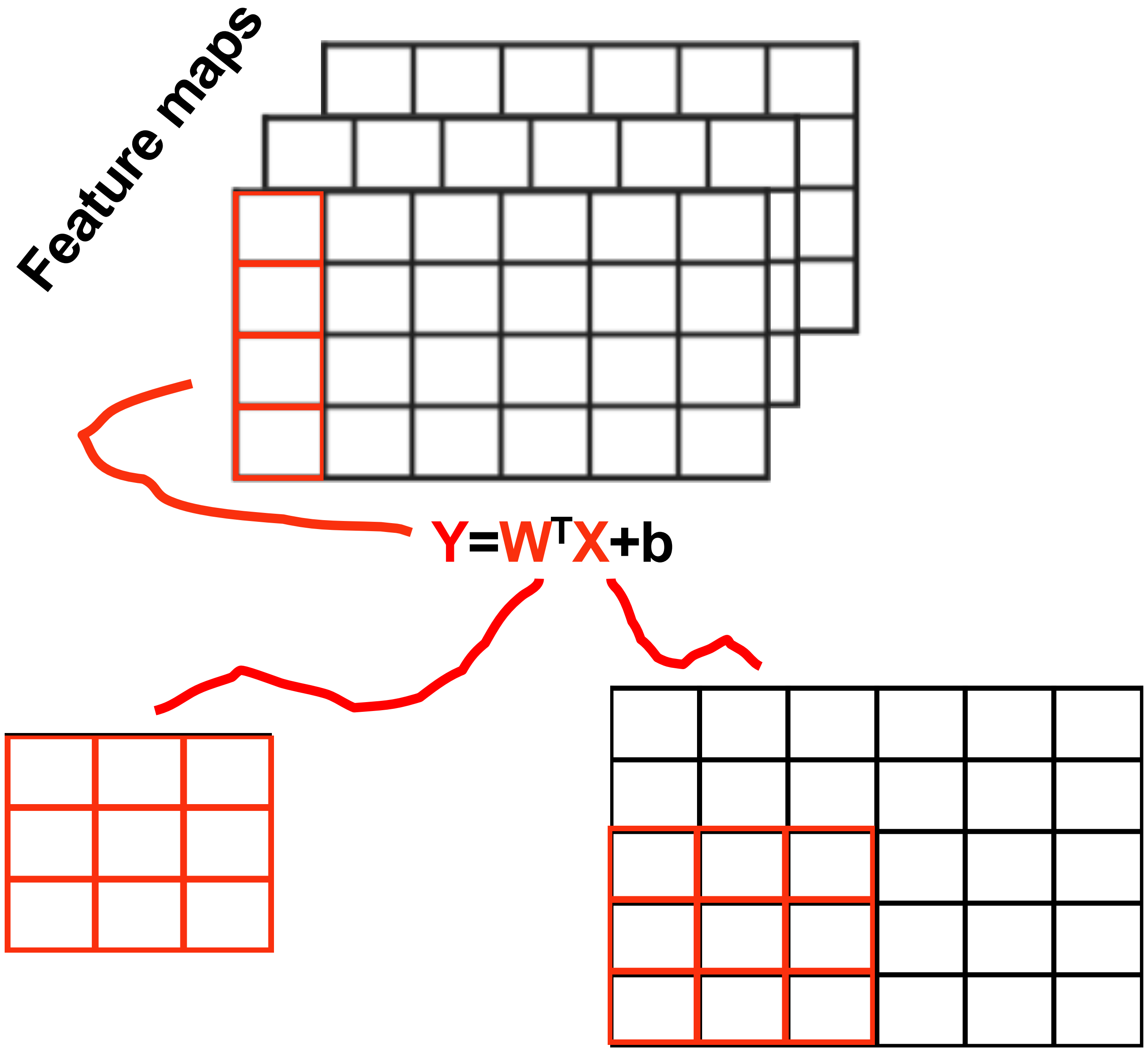
CNN Weight Multiplication



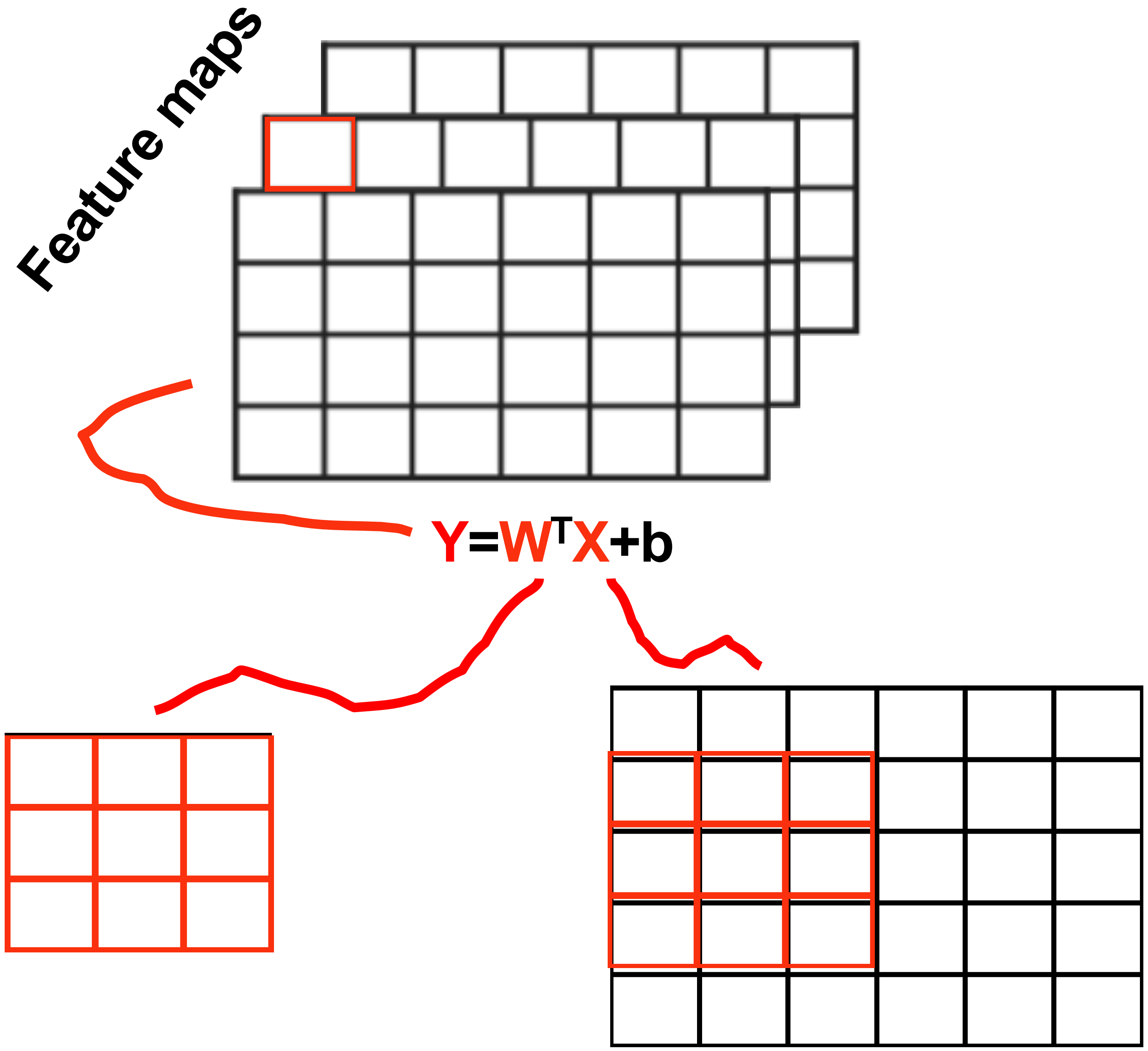
$$Y = W^T X + b$$



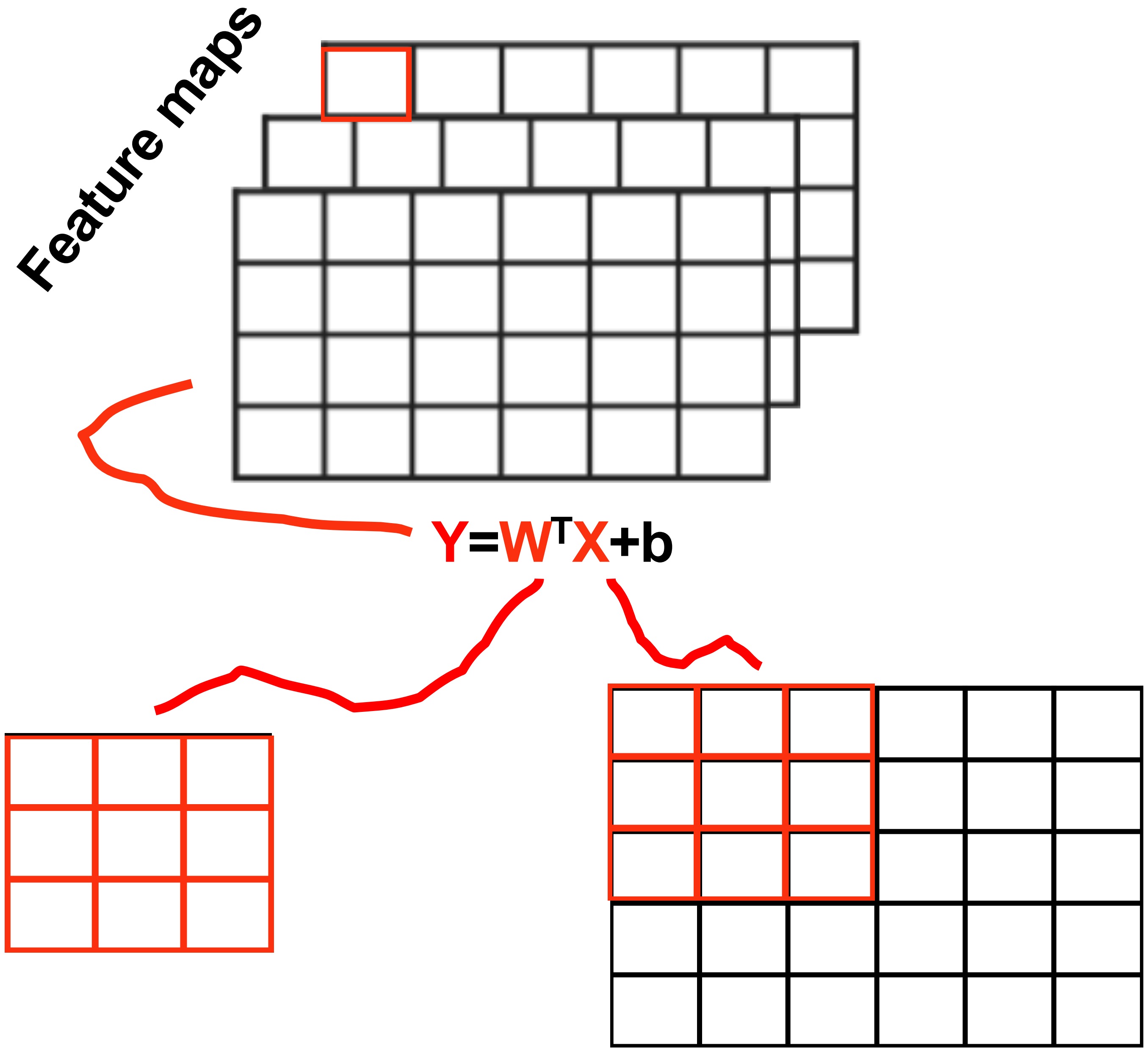
CNN Weight Multiplication



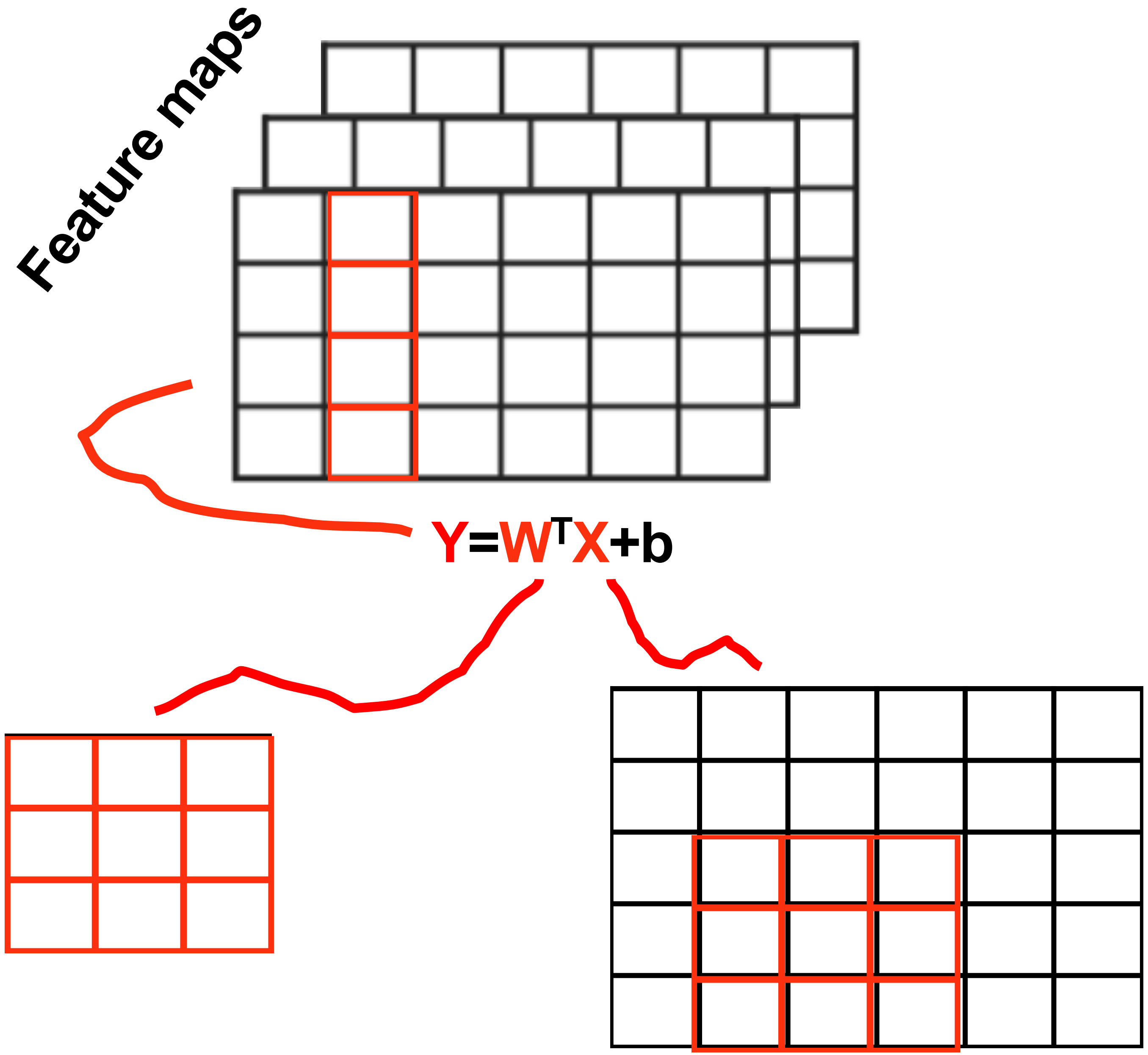
CNN Weight Multiplication



CNN Weight Multiplication



CNN Weight Multiplication



And what does all this do for us?

Results on SWITCHBOARD corpus…

System	300 Hours Training Data		2000 Hours Training Data	
	Cross-Entropy	Sequence	Cross-Entropy	Sequence
GMM	14.5			
DNN	14.1	12.5		
CNN	13.2	11.8	12.6	**10.4**

Remember this!

Recent Enhancements: Unfolded Recurrent NNs

- Feed-forward NNs have no memory over time: time traditionally captured with an HMM.
- A NN model for time varying behavior is an RNN:

$$\mathbf{y}_t = p(\boldsymbol{c}|\boldsymbol{x}_t) = \mathrm{softmax}(\mathbf{W}_{hy}\tilde{\mathbf{h}}_t)$$
