Update on Speech Research

Or…

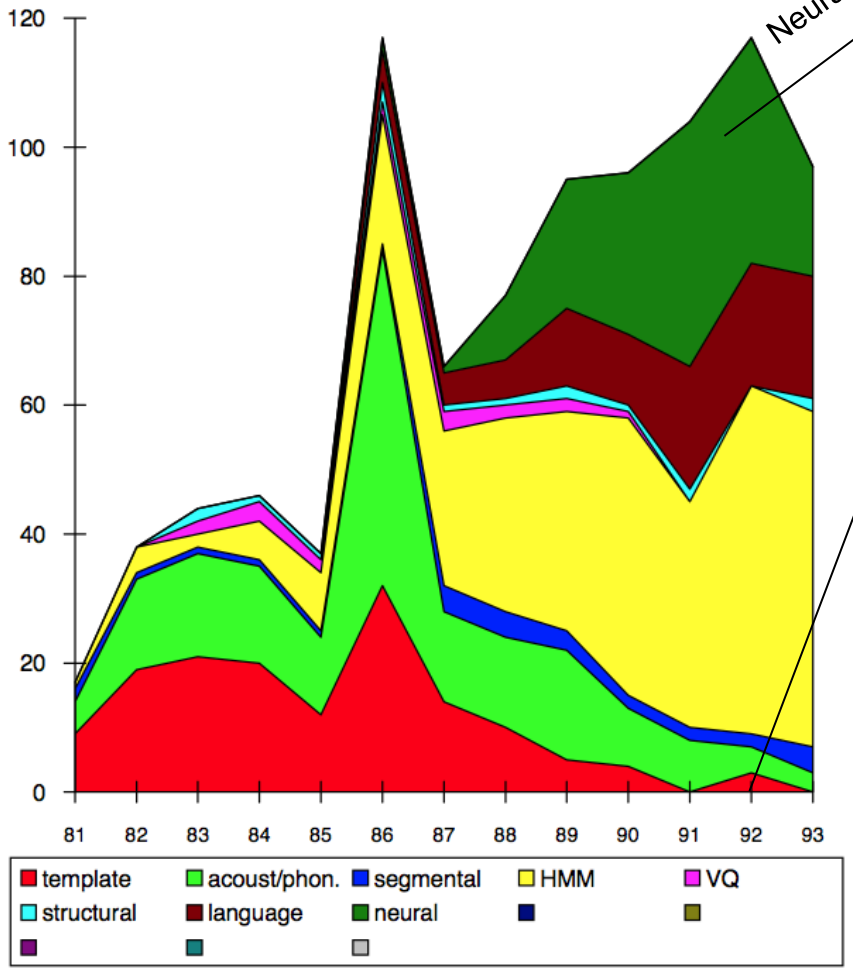
Sometimes they come back

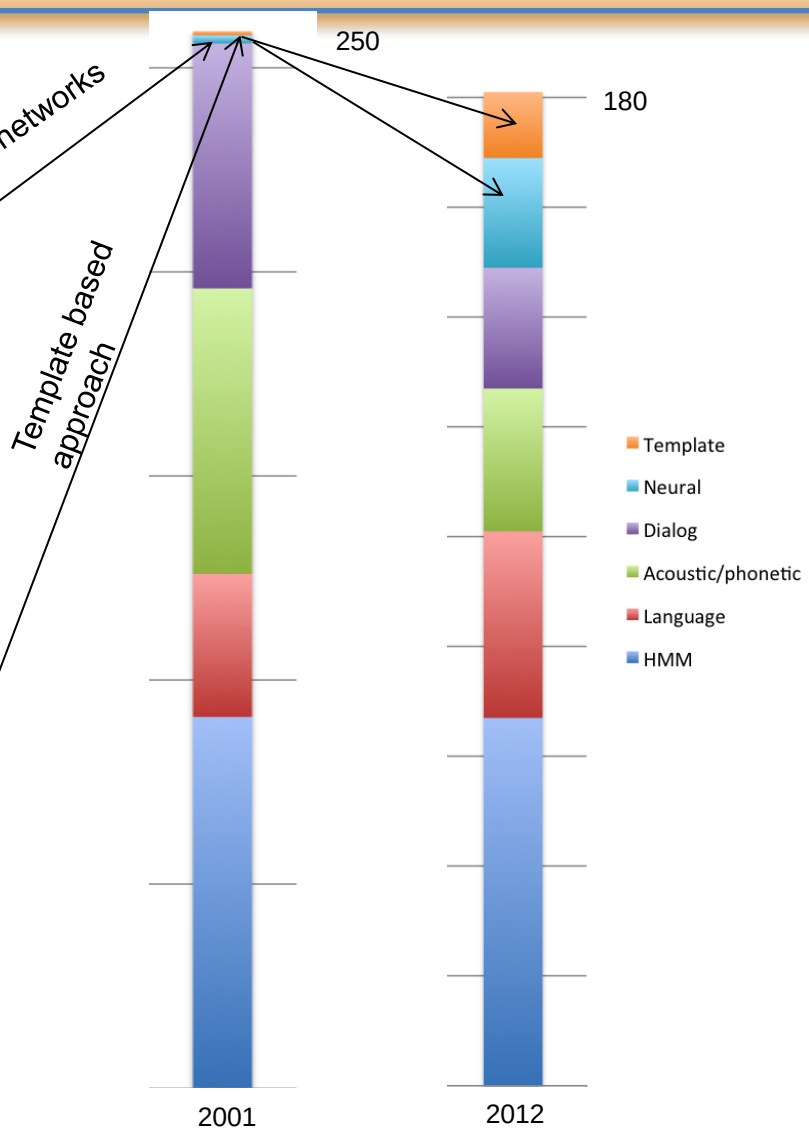
Roberto Pieraccini
CEO and Director
roberto@ICSI.berkeley.edu

What is ICSI?

