Whither speech recognition?
- Deep learning to deep thinking -

Sadaoki Furui
Toyota Technological Institute at Chicago
furui@ttic.edu

Outline

1. Generations of ASR technology
2. Recent success by deep learning (DNN)
3. J. R. Pierce: "Whither speech recognition?"
4. Speech recognition as a *prediction* process
   - Vowel reduction
   - Spectral dynamics and syllable perception
5. Multi-view learning of speech representations
6. Speech recognition by comprehensive knowledge processing
7. Conclusion

Outline

1. Generations of ASR technology
2. Recent success by deep learning (DNN)
3. J. R. Pierce: “Whither speech recognition?”
4. Speech recognition as a *prediction* process
   – Vowel reduction
   – Spectral dynamics and syllable perception
5. Multi-view learning of speech representations
6. Speech recognition by comprehensive knowledge processing
7. Conclusion

Generations of ASR technology

1950 1960 1970 1980 1990 2000 2010

1952 **1G** 1970

Heuristic approaches
(analog filter bank + logic circuits)

1970 **2G** 1980

Pattern matching
(LPC, FFT, DTW)

1980 **3G** 1990

Statistical framework
(HMM, n-gram, neural net)

1990 **3.5G** 2010

Discriminative approaches, machine learning, robust training, adaptation, rich transcription

2010 **4G**

Deep learning (DNN)

Prehistory ASR (1925)

Our research
NTT Labs (+Bell Labs), Tokyo Tech, Toyota Tech. Inst. at Chicago

Outline

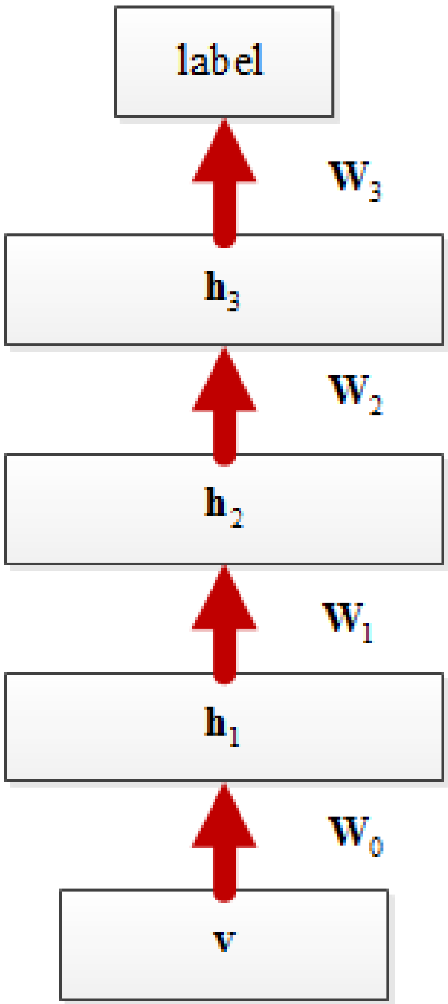
1. Generations of ASR technology
2. Recent success by deep learning (DNN)
3. J. R. Pierce: “Whither speech recognition?”
4. Speech recognition as a *prediction* process
   – Vowel reduction
   – Spectral dynamics and syllable perception
5. Multi-view learning of speech representations
6. Speech recognition by comprehensive knowledge processing
7. Conclusion

Deep neural network

- Multi-layer perceptron (MLP) with many hidden layers
- The last layer follows multinomial distribution

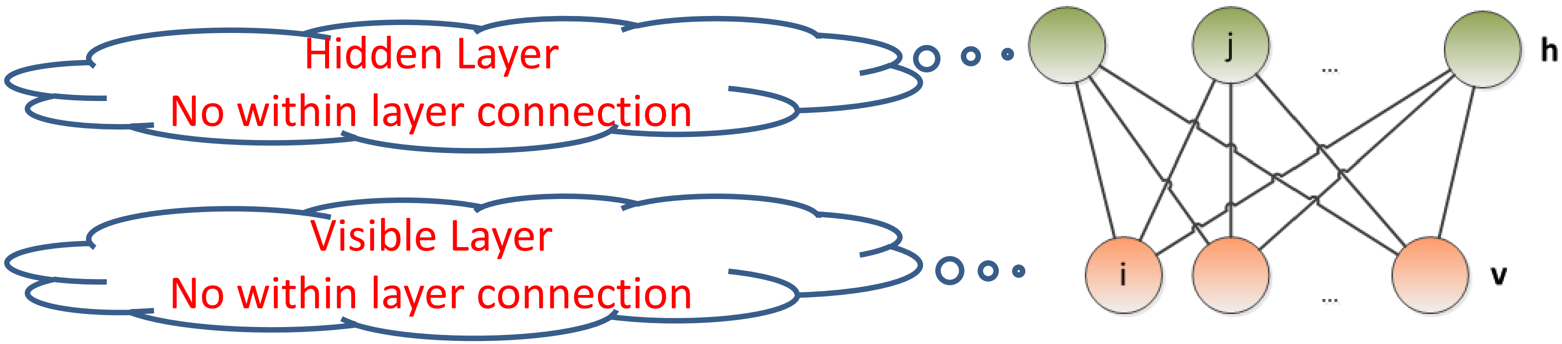
$$p(l = k|\mathbf{h};\theta) = \frac{exp\left(\sum_{i=1}^{H} \lambda_{ik} h_i + a_k\right)}{Z(\boldsymbol{h})}$$

- Nonlinear feature extraction: higher layer features are more invariant and discriminative than lower layer features
- Training deep neural network is hard: generative and discriminative pretrain



(Dong Yu, 2012)

Restricted Boltzmann machine



- Joint distribution $\mathrm{p}(\mathbf{v},\mathbf{h};\theta)$ is defined in terms of an energy function $\mathrm{E}(\mathbf{v},\mathbf{h};\theta)$

$$\mathrm{p}(\mathbf{v},\mathbf{h};\theta)=\frac{exp\left(-\mathrm{E}(\mathbf{v},\mathbf{h};\theta)\right)}{Z}$$

$$\mathrm{p}(\mathbf{v};\theta)=\sum_{\mathbf{h}}\frac{exp\left(-\mathrm{E}(\mathbf{v},\mathbf{h};\theta)\right)}{Z}=\frac{exp\left(-\mathrm{F}(\boldsymbol{v};\theta)\right)}{Z}$$