$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\,\mathbf{x}_t + \mathbf{W}_{hh}\tilde{\mathbf{h}}_{t-1})$$

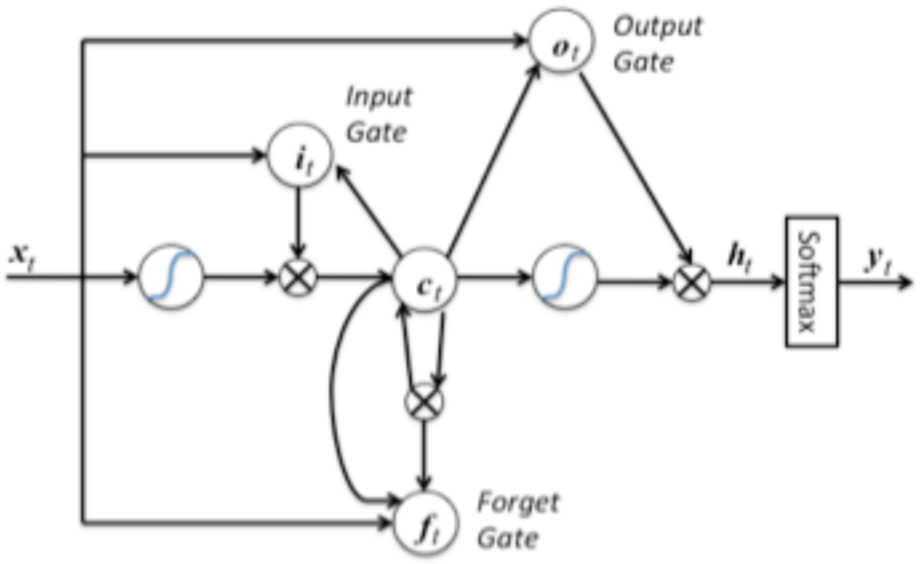
Above is iterated from 1 to T (number of input vectors)

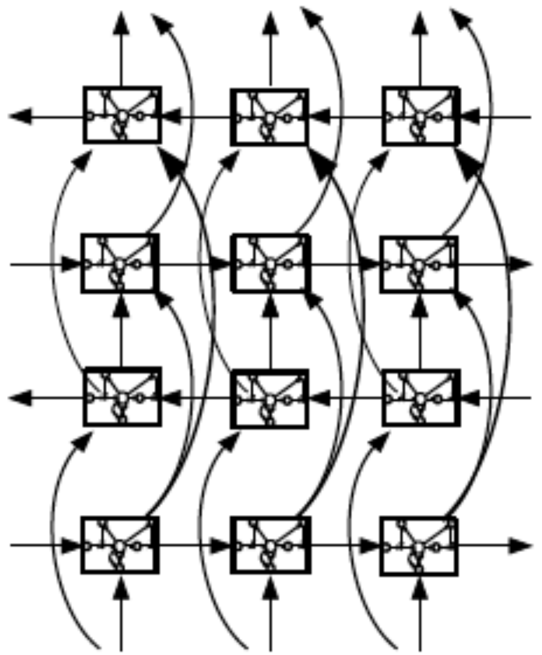
- For a simple RNN architecture as described above, it is possible to perform frame unrolling:

$$\begin{aligned}
\mathbf{h}_t &= \sigma(\mathbf{W}_{xh}\,\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}) \\
&= \sigma(\mathbf{W}_{xh}\,\mathbf{x}_t + \mathbf{W}_{hh}\,\sigma(\mathbf{W}_{xh}\,\mathbf{x}_{t-1} + \mathbf{W}_{hh}\mathbf{h}_{t-2})\,) \\
&\dots \\
&= \sigma(\mathbf{W}_{xh}\,\mathbf{x}_t + \mathbf{W}_{hh}\,\sigma\big(\dots + \mathbf{W}_{hh}\,\sigma(\mathbf{W}_{xh}\,\mathbf{x}_1 + \mathbf{W}_{hh}\mathbf{h}_0)\big))
\end{aligned}$$

- Effectively converts recursive network to a feed-forward network
- Permits leveraging of pre-existing training infrastructure