- Independent non-profit organization
- Located in downtown Berkeley
- Started in 1986 as a joint venture between UCB and the German Research Center for Information Technology
- Affiliation with UCB
- Focus on fundamental research in computer science
  - Networking and Security, Speech, Audio and Multimedia, Vision, AI, Computer Architectures, Brain research, Computational Genetics
- Yearly budget ~$12.5M (2012)
  - 85% US federal contracts/grants, 7.5% Industry, 7.5% International
- ~120 employees in 2012

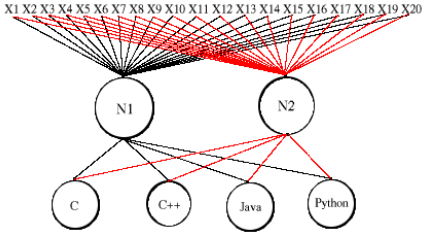
Trends of speech research





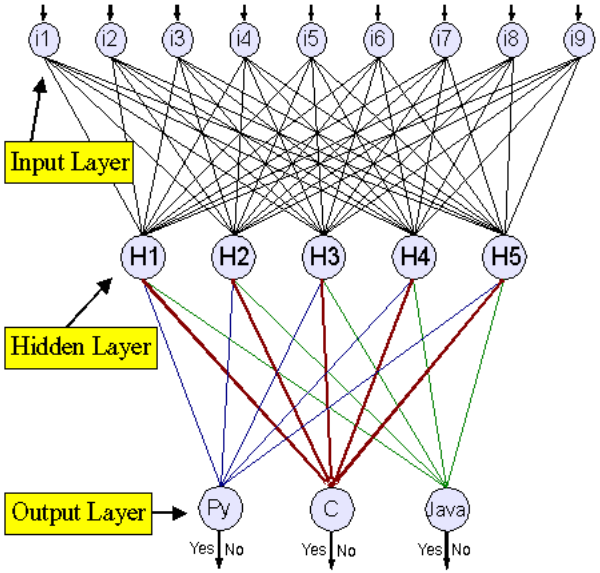
NEURAL NETWORKS IN SPEECH ARE HOT AGAIN

Artificial Neural Networks, a brief history



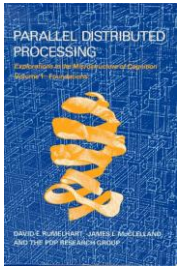
Perceptron (Frank Rosenblatt, 1957)

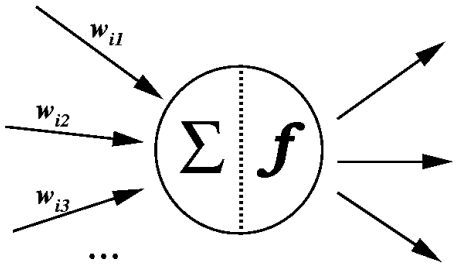
Perceptrons (Minsky-Papert, 1969) shows that single layer perceptrons are only capable of classifying linearly separable patterns, and computers were not powerful enough to handle large neural networks.



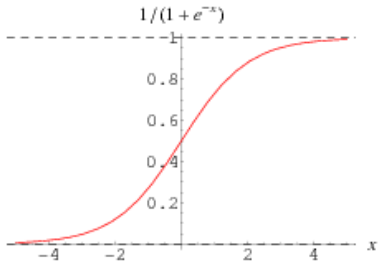
Invention of the back propagation algorithm and its use on multi-layer perceptrons (MLP) (or *artificial neural networks ANN)* (work by Werbos, Rumelhart, Hinton, Williams, 1974 and later)

Rumelhart and McClelland publish *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* in 1986. Neural Networks become popular and mainstream. (Scientific American Hopfield)