- Conditional independence

$$p(\mathbf{h}|\mathbf{v})=\prod_{\mathrm{j}=0}^{H-1}p(h_j|\mathbf{v})$$

$$p(\mathbf{v}|\mathbf{h})=\prod_{\mathrm{i}=0}^{V-1}p(v_i|\mathbf{h})$$

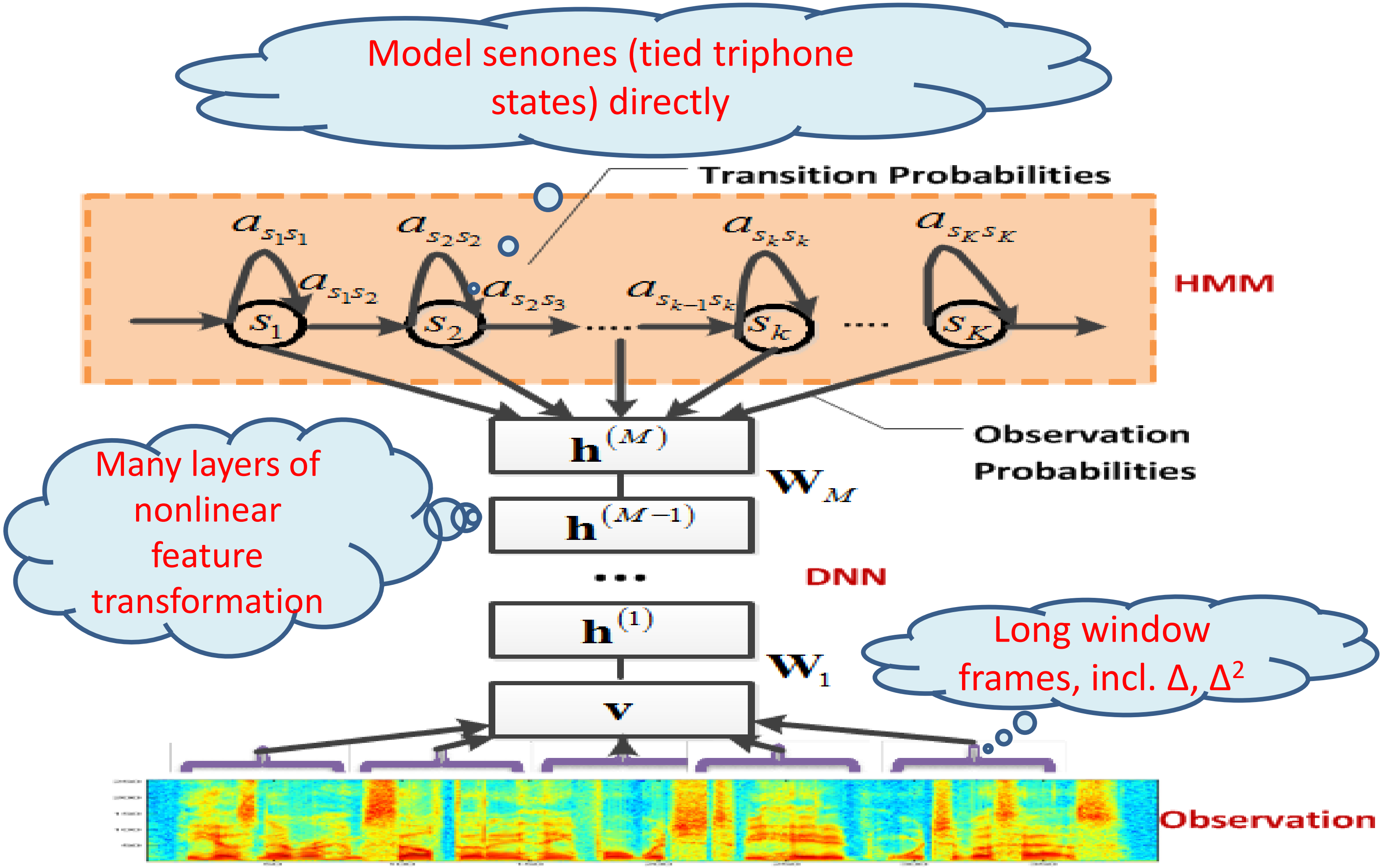
(Dong Yu, 2012)

Why deep network is helpful

- Many simple non-linearities = One complicated non-linearity
- More efficient in representation: need fewer computational units for the same function
- More constrained space of transformations determined by the structure of the model – less likely to overfit
- Lower layer features are typically task independent (e.g., edges) and thus can be learned in an unsupervised way.
- Higher layer features are task dependent (e.g., object parts or object) and are easier to learn given the low-level features.
- Higher layers are easier to be classified using linear models.

(Dong Yu, 2012)

CD-DNN-HMM: 3 key components



(Dong Yu, 2012)

Empirical evidence: Summary

(Dahl, Yu, Deng, Acero 2012, Seide, Li, Yu 2011 + new result)

- Voice Search SER (24 hours training)

AM	Setup	Test
GMM-HMM	MPE	36.2%
DNN-HMM	5 layers x 2048	30.1% (-17%)

- Switch Board WER (309 hours training)

AM	Setup	Hub5'00-SWB	RT03S-FSH
GMM-HMM	BMMI (9K 40-mixture)	23.6%	27.4%
DNN-HMM	7 x 2048	15.8% (-33%)	18.5% (-33%)

- Switch Board WER (2000 hours training)

AM	Setup	Hub5'00-SWB	RT03S-FSH
GMM-HMM (A)	BMMI (18K 72-mixture)	21.7%	23.0%
GMM-HMM (B)	BMMI + fMPE	19.6%	20.5%
DNN-HMM	7 x 3076	14.4% (A: -34% B: -27%)	15.6% (A: -32% B: -24%)

(Dong Yu, 2012)

Deeper models more powerful?

(Seide, Li, Yu 2011, Seide, Li, Chen, Yu 2011)

L×N	DBN-Pretrain	BP	LBP	Discri-Pretrain	1×N	DBN-Pretrain
1×2k	24.2	24.3	24.3	24.1	1×2k	24.2
2×2k	20.4	22.2	20.7	20.4	-	-
3×2k	18.4	20.0	18.9	18.6	-	-
4 ×2k	17.8	18.7	17.8	17.8	-	-
5×2k	17.2	18.2	17.4	17.1	1×3772	22.5
7 ×2k	17.1	17.4	17.4	16.8	1×4634	22.6
9×2k	17.0	16.9	16.9	-	-	-
9× 1k	17.9	-	-	-	-	-
5×3k	17.0	-	-	-	-	-
					1× 16k	22.1

Compare BP with DBN pre-training, pure backpropagation (BP), layer-wise BP-based model growing (LBP), and discriminative pretraining. Shown are word-error rates in %. ML alignment.

(Dong Yu, 2012)

Improving deep learning

- Better optimization
- Better types of neural activation function and better network architectures
- Better ways to determine the myriad hyper-parameters of DNNs
- More appropriate ways to preprocess speech for DNNs
- Ways of leveraging multiple languages or dialects that are more easily achieved with DNNs than with GMMs
- Using more computing power, more training data, and better software engineering (e.g., distributed frameworks)

(Li Deng et al., 2013)

Discoveries

- When there is a large amount of labeled data, the main effect of the pre-training is just to get the initial weights to be about the right scale so that back-propagation works well. But there are simpler ways of doing this.
- DNNs work well for noisy speech.
- DNNs work much better for acoustic modeling if we use one or more convolutional layers that do weight-sharing across nearby frequencies and then pool the filter responses to similar frequencies.
- Rectified linear units and "dropout" are very effective.
- The same methods can be used for application other than acoustic modeling (e.g., language modeling).
- The DNN architecture can be used for multi-task learning in several different ways.