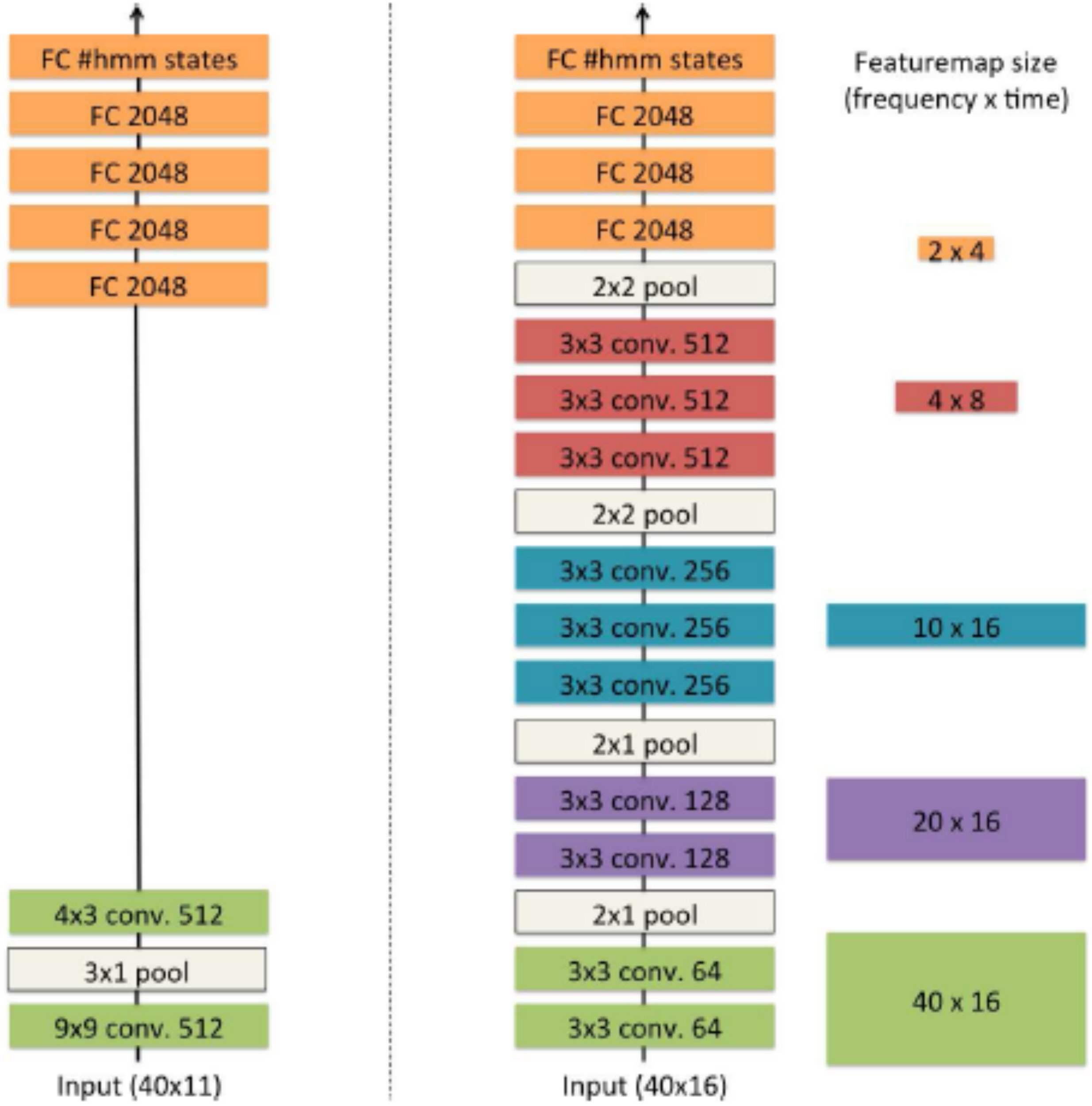
Recent Enhancements: LSTM Networks





- In the RNN, the gradients decay exponentially in time making it hard to capture long term dependencies
- The LSTM ("Long-Short-Term-Memory") network adds trainable gates that allow information to be stored for long periods of time.
- Best systems employ bidirectional LSTMs - 4/5 layers now typical

Recent Enhancements: VGG Networks



Language Modeling Improvements

All previous results used a 4-gram LM with 4M ngrams and a vocabulary of 30.5K words

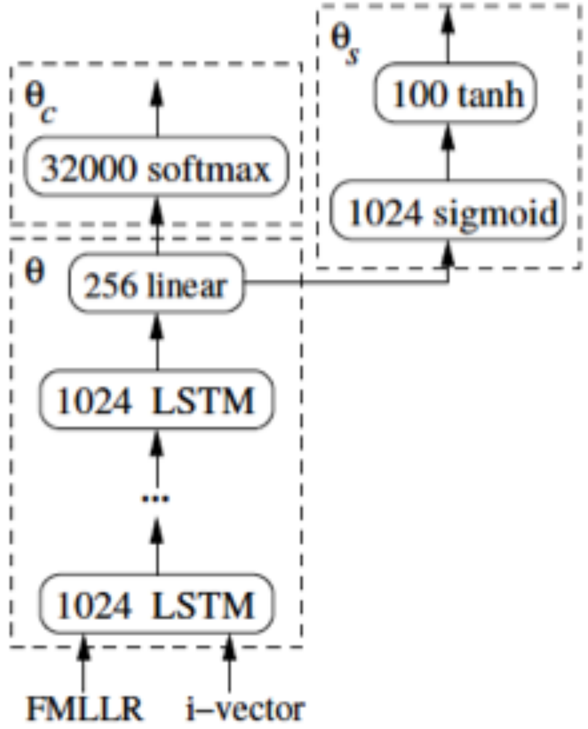
Enhancement: Combine Three LMs with a vocabulary of 85K words

- 4-gram with 36M n-grams
- Feed-forward neural network LM
- MaxEnt class-based LM called ("Model M")
  - $p(w_j \mid w_{j-1}\, w_{j-2}) = p(w_j \mid c_j\, w_{j-1}\, w_{j-2}) \times p(c_j \mid c_{j-1}\, c_{j-2},\, w_{j-1}\, w_{j-2})$

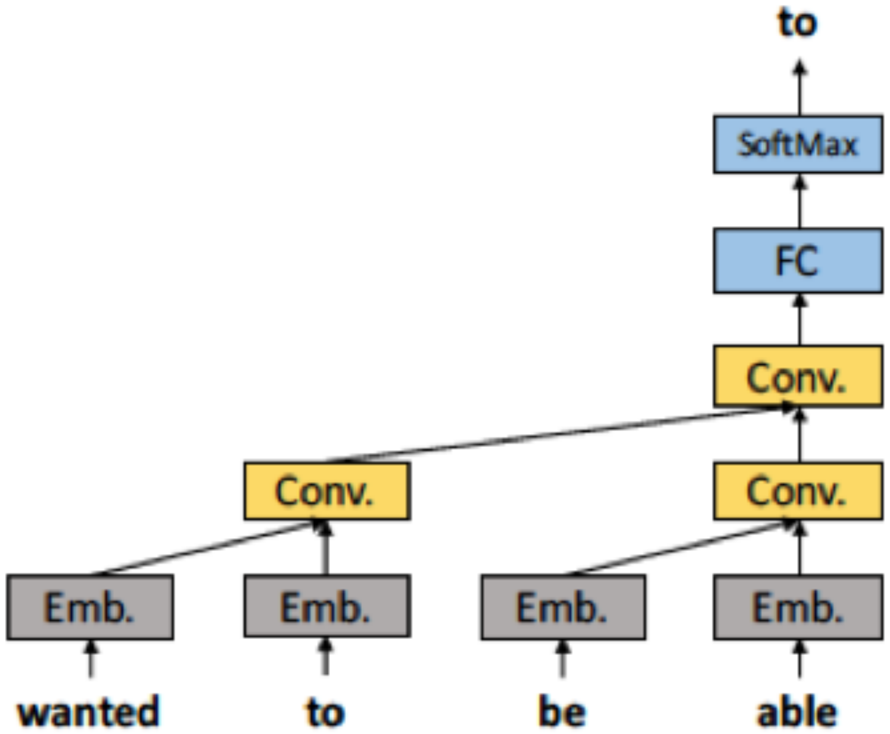
2017 Progress in Speech Recognition

Advanced Deep Learning

Adversarial Learning

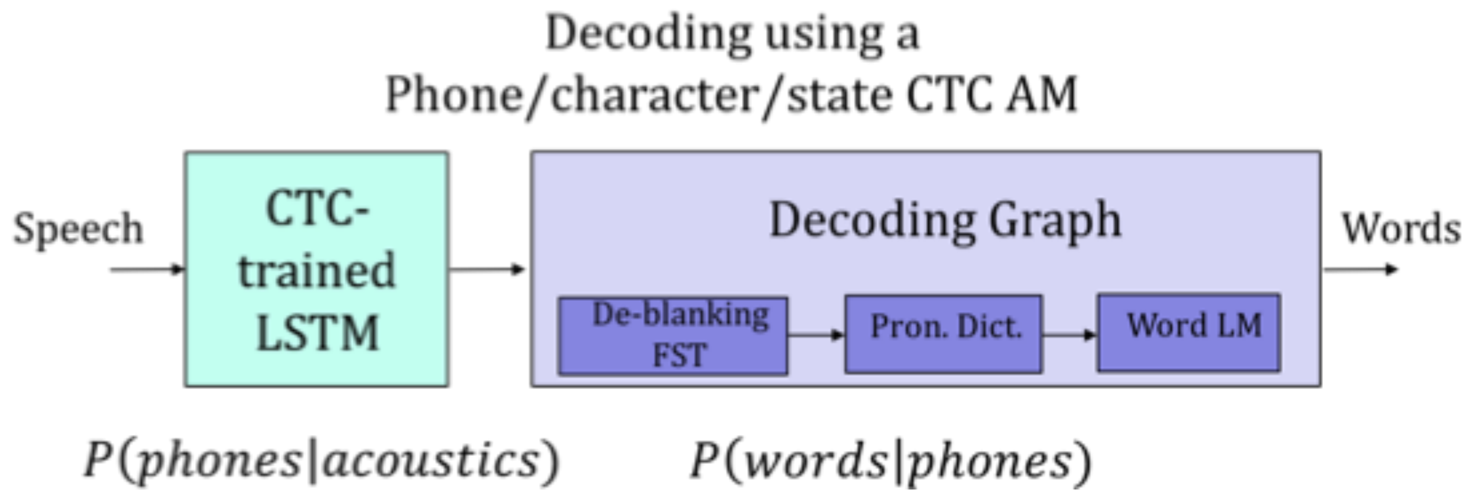


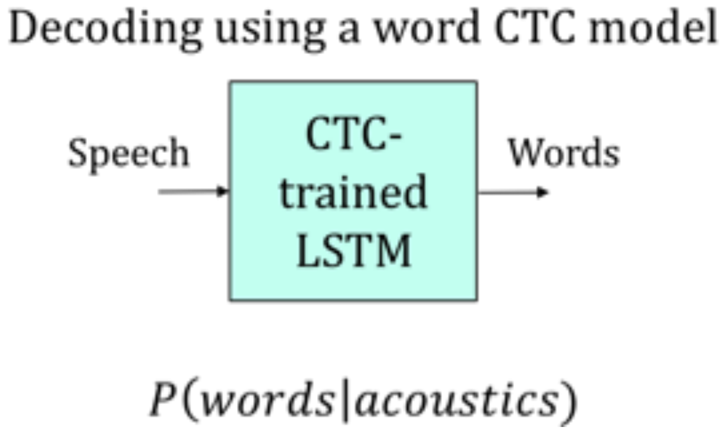
Convolution-Inspired NN LMs



Direct Acoustics-to-Word Automatic Speech Recognition

- New direction eliminating all modeling assumptions relying purely on Deep Learning
- Scalable: Formerly large complex speech engine reduced to single NN architecture





Conventional sub-word based ASR uses phones, dictionary, and language model during decoding → **not end-to-end.**

Direct acoustics-to-word ASR uses no dictionary, language model, or decoder → **True end-to-end**

Impact of Deep Learning

Model	Word Error Rate	Described in IBM Publication
1. CNN	**10.4**	[TS2013b]
2. RNN	9.9	[Saon2014,Saon2015]
3. VGG	9.4	[Sercu,2016]
4. RNN+VGG+LSTM	8.6	[Saon,2016]
5. (4) +More Ngrams+ModelM	7.0	[Chen2009, Saon2016]
6. (4) +More Ngrams+ModelM +NNLM	6.6	[Mangu2007, Chen2009, Saon2016]
7. Adversarial Learning + Resnet + LSTM	6.7	[Saon2017]
8. (7) + (6) + LSTM LMs + Wavenet LM	5.5	[Saon2017,Kurata2017]

How Well do Humans Do?

	WER SWB
Transcriber 1 raw	6.1
Transcriber 1 QC	5.6
Transcriber 2 raw	5.3
Transcriber 2 QC	**5.1**
Transcriber 3 raw	5.7
Transcriber 3 QC	5.2

So Are We Done?