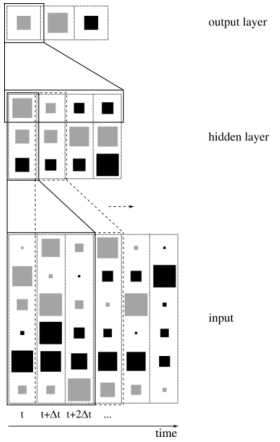


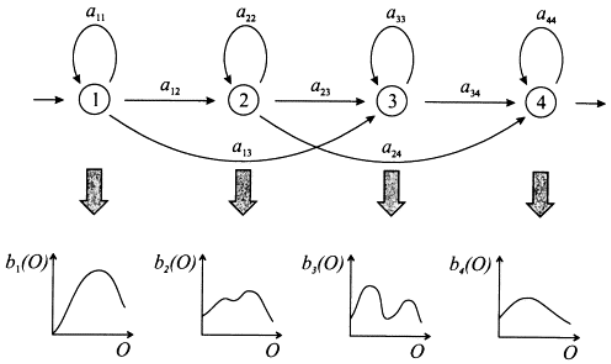
$$y_i = f(net_i)$$

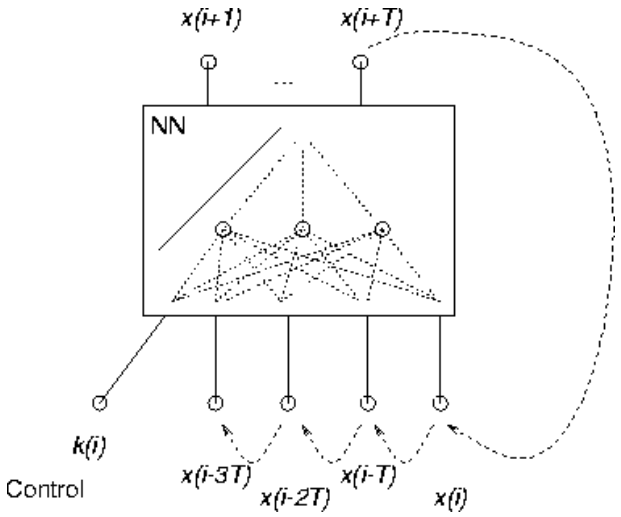


Artificial Neural Networks in Speech

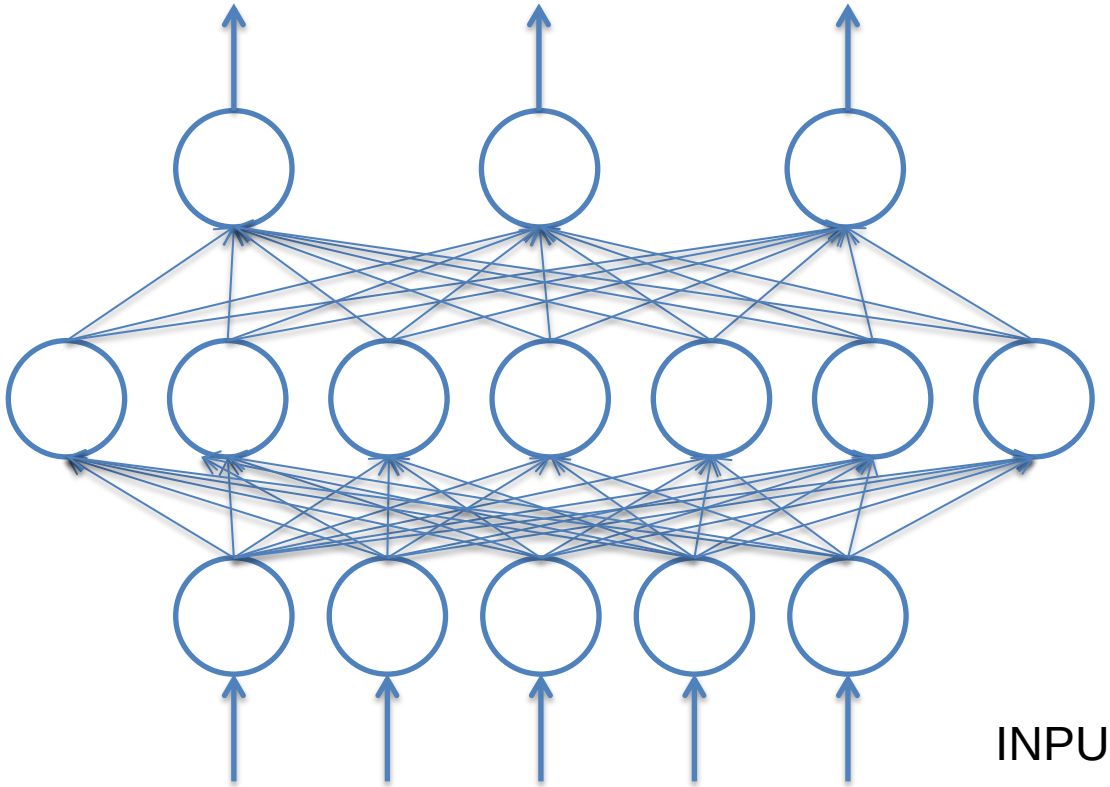
- Neural Networks do not have an in-built capability to handle time warping, so their adoption in speech required special modifications.
- Speech/non-speech classification (Morgan, 1983)
- Speech event classification (Makino, 1983)
- Recurrent ANN (Fallside, Robinson, 1989)
- Time-Delay neural Networks (Alex Waibel et als., 1989)
- Hybrid HMM/ANN (Morgan, Bourlard, 1989)
- Hidden Control Neural Networks (Esther Levin, 1990)
- Eventually used in a time-static manner to estimate probability distributions of observations in HMMs