(Li Deng et al., 2013)

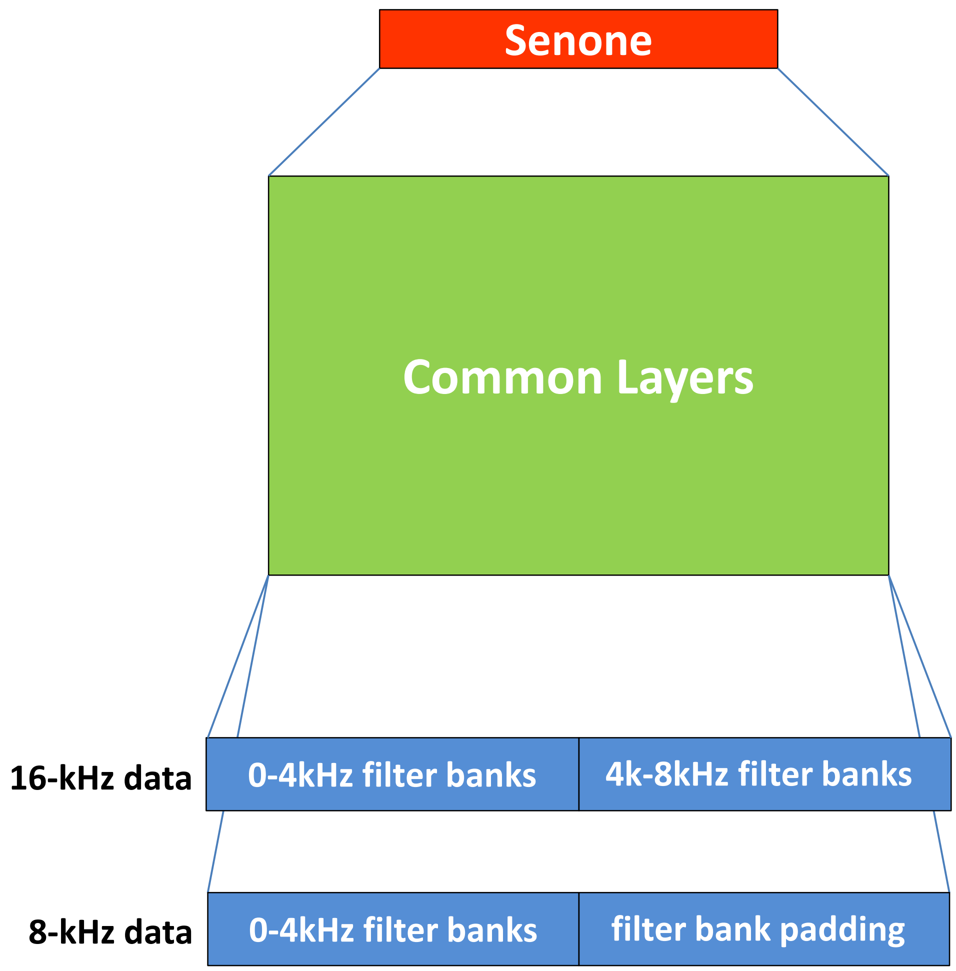
A comparison of several systems in the literature to a DNN system on the Aurora 4 task (word error rate(%))

Systems	Distortion				AVG
	None (clean)	Noise	Channel	Noise+ channel	
GMM baseline	14.3	17.9	20.2	31.3	23.6
MPE-NAT + VTS	7.2	12.8	11.5	19.7	15.3
NAT + Derivative Kernels	7.4	12.6	10.7	19.0	14.8
NAT + Joint MLLR/VTS	5.6	11.0	8.8	17.8	13.4
DNN (7x2048)	5.6	8.8	8.9	20.0	13.4

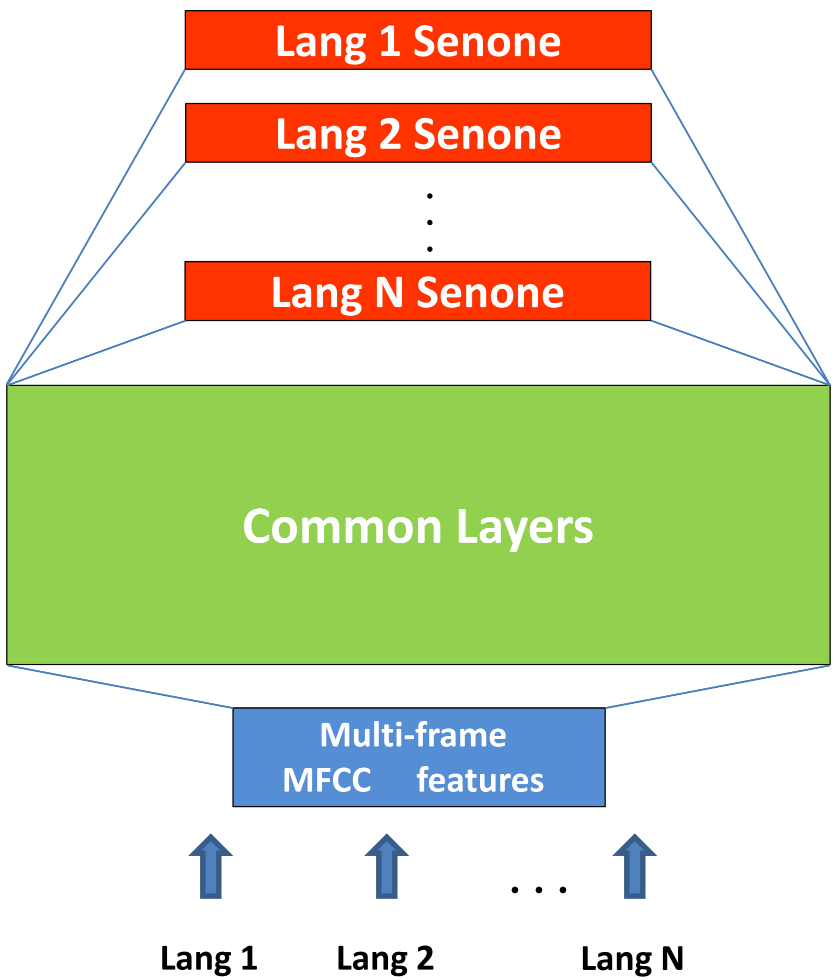
The DNN result (multi-condition training) was obtained in a single pass, while the above two systems require multiple passes for adaptation.
NAT: noise adaptive training