- Results strongly tailored to this individual corpus
  - Trained on 2000 hours of strongly targeted data both for LM and for AM
  - Relatively high quality (if telephony based) speech
  - Relatively accent free
  - Nature of conversations somewhat stilted

So Are We Done?

- What happens when speech systems have to deal with variations in
  - –Accent
  - –Noise
  - –Speaking Style
  - –Domain Switching

Corpus	WER Relative Increase
LDC-Switchboard	x1.0
LDC-Broadcast News	x1.4
LDC-Call Home	x2.0
Customer-Agent	x2.1
Emotional Speech	x2.8
Noisy Speech	x3.4
Accented Speech	x3.4

- We know that task specific data would help a lot, but do we really have to put in this level of effort for each language for each domain?
- And what are human abilities in terms of being able to cope with these variations?

Rest of Talk

- Look at following areas
  - Noise
  - Speaking Style
  - Accent
  - Domain Robustness
  - Language Learning Capabilities

- Review state of human and machine performance in these areas

- **Goal:** Try to make the case that we have a long way to go in speech recognition – so let’s keep doing research!

Perception of Noisy Speech

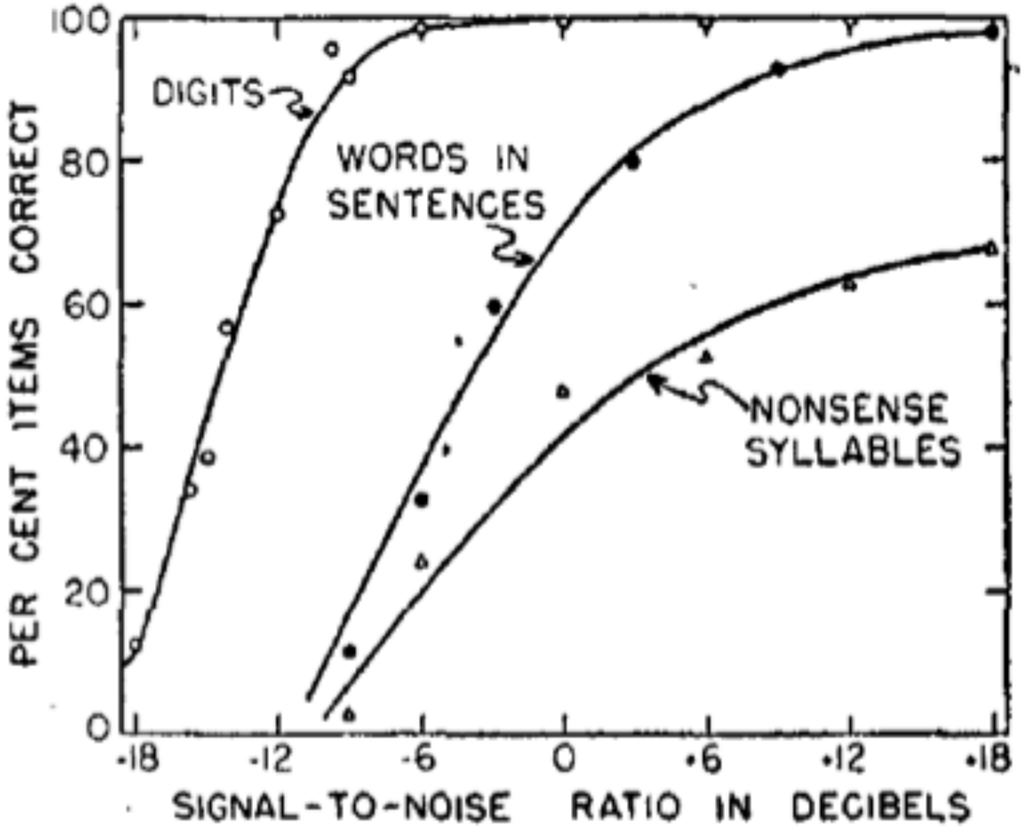
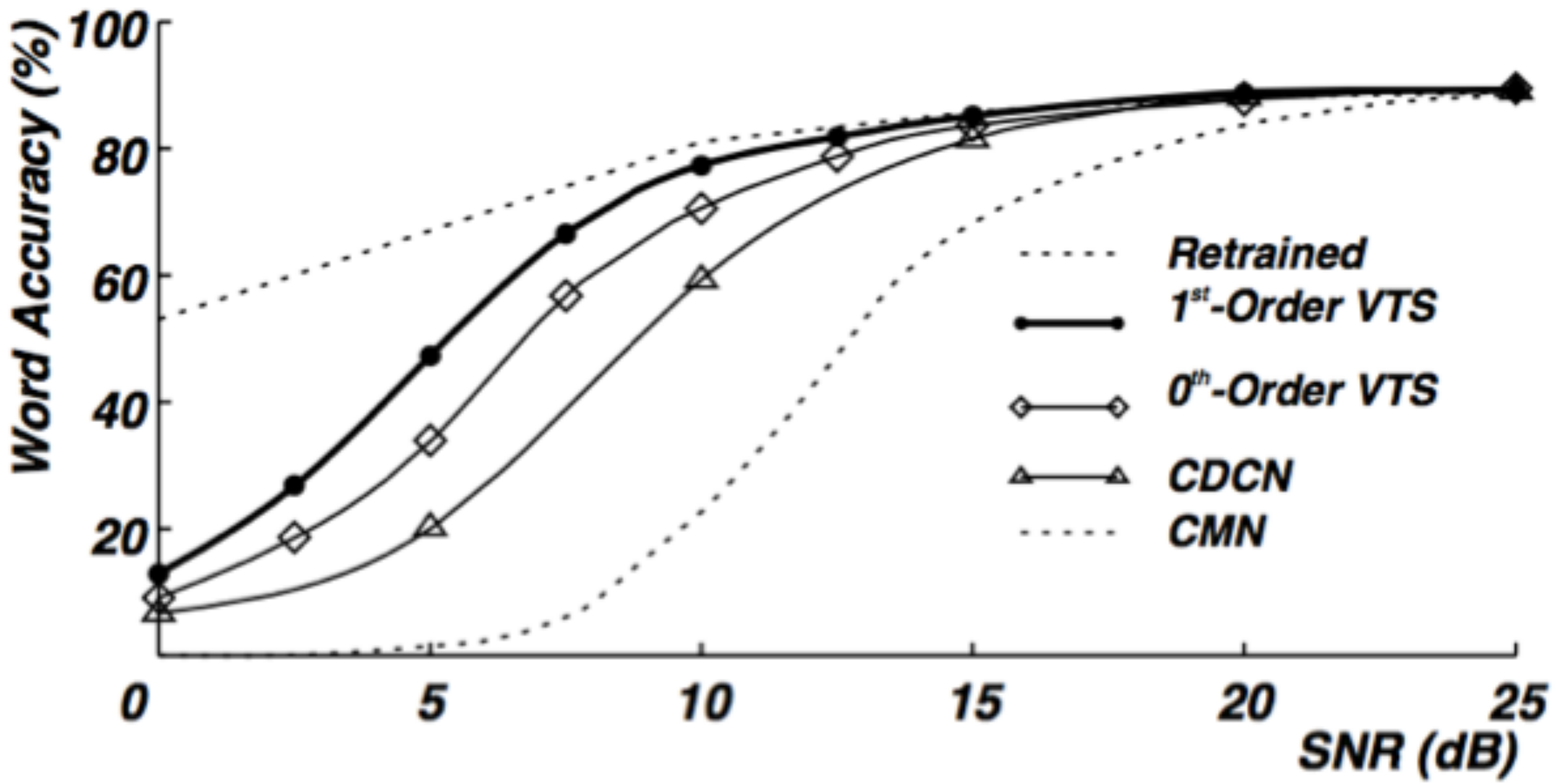


Fig. 1. Relative intelligibility of different test materials

- Intelligibility depends on the predictability of the materials
- Starts decreasing at 10 dB SNR; 0% by -7 db SNR

Recognition of Noisy Speech

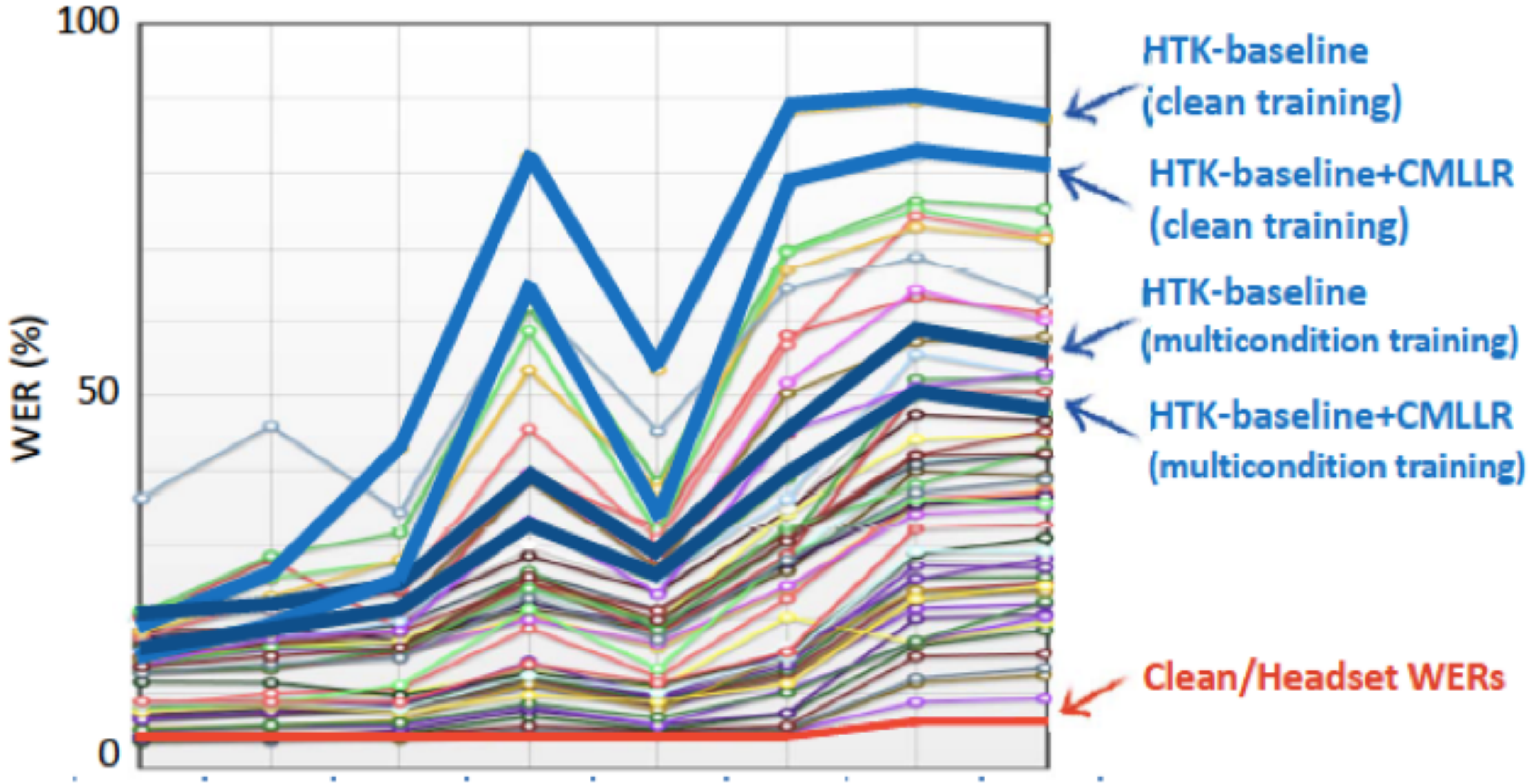
Results on WSJ-84, 5000 word vocabulary test set.