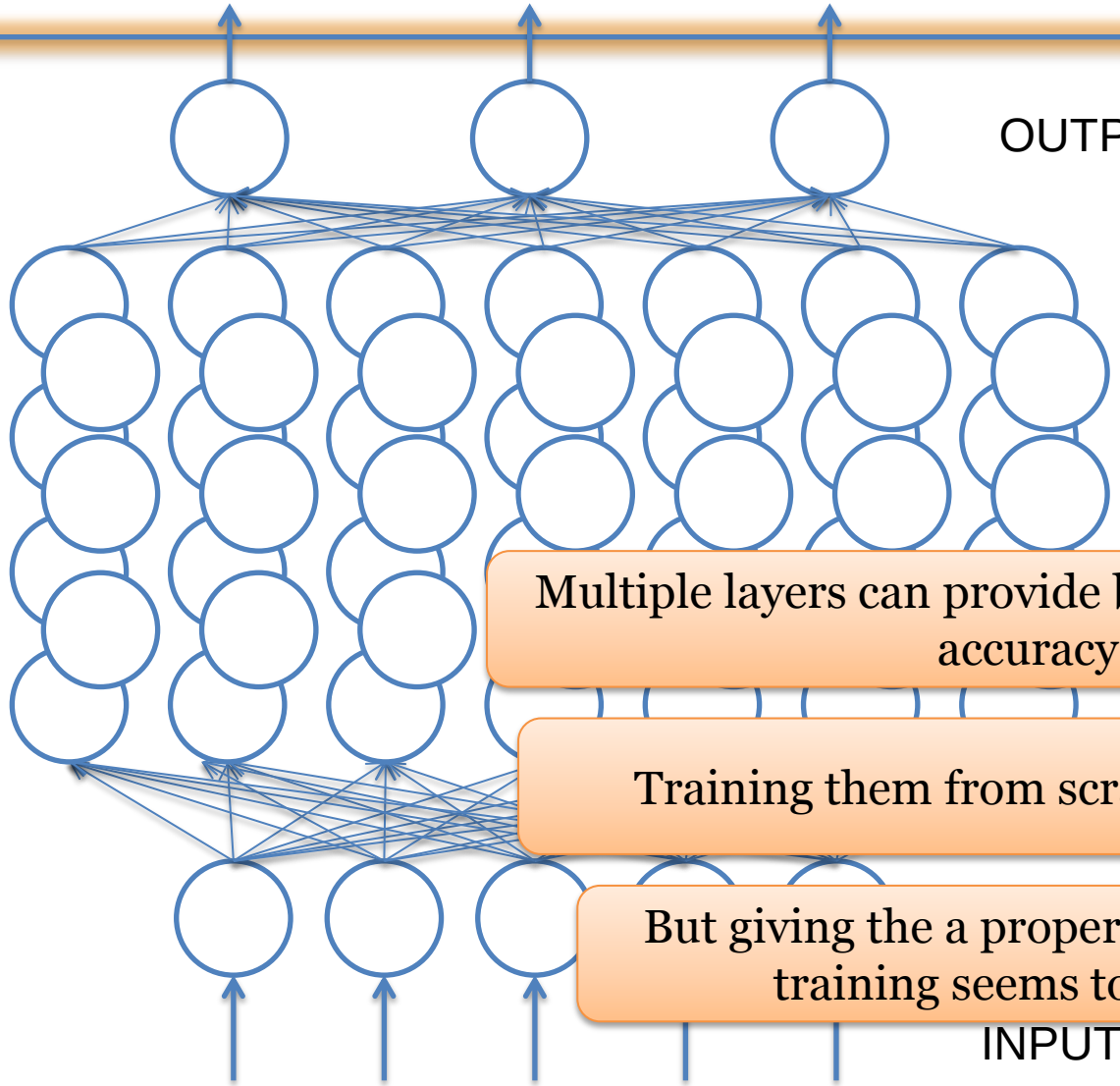




Deep Neural Networks



Deep Neural Networks



Multiple layers can provide better classification accuracy

Training them from scratch did not work

But giving the a proper initialization and then training seems to work quite well

Significant speech recognition improvement

Configuration	Test WER
CD-GMM-HMM (BMMI)	34.8%
2kx5	27.4%
2kx2-(64:64)x1-2kx2	26.8%
2kx4-(64:64)x1	26.4%
2kx4-(96:96)x1	26.2%

Microsoft Research, Switchboard (Dong, Deng, Seide, Interspeech 2012)

Training	WER	BMMI objfun.
ML	23.6%	0.16
FMMI	20.3%	0.18
FMMI-BMMI	18.7%	0.20

IBM research, broadcast news (Saon, Kingsbury, Interspeech 2012)