(M. Seltzer et al., 2013)

Multi-task learning by DNN



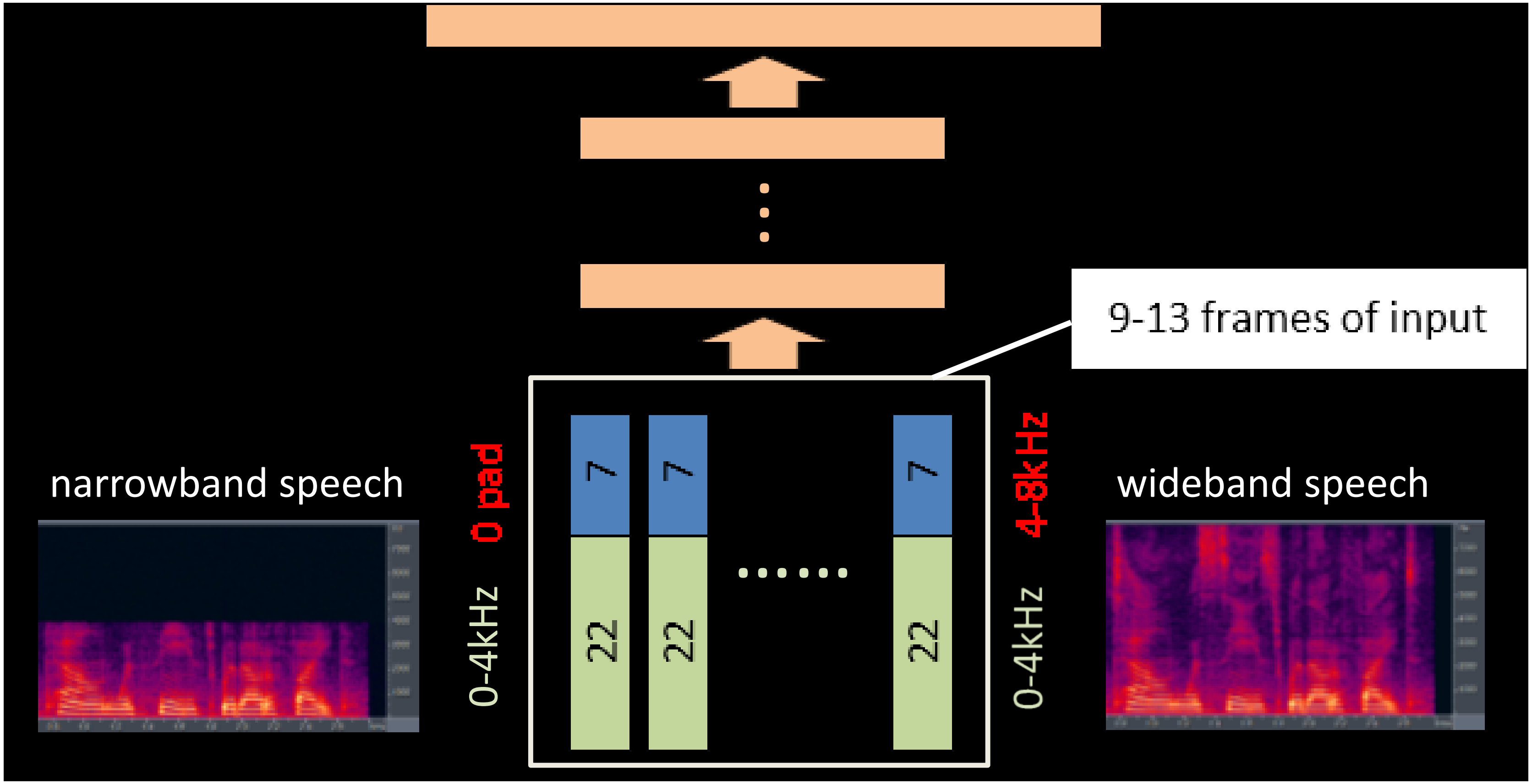
16-kHz and 8-kHz mixed band modeling



Multilingual modeling

(Li Deng et al., 2013)

Illustration of mixed-bandwidth speech recognition using DNN



DNN acoustic adaptation

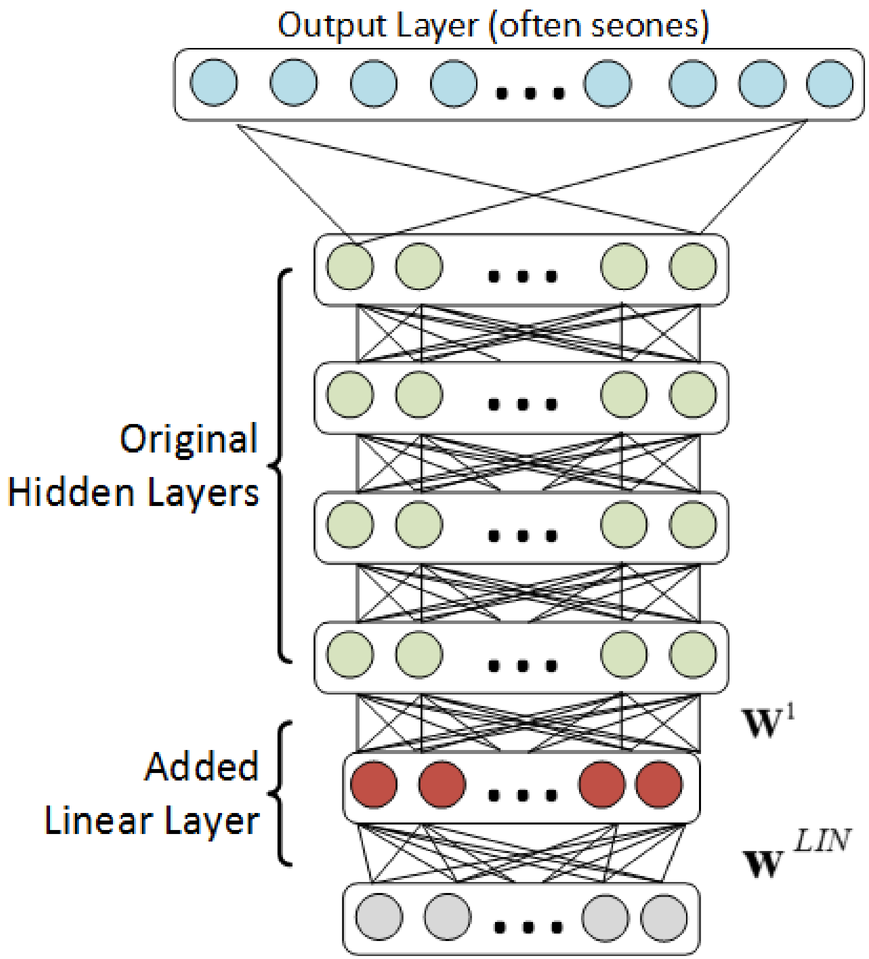
- Feature space transforms (feature normalization)
  - fMLLR
  - fDLR (feature-space discriminative linear regression)
  - LIN (linear input network): An additional speaker dependent layer between the input features and the 1st hidden layer
- Auxiliary features
  - i-vectors: The basis vectors which span a subspace of speaker variability
  - Speaker-specific bottleneck features
- Model-based adaptation
  - DNN parameters are adapted directly
  - Factorization based on SVD
  - Various ways to reduce the DNN weights to be modified