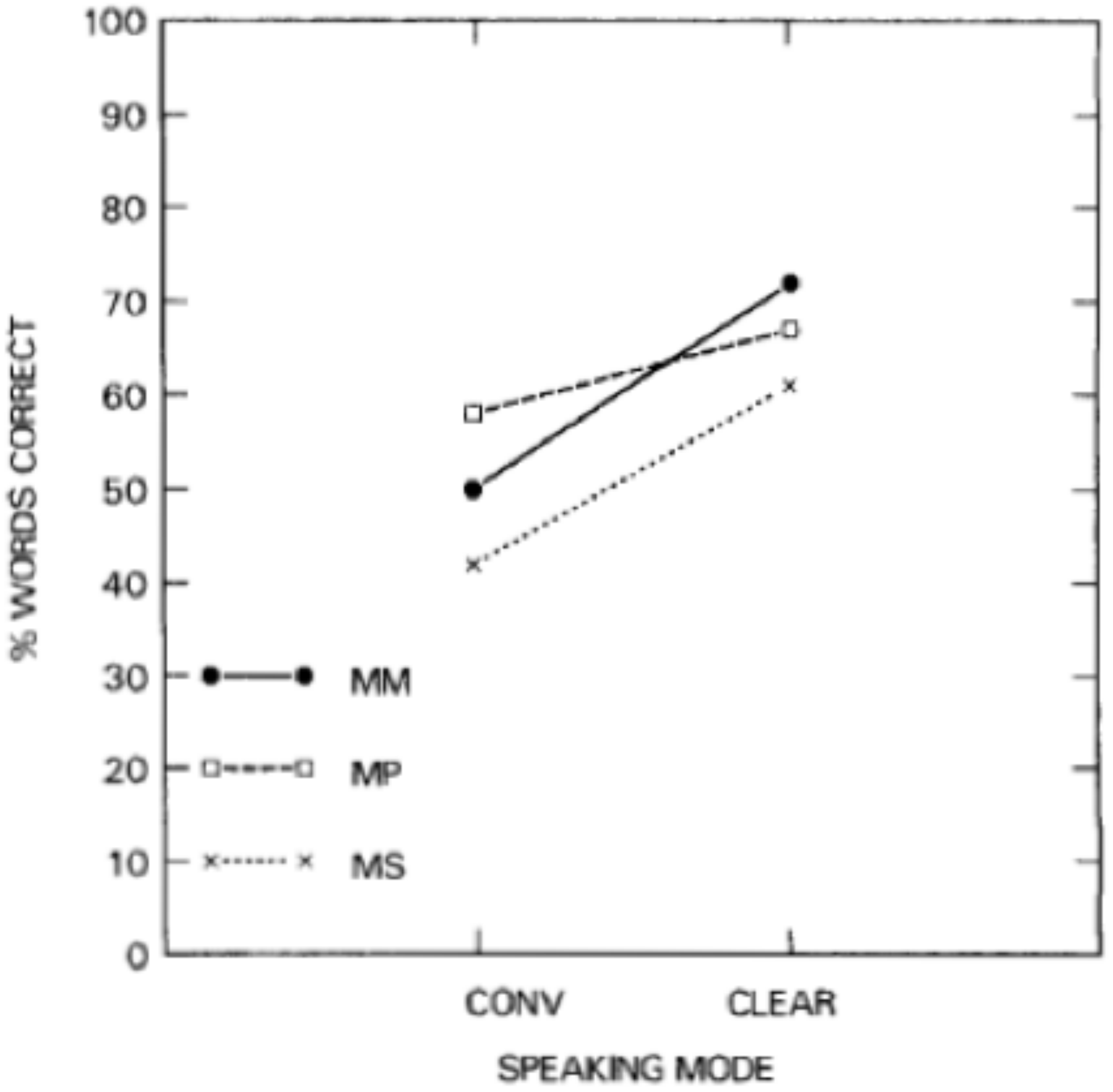
- Typical feature-based methods start losing accuracy at 10 dB; reaches chance by 0 dB
- **Multi-style training** maintains robustness over larger SNR ranges.

More Recent Results in Noisy Speech Recognition

- Deep Learning improves speech recognition performance but no special advantage seen for noisy/reverberant speech.
- Recent Noisy/Reverberant Speech Challenges (REVERB, CHIME. ASPIRE) achieve best results by combining a variety of techniques
  - –Multimike processing, Multistyle training, Multiple systems



Effect of Speaking Style on Speech Intelligibility



Informal speech is harder to recognize than clearly articulated speech

Effect of Speaking Style on Speech Intelligibility

TABLE 1. Speaking rates (words/min) for all 3 speakers.

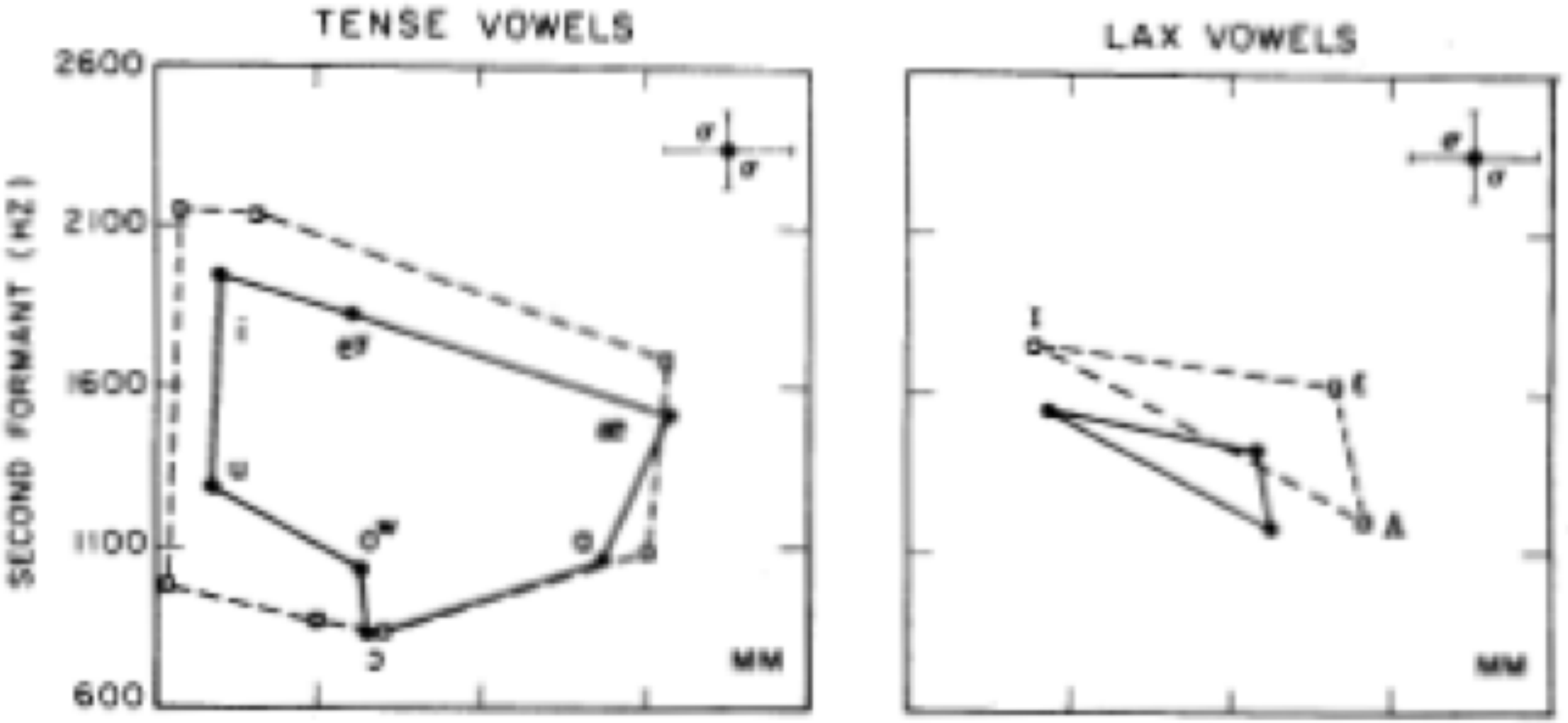
Speaker	Conversational speech	Clear speech
MM	205 (3.9)	101 (1.9)
MP	160 (3.0)	91 (1.7)
MS	199 (3.8)	101 (1.9)

Effect of Speaking Style on Speech Intelligibility

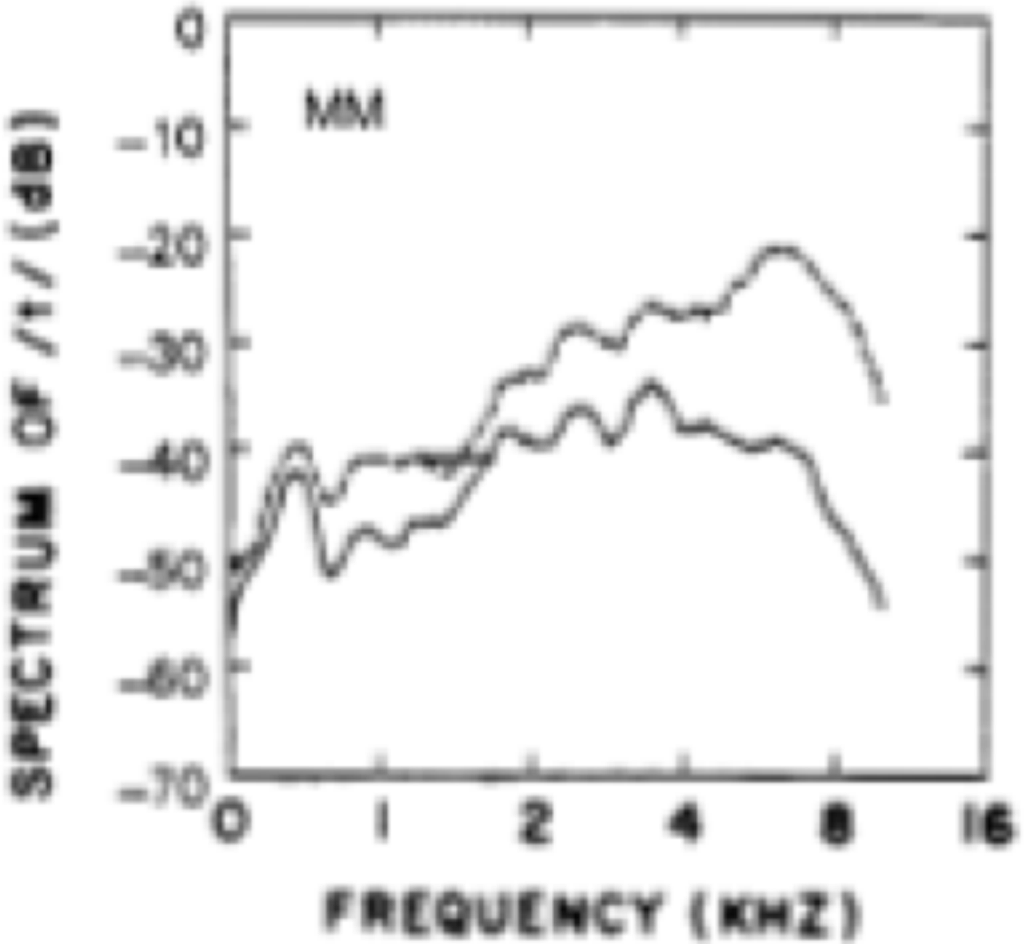
TABLE 2. Phonological phenomena occurrences.