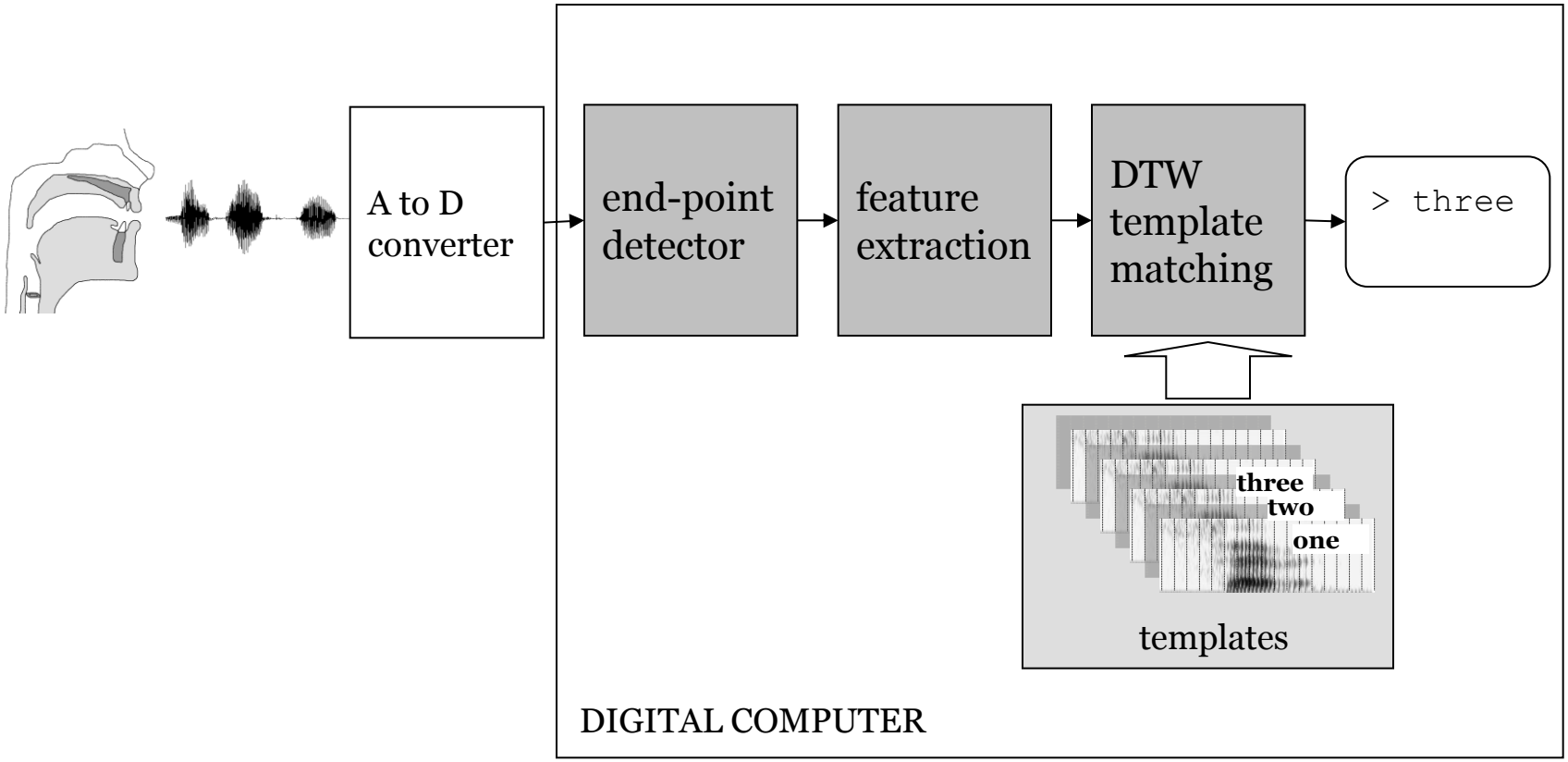
Name	Model	WER(%)
Voice Search	GMM-HMM baseline	16.0
	DBN pretrained ANN/HMM with sparsity	12.3
	+ *MMI*	12.2
	+ *system combination with SCARF*	**11.8**
YouTube	GMM-HMM baseline	52.3
	DBN pretrained ANN/HMM with sparsity	47.6
	+ *MMI*	47.1
	+ *system combination with SCARF*	**46.2**

Google (Jaitly et als, Interspeech 2012)