Comparison of feature-transform based speaker-adaptation techniques for GMM-HMMs, a shallow, and a deep NN. Word-error rates in % for Hub5'00-SWB, ( ): relative change

Adaptation technique	GMM-HMM 40mix	CD-MLP-HMM 1x2k	CD-DNN-HMM 7x2k
Speaker independent	23.6	24.2	17.1
+ VTLN	21.5 (-9%)	22.5 (-7%)	16.8 (-2%)
+ {fMLLR/fDLR} x4	20.4 (-5%)	21.5 (-4%)	16.4 (-2%)

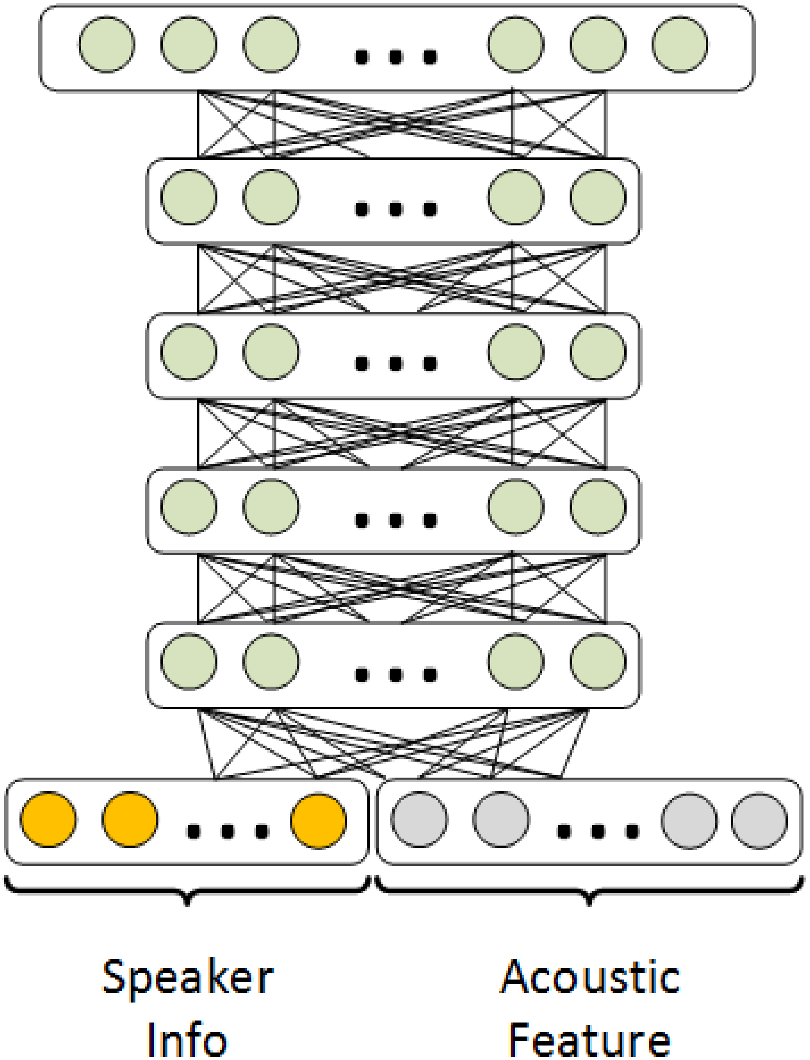
(F. Seide, et al., 2011)

Feature normalization by LIN (linear input network)



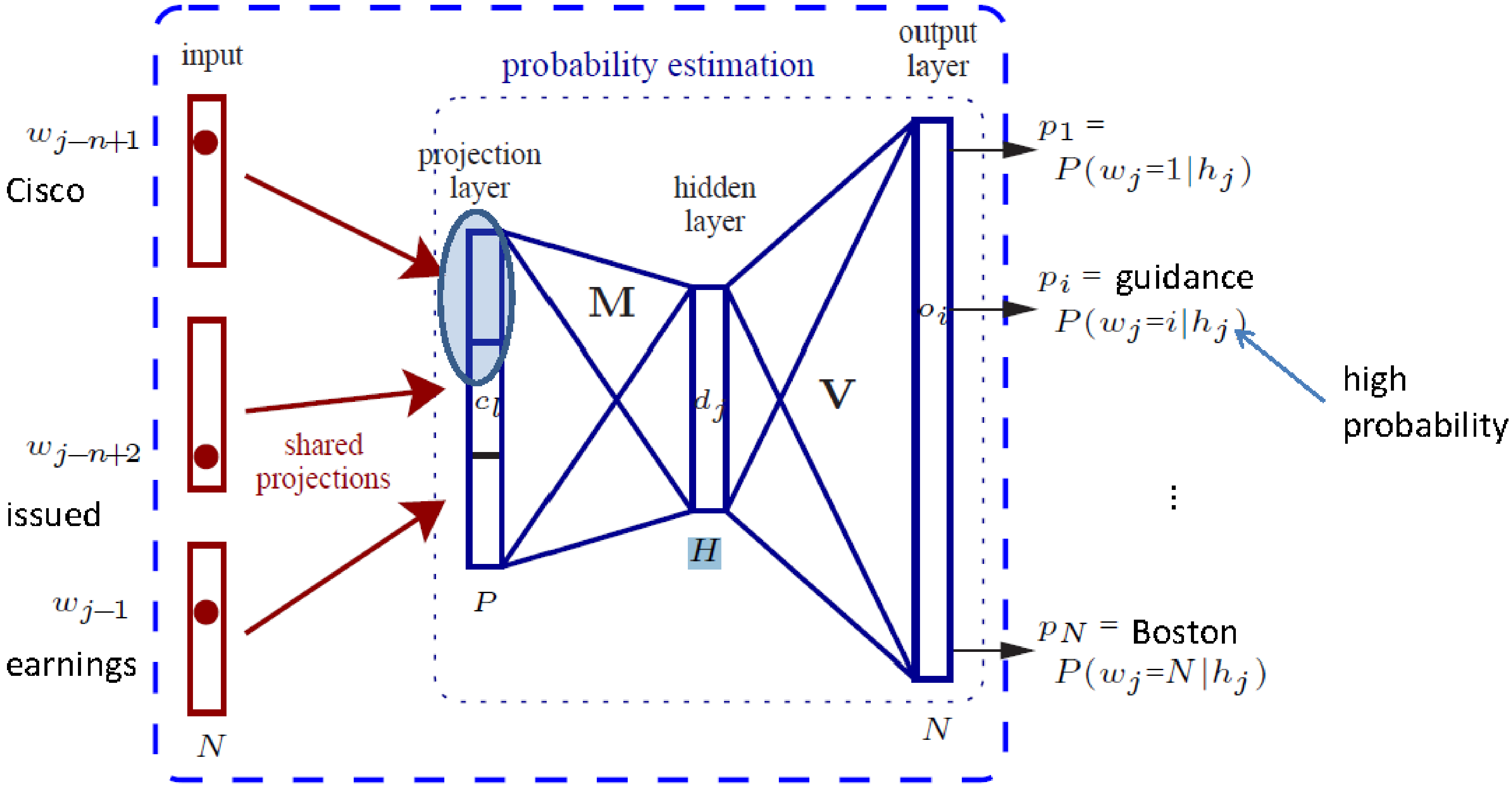
(D. Yu and L. Deng, 2014)