Phonological type	MM			
	Conv		*Clear*	
	Con	*Fun*	*Con*	*Fun*
VM	28	88	18	47
BE	39	9	8	9
DG	6	1	0	1
AF	4	5	2	1
SI	1	0	38	0
MSD	9	13	2	6

Effect of Speaking Style on Speech Intelligibility

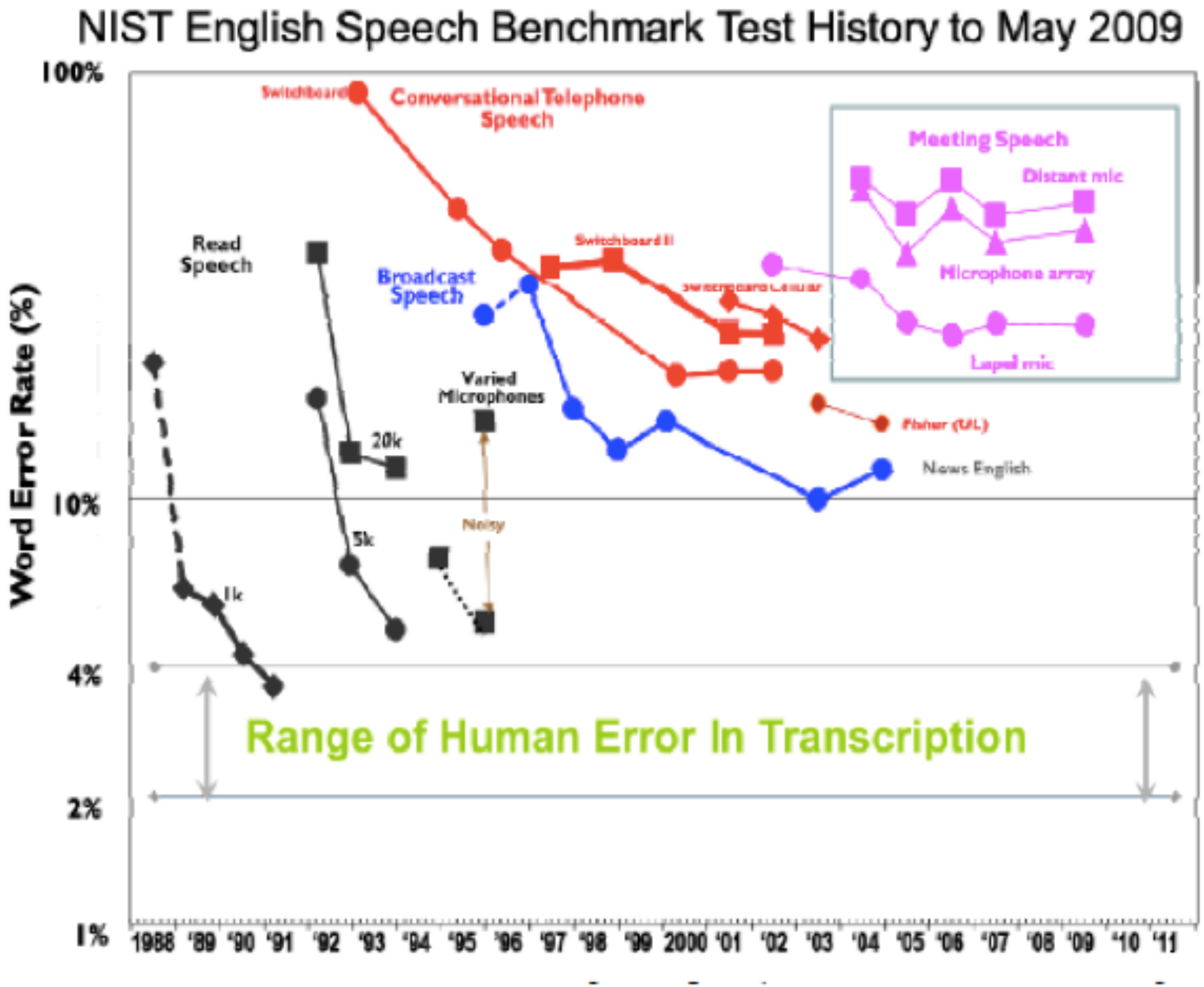


Effect of Speaking Style on Speech Intelligibility



- **Significant acoustic changes when you speak conversationally**
- **Impacts both human and machine recognition performance**

Effect of Speaking Style on Speech Recognition Performance



Effect of Speaking Style on Speech Recognition Performance

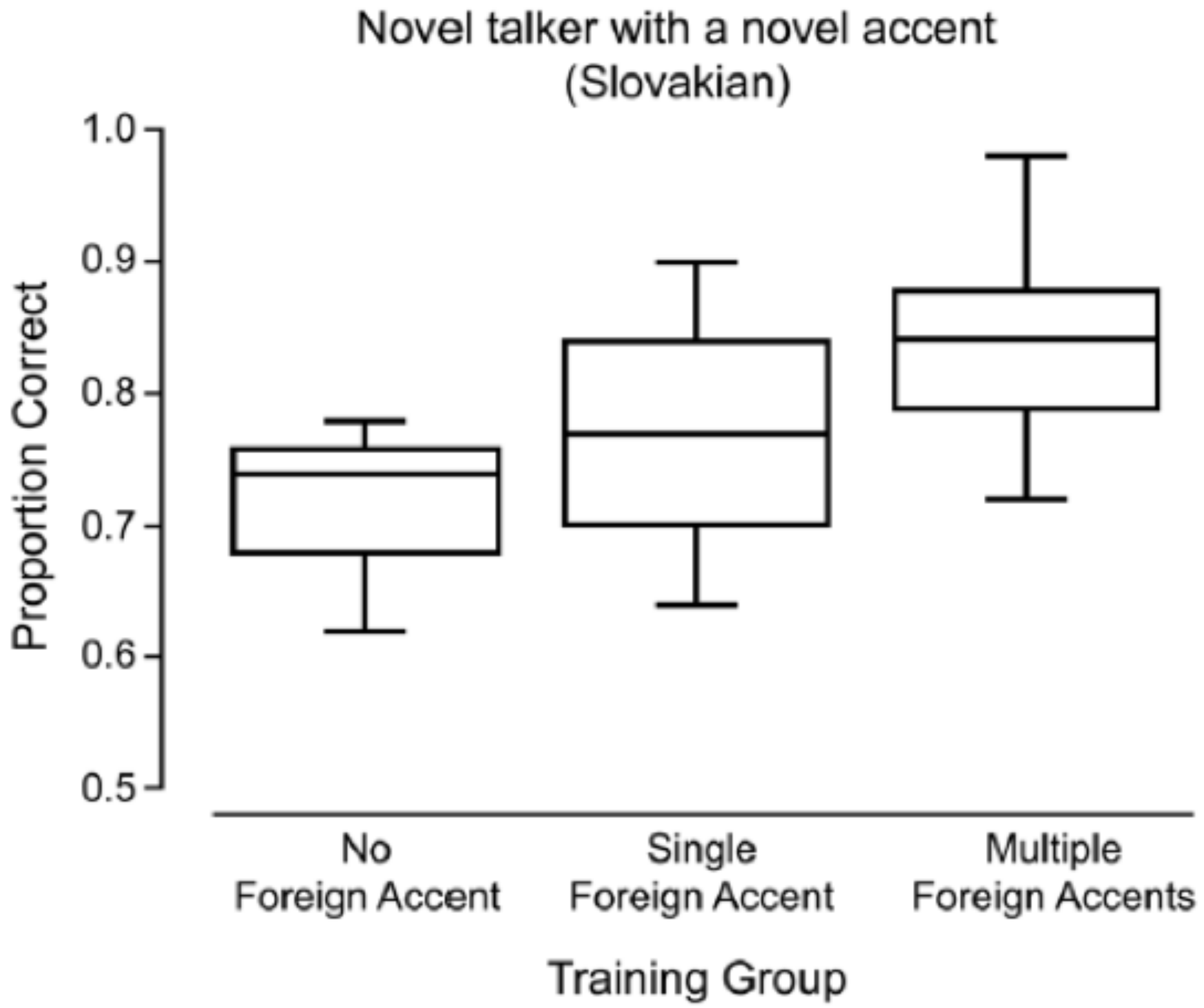
- Speaking style clearly affects speech recognition performance
- In order of difficulty: read speech, formal speech, person-to-person speech, many-person (meeting room) speech

Table 4. Word Error Rates (%) on AMI – IHM

System	AMI
BMMI GMM-HMM (LDA+STC, SAT)	29.6
DNN – Sigmoid	26.6
DNN – ReLU	25.5
DNN – Maxout	26.3
CNN – Sigmoid	25.6
CNN – ReLU	24.9
CNN – Maxout	25.0

- Meeting speech clearly difficult, even with recent DL advances
- Unlike SWB; no human benchmarks exist

Perception of Accented Speech



Intelligibility of Accented Speech Depends on Accent Exposure

Recognition of Accented Speech

Train on lots of data and Leverage Grapheme Knowledge

Model	Indian Accent (US)
EnUs CTC-P	15.2
EnUs Adapted Multi-Dialect CTC-P	11.2
Multi-Dialect HCTC-G	8.5
EnUs Adapted Multi-Dialect HCTC-G	8.7

Table 4: WER (%) performance of various models on an Indian accented US queries test set.

Need lots of data to train (have ~3000 hours per accent here (!))
Grapheme effects may be unique to English

Domain Robustness

- Systems now are trained on **thousands of hours of speech** and **billions of words of text.** Humans recognize a large variety of contexts by the time they are 20.
- How much speech does a person typically hear by the time they are 20?
  – Yahoo answers: A Human usually hears about 50000 words a day and you use about 25000 a day depending on how talkative you are
- By 20 have heard 365, 000,000 words (!) give or take a factor of 4 ☺. At 2.5 words a second, this is about 25,000 hours of speech.