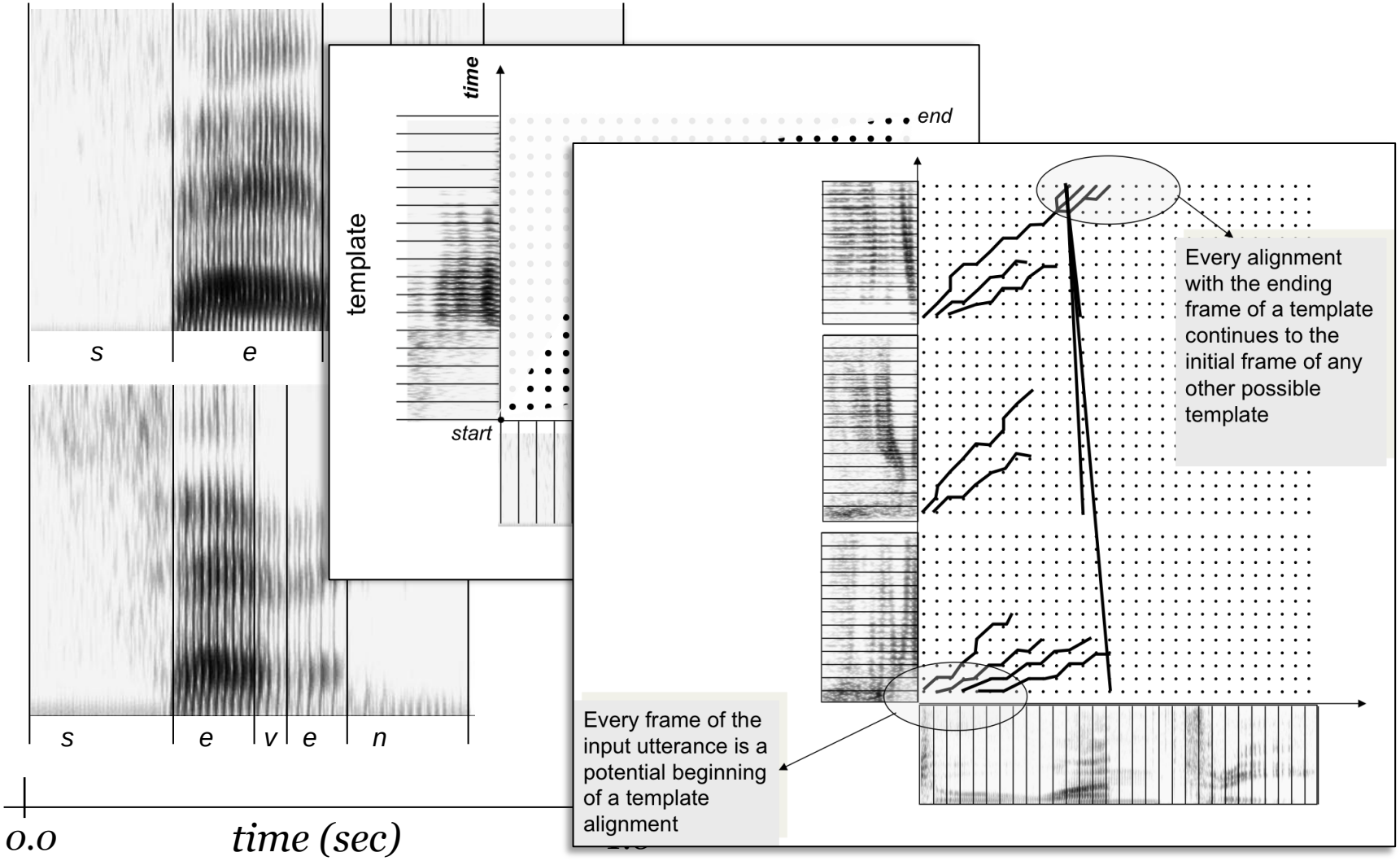
TEMPLATES IN SPEECH ARE HOT AGAIN

Template-matching approach

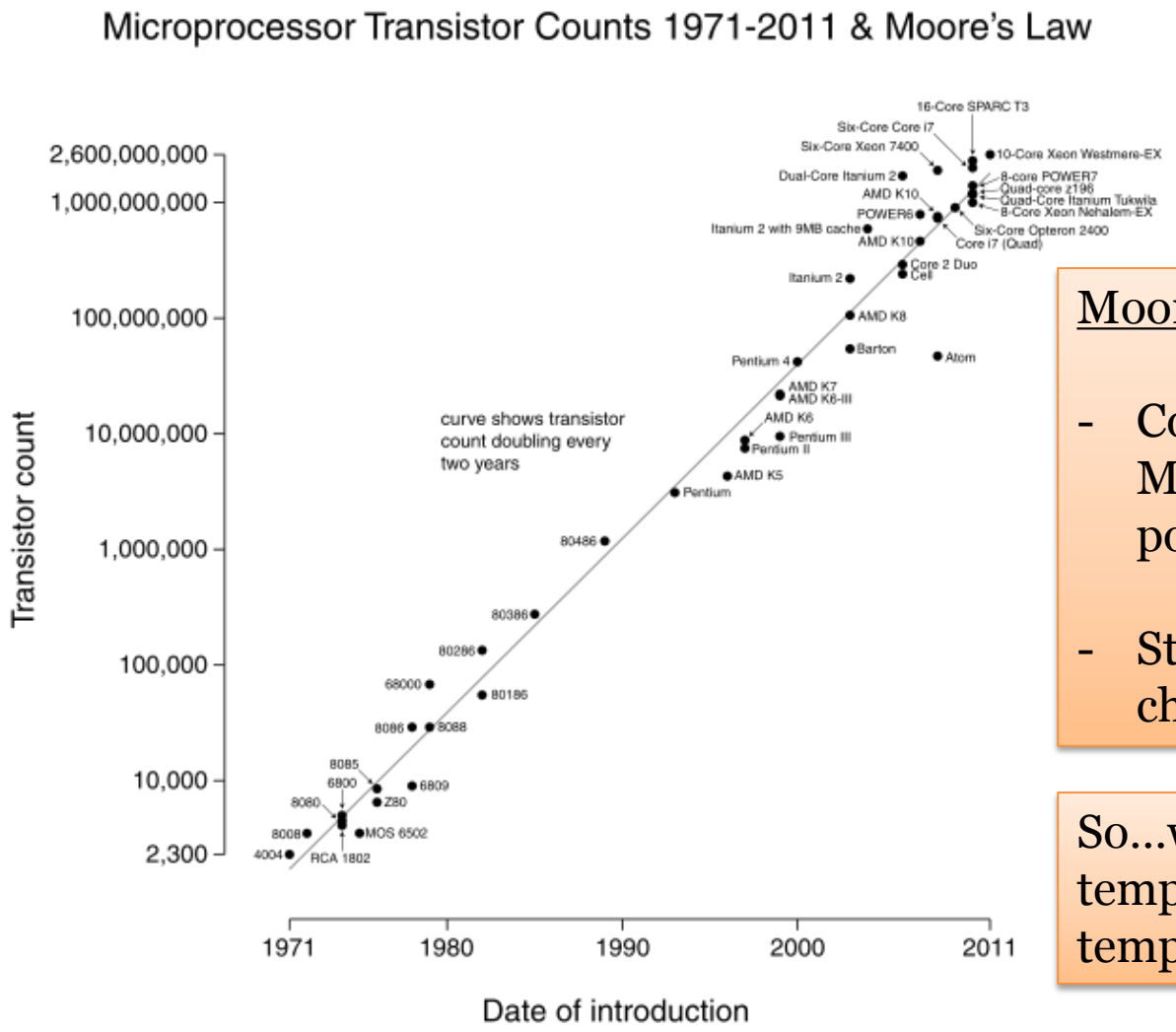
the mother of modern speech recognition



Matching with Dynamic Time Warping



So what's new?



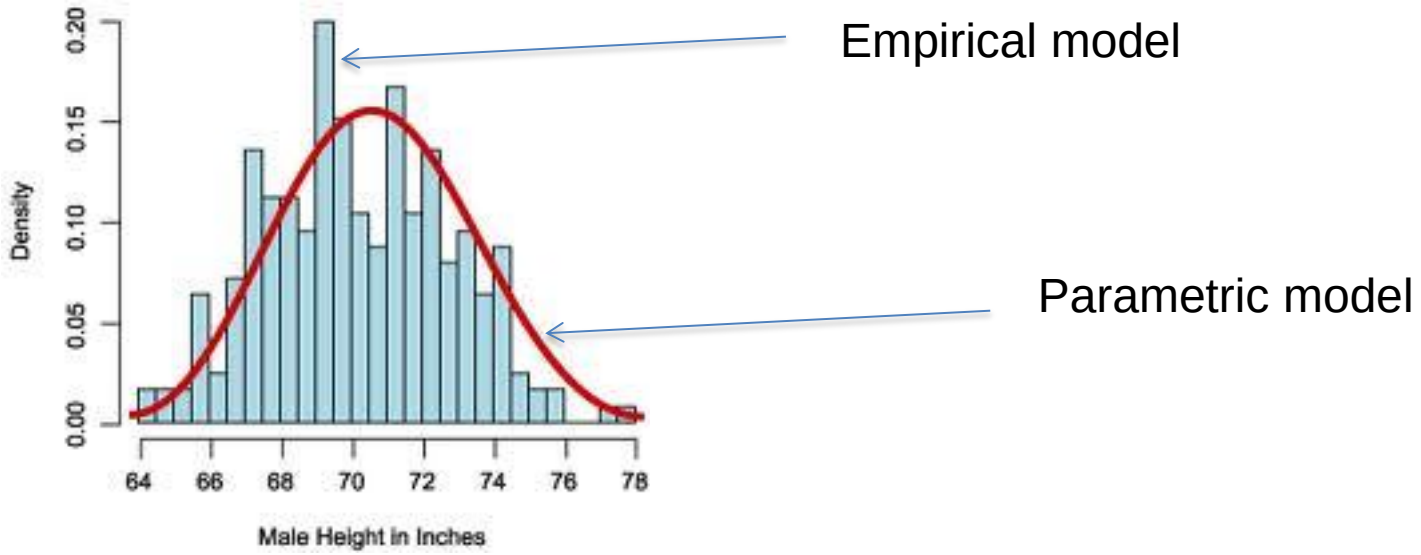
Moore's Law:

- Computers today are 1 Million times more powerful than in 1970
- Storage is 1 Million times cheaper

So…why not using more templates? Way more templates?

Why templates? rather than statistical models?

- Statistical models try to force data to fit a model.
- The model may make wrong assumptions.
- The resulting model does not represent the data accurately because we try to fit the data to a model (for instance a bell curve)