Adaptation using auxiliary features



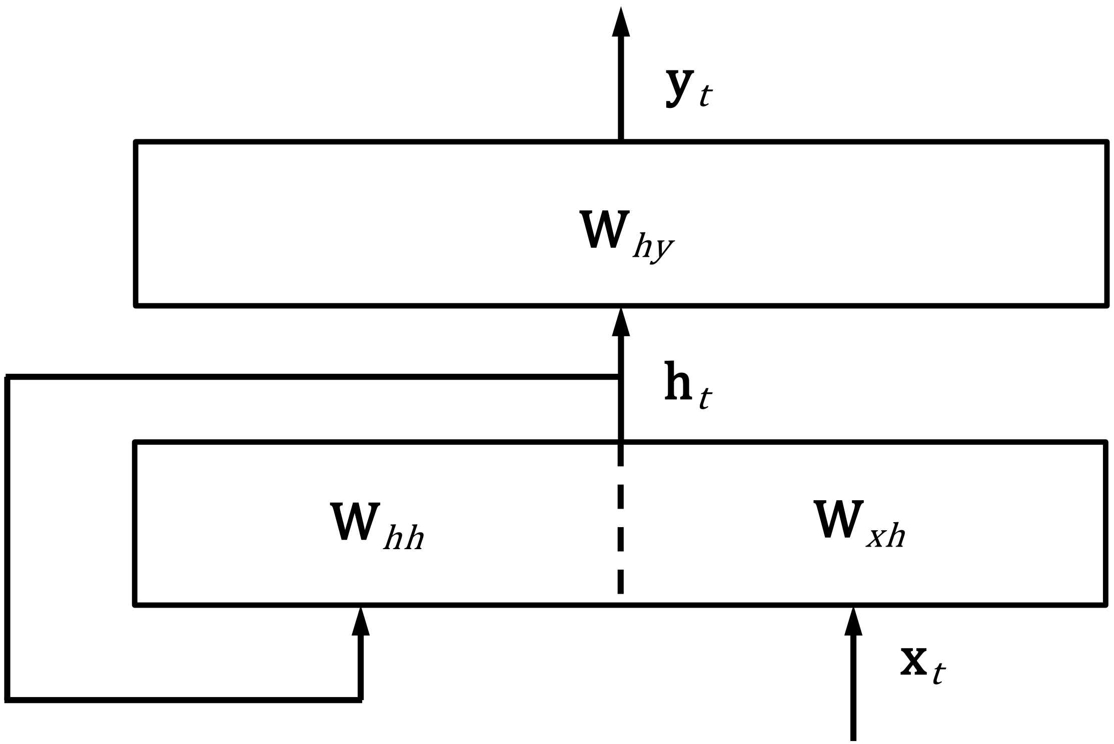
(D. Yu and L. Deng, 2014)

Neural-network language models



(X. He, et al., 2014)

RNN (Recurrent NN) architecture with one recurrent layer



$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t\text{-}1})$$

$$\mathbf{y}_t = \mathrm{softmax}(\mathbf{W}_{hy}\mathbf{h}_t)$$

Bi-directional RNN, LSTM (Long Short-term Memory) RNN, Hierarchical RNN, Deep RNN

LSTM (Long Short-term Memory) RNN

- One of the main advantages of RNN over feedforward NN is that no explicit dependence on a pre-defined context length has to be assumed.
- However, standard gradient-based training algorithms fall short of learning RNN weights, due to vanishing and exploding gradients.
- By replacing the standard recurrent hidden layer with an LSTM layer, the problem can be avoided, while it can be trained with conventional RNN learning algorithms.

LSTM (Long Short-term Memory) RNN

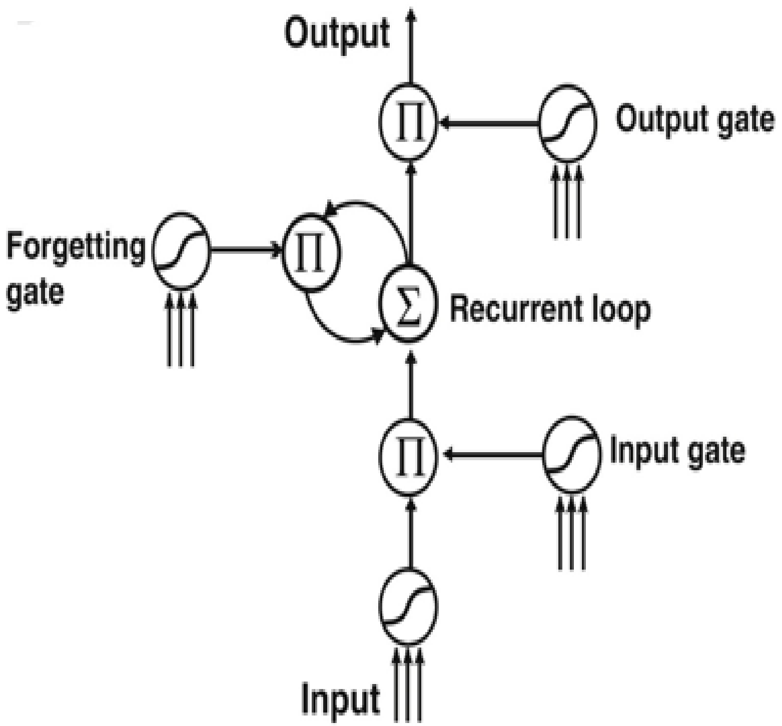
$$\mathbf{i}_t = \sigma\left(\mathbf{W}^{(xi)}\mathbf{x}_t + \mathbf{W}^{(hi)}\mathbf{h}_{t-1} + \mathbf{W}^{(ci)}\mathbf{c}_{t-1} + \mathbf{b}^{(i)}\right)$$

$$\mathbf{f}_t = \sigma\left(\mathbf{W}^{(xf)}\mathbf{x}_t + \mathbf{W}^{(hf)}\mathbf{h}_{t-1} + \mathbf{W}^{(cf)}\mathbf{c}_{t-1} + \mathbf{b}^{(f)}\right)$$

$$\mathbf{c}_t = \mathbf{f}_t \bullet \mathbf{c}_{t-1} + \mathbf{i}_t \bullet \tanh\left(\mathbf{W}^{(xc)}\mathbf{x}_t + \mathbf{W}^{(hc)}\mathbf{h}_{t-1} + \mathbf{b}^{(c)}\right)$$