Domain Robustness

- .How many words does a person typically read by the time they are 20?
  - https://techcrunch.com/2009/12/09/study-americans-consume-34-gigabytes-of-information-per-day/ "Americans consume 100,000 words per day on average. That includes all words read, all words heard, etc."
  - ~365,000,000 words in 20 years (taking half of above)

Not unreasonable to be training systems on at least 10000 hours of speech….but implies 400M words of exposure may be enough to understand all domains…so **why do our language models need billions of words?**

Value of Domain Adaptation to Speech Recognition

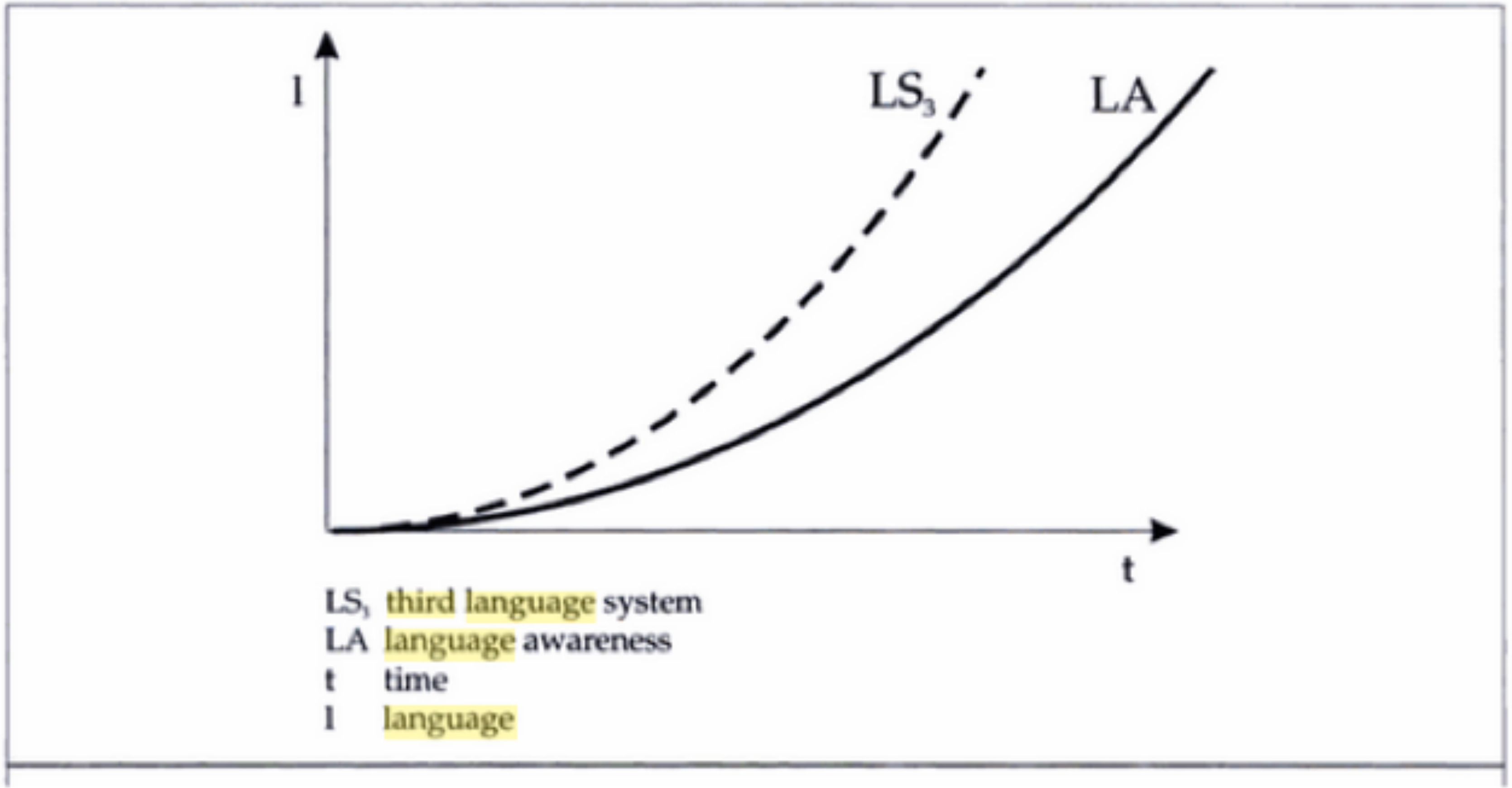
	Healthcare	Insurance	Hospitality
System	**.5 hrs**	**3 hrs**	**140 hrs**
Baseline	31.0	24.8	13.8
+ AM-Unsupervised	22.0	23.4	
+ AM-Supervised	16.5	21.9	10.0
+ LM	12.8	19.5	10.8
+ AM-Supervised	9.6	18.9	9.4

- Domain Adaptation Helps a Lot, particularly LM adaptation
- Not that much data is needed per domain on top of a good base
- Unclear how many domains can be simply interpolated together
  - Do we need more work on dynamic adaptation method?
    - Have been attempts in the past, but on much older technology bases.

Learning New Languages

"If, for the sake of argument, we consider fluency to be the same as being an "expert" in speaking a language, then a learner may well invest 10,000 hours in language studies to attain fluency."

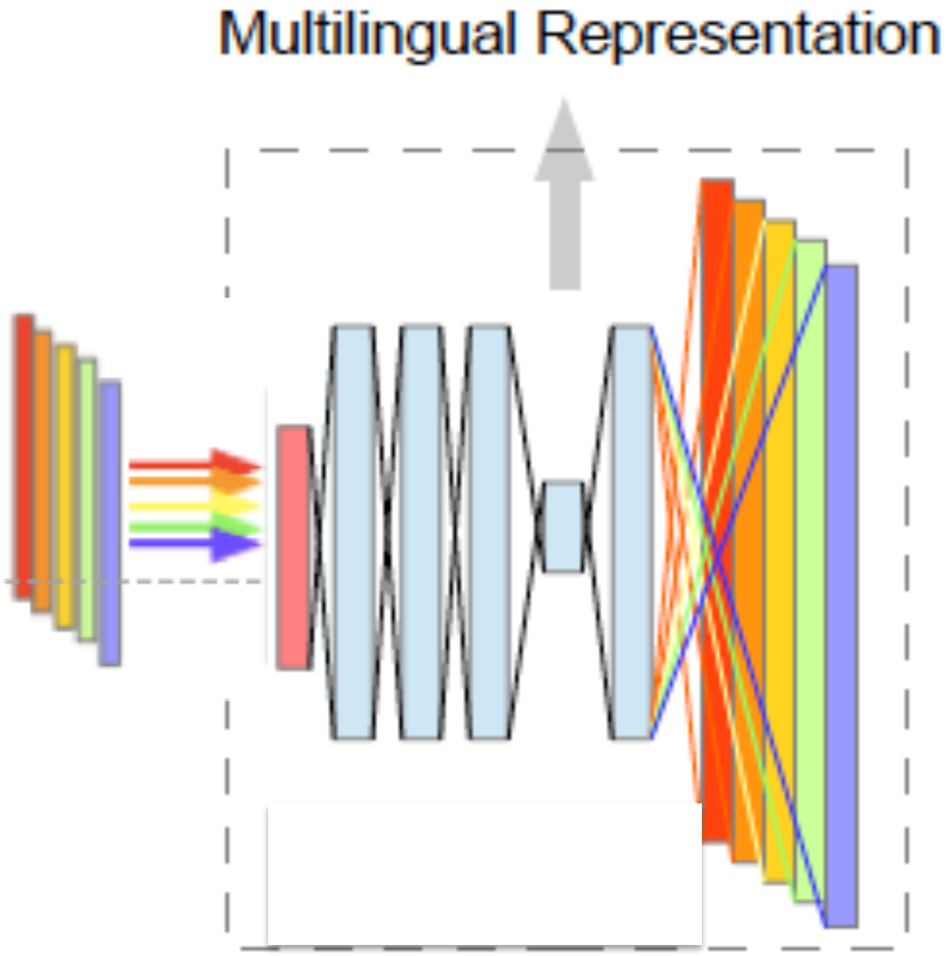
Learning New Languages



- It takes thousands of hours of exposure to learn a second language
- Third language learning may be somewhat faster, with even more ease for more languages
- Very little quantification exists, especially for 3+ languages

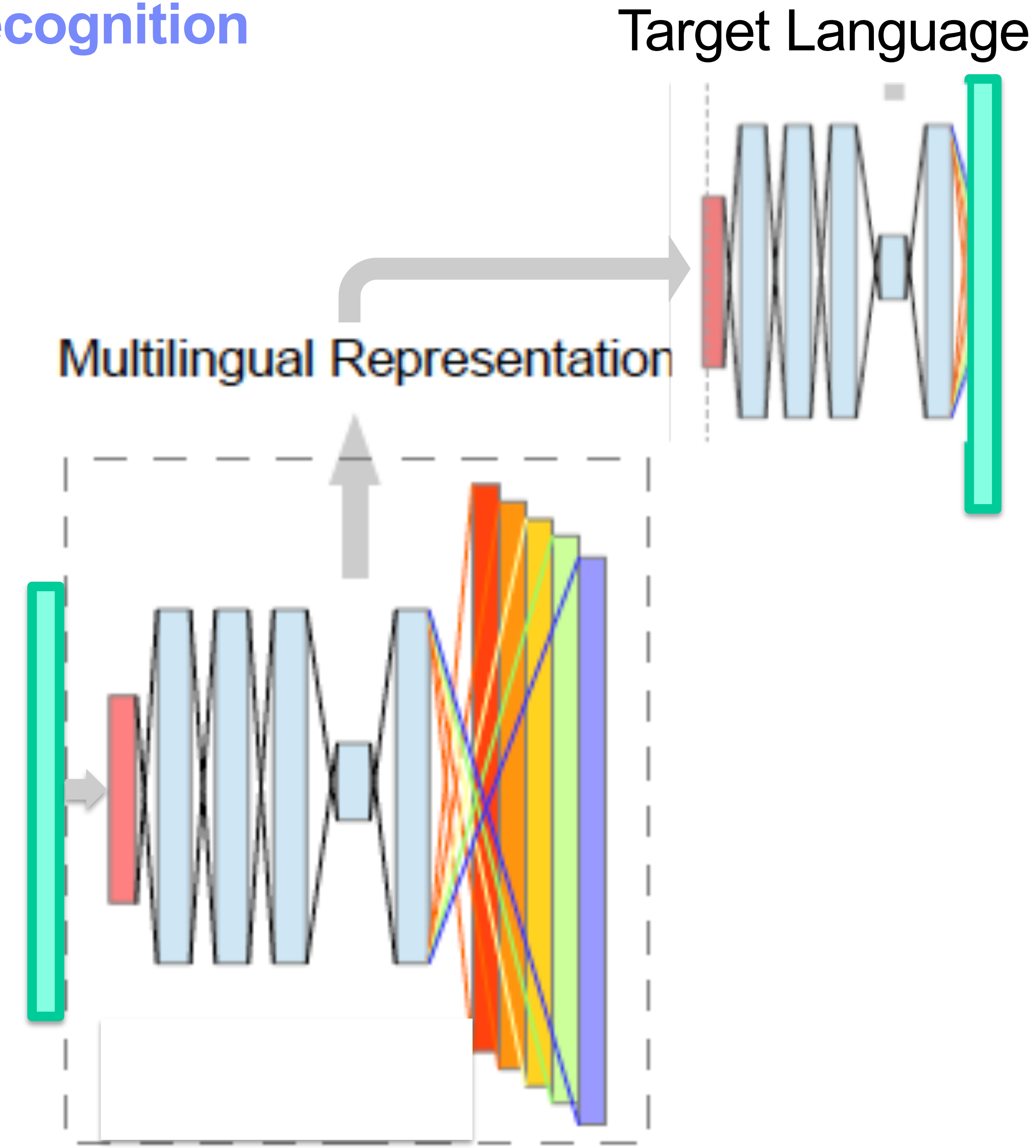
Learning New Languages – Speech Recognition