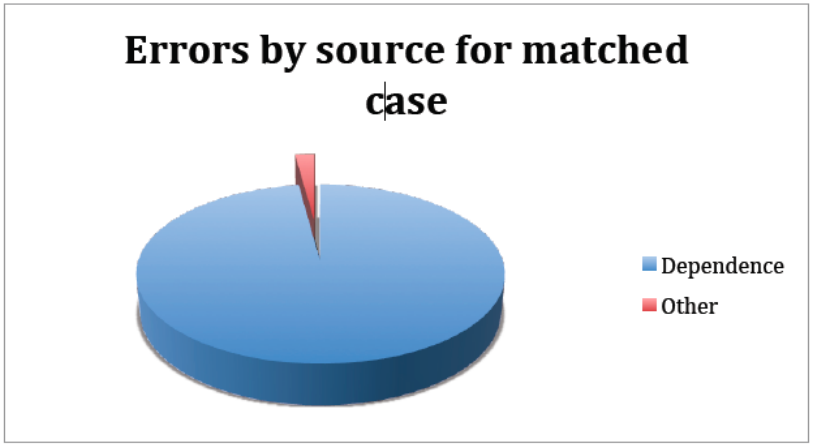


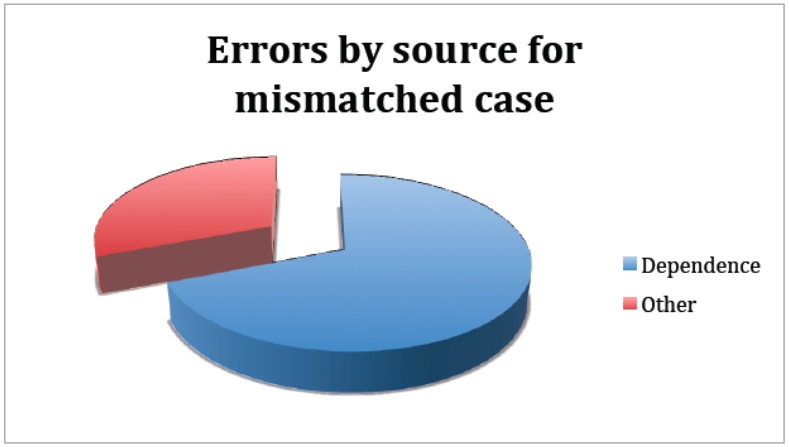
- If we had enough samples, we could use the data itself as a model, rather than making assumptions on the underlying distribution.

Large scale study on the effect of wrong model assumptions the OUCH project at ICSI (Wegmann, Morgan, Cohen, 2013)

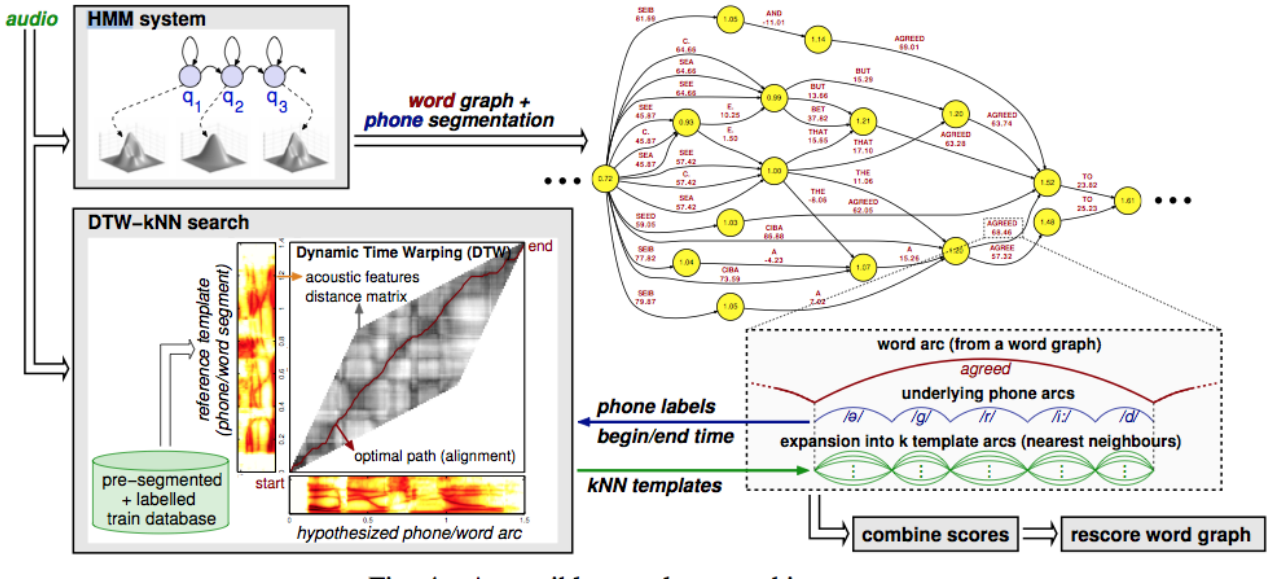
Resampling method	WER (%)	Standard Error	Δ WER (%)
Original data	44.7	-	19

ICSI meeting parallel corpus, near field, far field microphones.





Template based speech recognition research today



From "Exemplar-Based Processing for Speech Recognition, Sainath et als", IEEE Signal Processing Magazine, 2012

Work on template-based approach by Van Compernolle et. als (Univ of Leuven, Belgium), Nguyen and Zweig (MS Research), Sainath, Ramabhadran, Nahamoo, Kanesky et als (IBM Research)

Latest results show that a large number of templates (in the millions) can perform better than HMMs in simple tasks (e.g. 1000 word vocabulary) and comparably in larger tasks

Conclusions

- 1,000,000 times more powerful computers and storage give us a chance to review old techniques which were dormant for decades
- The Deep (as in Deep Learning) return of Artificial Neural Networks
- The Big (as in Big Data) return of Template Matching
- Work in progress