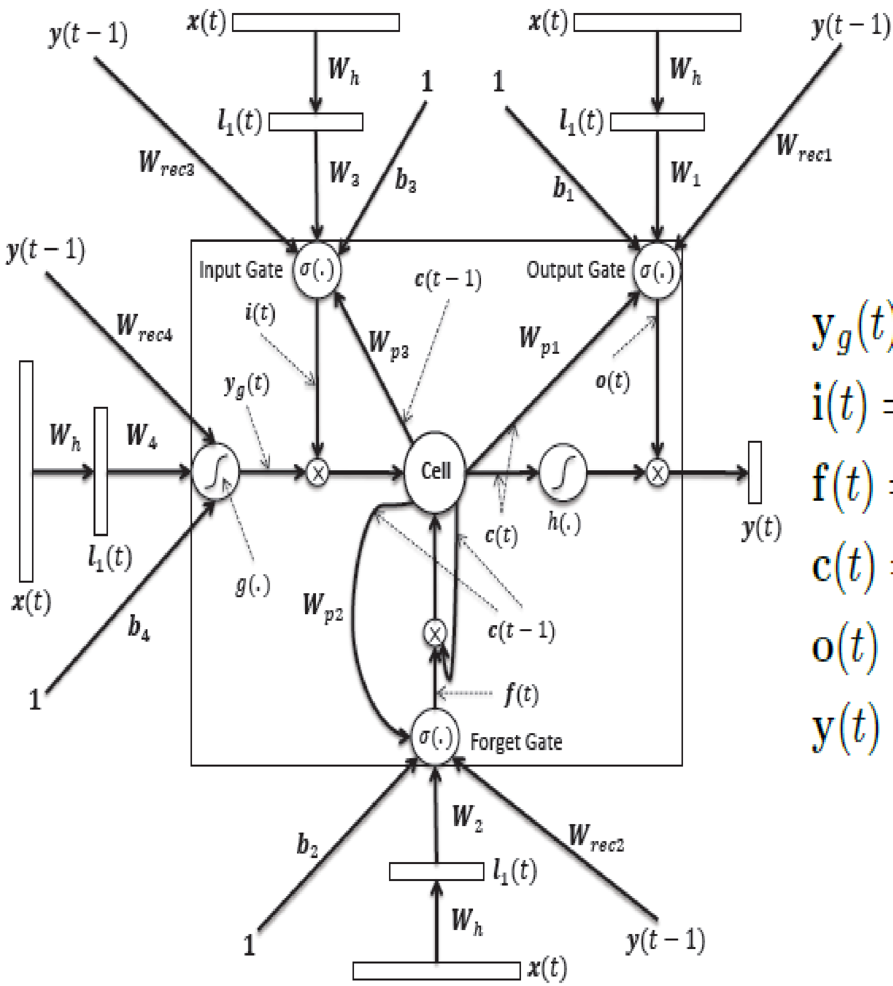
$$\mathbf{o}_t = \sigma\left(\mathbf{W}^{(xo)}\mathbf{x}_t + \mathbf{W}^{(ho)}\mathbf{h}_{t-1} + \mathbf{W}^{(co)}\mathbf{c}_t + \mathbf{b}^{(o)}\right)$$

$$\mathbf{h}_t = \mathbf{o}_t \bullet \tanh\left(\mathbf{c}_t\right),$$



(D. Yu and L. Deng, 2015)