- Human perception suggests we need 10000 hours of speech
- Perceptual evidence - humans leverage knowledge from other languages. Can machines?
- Babel program looked at this for small-scale amounts of training data but lots of languages (28)



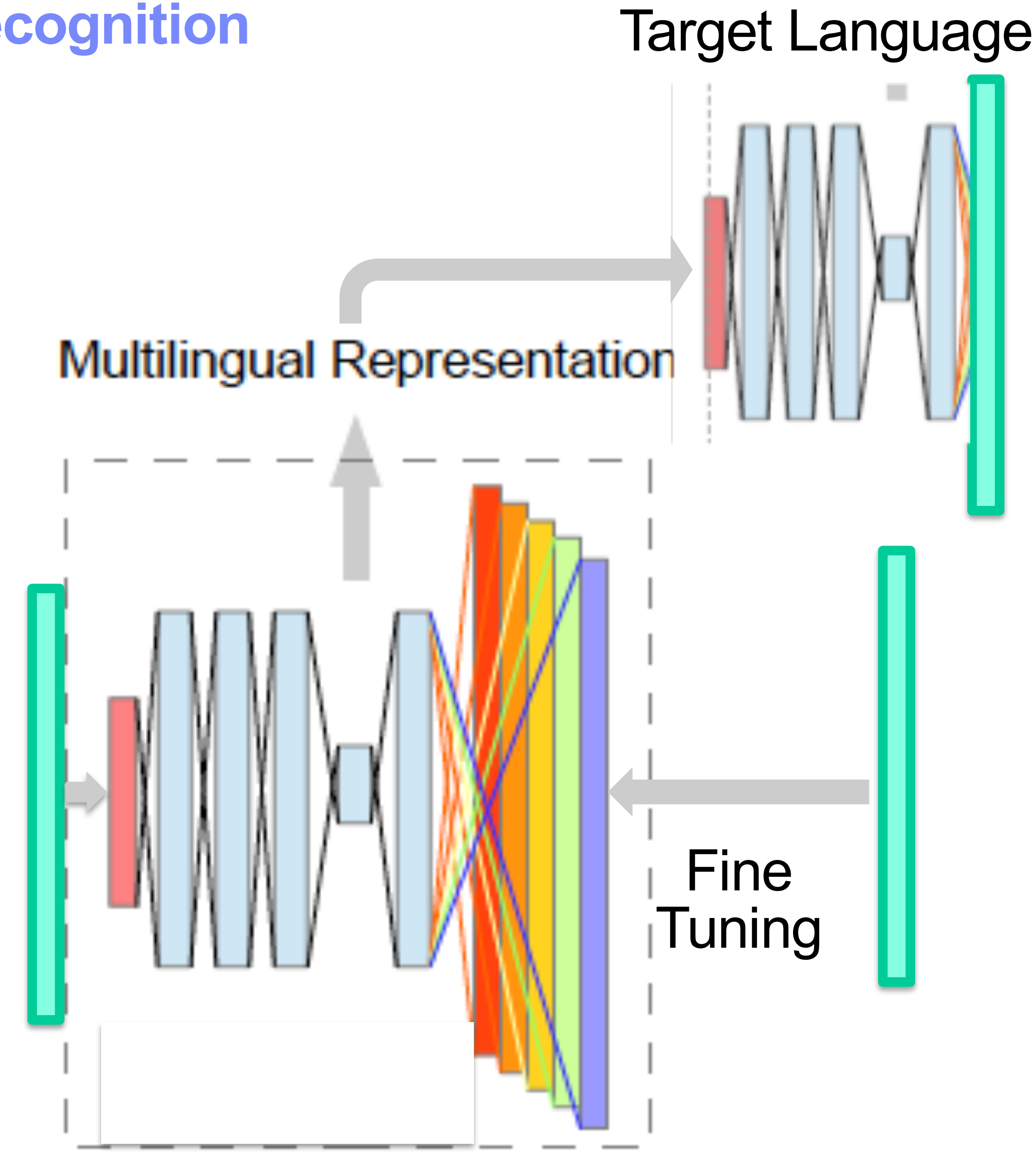
Learning New Languages – Speech Recognition

- Human perception suggests we need 10000 hours of speech
- Perceptual evidence - humans leverage knowledge from other languages. Can machines?
- Babel program looked at this for small-scale amounts of training data but lots of languages (28)



Learning New Languages – Speech Recognition

- Human perception suggests we need 10000 hours of speech
- Perceptual evidence - humans leverage knowledge from other languages. Can machines?
- Babel program looked at this for small-scale amounts of training data but lots of languages (28)



Performance vs. Number of Languages

# of Languages	Training Data (Hours)	WER	
1	41	62.3	
		w/o Fine Tuning	w Fine tuning
11	601	59.6	
17	834	57.2	55.4
24	1110	56.5	
28	1793	56.2	55.1

Javanese, 41 hours of training data

- More languages seem to help performance
- Less clear what happens when we build systems with much more data

Summary

- With a lot of domain-specific data, we can now build systems that rival human performance in that domain.
  - Driven by advances in Deep Learning

- Noise and reverberation robustness seems to have made serious strides as well in terms of being comparable to humans
  - Techniques include multi-style training and multi-microphone processing

- In other areas Humans still seem to be much more capable
  - Adapt quickly to accents
  - More flexible in handling a wide variety of domains
  - Learn languages robustly with considerably less data

- Extremely informal speech such as what we see in meetings is still very challenging
  - No surprise, given the extent to which the acoustic properties of the speech change!

- **Conclusion:** There is still a lot of things for speech recognition researchers to work on!!!

Teşekkür ederim!

References

- [Pad2002] Padmanabhan, M., and Picheny, M. (2002). Large-Vocabulary Speech Recognition Algorithms. IEEE Computer Magazine, April 2002, 42–50.
- [TS2013a] Adapted from T. Sainath , „Deep Neural Networks: Applications for Speech and Language Processing“, Lecture Notes for MIT Course 6.345 Spring 2013.
- [King2012] B. Kingsbury, T. N. Sainath, and H. Soltau, (2012) "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in Proc. Interspeech, 2012.Portland, OR.
- [TS2013b] Tara Sainath et al (2013) "Deep Convolutional Neural Networks for LVCSR, ICASSP 2013, Prague.
- [Saon2014] G. Saon, H. Soltau, A. Emami, and M. Picheny, "Unfolded recurrent neural networks for speech recognition," in *Interspeech 2014*, Singapore, 2014, pp. 343-347.
- [Rennie2014] S. Rennie, V. Goel, and S. Thomas, "Annealed dropout training of deep networks", in Spoken Language Technology (SLT) IEEE Workshop, 2014.
- [Graves2013] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *ICASSP*, pp. 6645-6648, 2013.
- [Sercu2016] T. Sercu, "Very deep multilingual convolutional neural networks for LVCSR," in *ICASSP 2016*, Shanghai, China, 2016, pp. 4955-4959.
- [Chen2009] S. F. Chen, "Shrinking exponential language models," in *Proc. NAACL-HLT*, 2009, pp. 468-476.

References

- [Mangu2007] L. Mangu and A. Emami, "Empirical study of neural network language models for Arabic speech recognition," in *Proc. ASRU*, Kyoto, Japan, 2007, pp. 147–152.
- {Saon2017] G . Saon et. all (2017) "English Conversational Telephone Speech Recognition by Humans and Machines" Interspeech Aug. Stockholm
- [Kurata2017] G. Kurata et all (2017) "Empirical Exploration of Novel Architectures and Objectives for Language Models " Interspeech Aug. Stockholm
- [TS2013c] Tara Sainath et al (2013) "Deep Convolutional Neural Networks for LVCSR, ICASSP 2013, Prague.
- [Saon2015] G. Saon, H.K.J. Kuo, S. Rennie, and M. Picheny, (2015) "The IBM 2015 English Conversational Telephone Speech Recognition System," in *Interspeech 2015*, Lyon pp. 3140-3144
- [Saon2016] G. Saon, T. Sercu, S. Rennie, H.K.J. Kuo, (2016) "The IBM 2016 English Conversational Telephone Speech Recognition System," in *Interspeech 2016*, San Francisco
- [Miller1951] Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology: General*, *41*, 329-335.
- [Moreno1996] Pedro J. Moreno (1996)) Speech Recognition in Noisy Environments (1.3MB), (PDF format) Ph.D. Thesis, ECE Department, CMU,
- [Kino2016] K.Kinoshita, et. (2016). A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 7.
- [Picheny1985] Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of speech and hearing research*, *28*(1), 96-103.
- [Picheny1986] Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of speech and hearing research*, *29*(4), 434-446.
- [Harper2015] Harper, M. (2015, December). The automatic speech recogition in reverberant environments (ASpIRE) challenge. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on* (pp. 547-554). IEEE.

References

- [Renals2014] Renals, S., & Swietojanski, P. (2014, May). Neural networks for distant speech recognition. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on* (pp. 172-176). IEEE.
- [BB2013] M. Baese-Berk (2013) "Accent-independent adaptation to foreign accented speech", *J. Acoust Soc. Amer.* 133:3 EL174
- [Rao2017] K. Rao and H. Sak (2017) "Multi-Accept Speech Recognition with Hierarchical Grapheme based Models" Proc. ICASSP New Orleans LA, pp 4815-4819
- [Eaton2011] Eaton, S. E. (2011). How Long Does It Take to Learn a Second Language?: Applying the. *Online Submission* to Eric https://eric.ed.gov/?id=ED516761
- [Cenoz2001] Cenoz, J. (2001). The effect of linguistic distance, L2 status and age on cross-linguistic influence in third language acquisition. *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*, *111*(45), 8-20.
- [Cui2015] J. Cui et al "Multilingual Representations for Low Resource Speech Recognition and Keyword Search" IEEE ASRU 2015, Phoenix AZ
- [So2016] H. Soltau et al "Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition" arXiv:1610.09975
- {KA2017] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, D. Nahamoo (2017) "*Direct acoustics-to-word models for English Conversational Speech Recognition*", Proc. Interspeech, Stockholm, August.
- [KA2018] K. Audhkhasi, B. Kingsbury, B. Ramabhadran, G. Saon, M. Picheny (2018), *"Building competitive direct acoustics-to-word models for English conversational speech recognition",* ICASSP-2018, Calgary CA