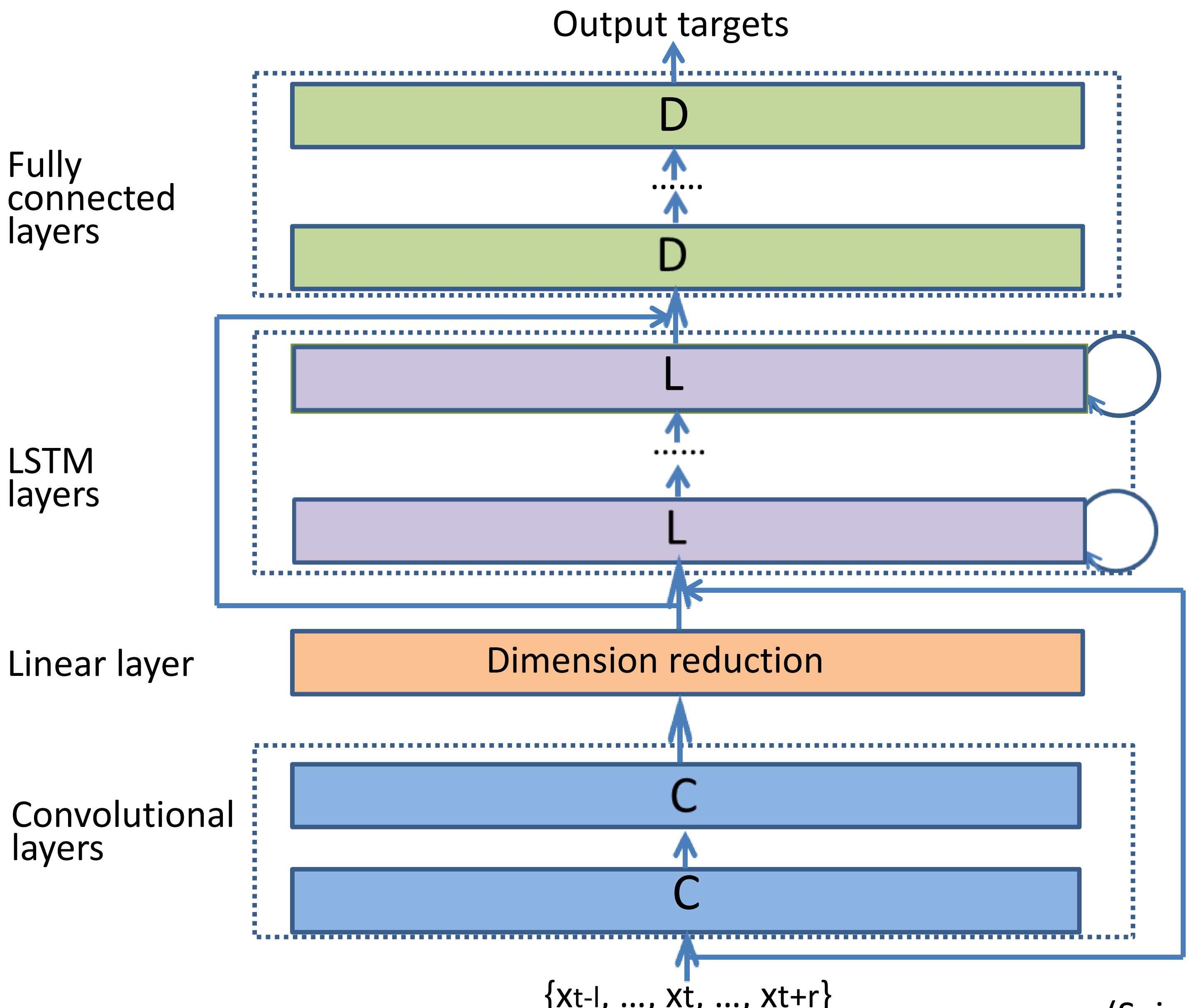
LSTM: for informational retrieval



$$\mathbf{y}_g(t) = g(\mathbf{W}_4\mathbf{l}_1(t) + \mathbf{W}_{\text{rec}4}\mathbf{y}(t-1) + \mathbf{b}_4)$$
$$\mathbf{i}(t) = \sigma(\mathbf{W}_3\mathbf{l}_1(t) + \mathbf{W}_{\text{rec}3}\mathbf{y}(t-1) + \mathbf{W}_{p3}\mathbf{c}(t-1) + \mathbf{b}_3)$$
$$\mathbf{f}(t) = \sigma(\mathbf{W}_2\mathbf{l}_1(t) + \mathbf{W}_{\text{rec}2}\mathbf{y}(t-1) + \mathbf{W}_{p2}\mathbf{c}(t-1) + \mathbf{b}_2)$$
$$\mathbf{c}(t) = \mathbf{f}(t) \circ \mathbf{c}(t-1) + \mathbf{i}(t) \circ \mathbf{y}_g(t)$$
$$\mathbf{o}(t) = \sigma(\mathbf{W}_1\mathbf{l}_1(t) + \mathbf{W}_{\text{rec}1}\mathbf{y}(t-1) + \mathbf{W}_{p1}\mathbf{c}(t) + \mathbf{b}_1)$$
$$\mathbf{y}(t) = \mathbf{o}(t) \circ h(\mathbf{c}(t))$$

(L. Deng, 2015)

CLDNN (Convolutional, LSTM-DNN)



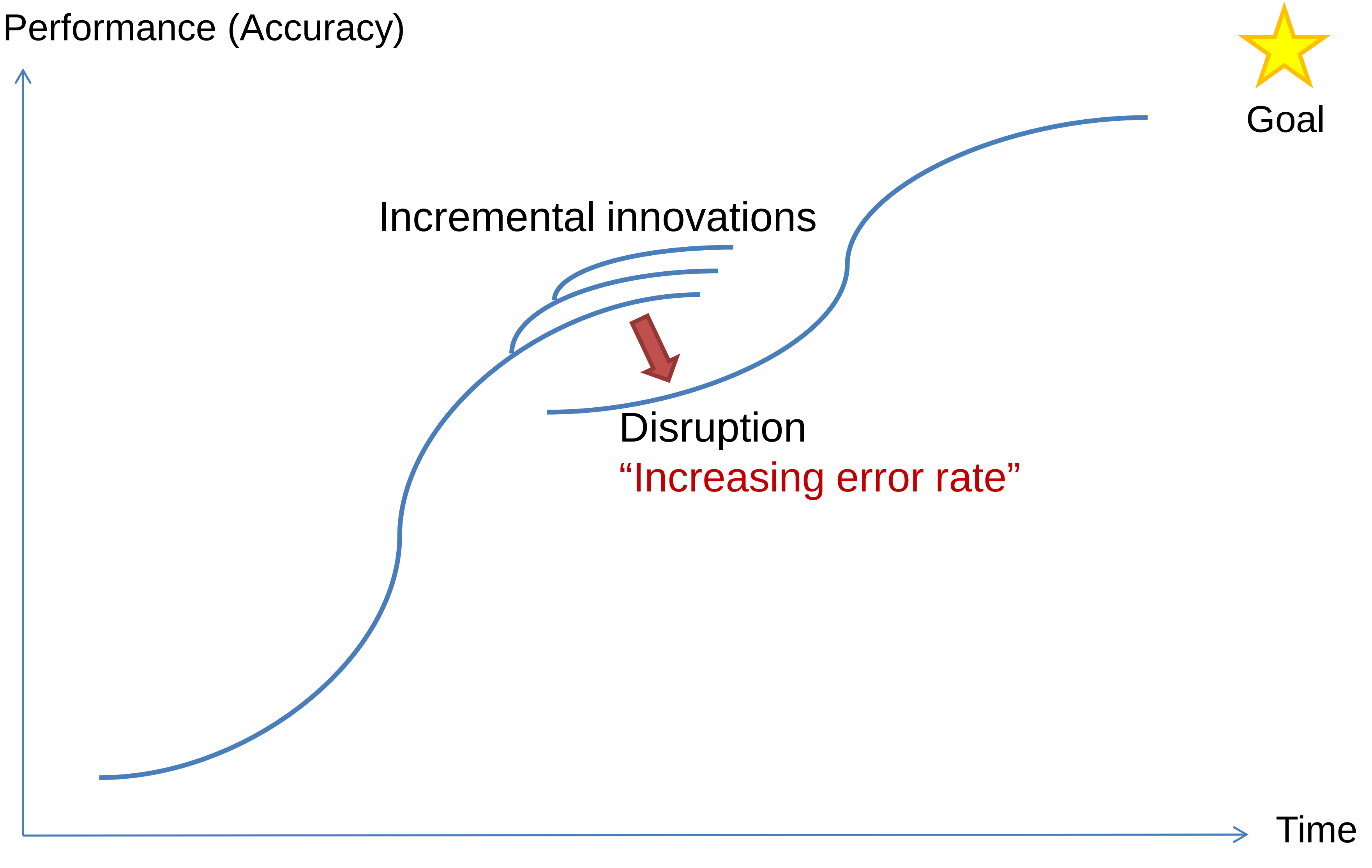
(Sainath et al., 2015)

Performance of different types of language models for an English test data. (Neural network LMs are always interpolated with the large count LM.)

LM	Hidden Layers	Perplexity	Character Error Rate (%)	Word Error Rate (%)
Count-based	-	131.2	7.6	12.4
+Feedforward	100	121.1	7.5	11.8
	600	112.5	7.2	11.5
	2x 100	121.2	7.5	11.9
	2x 600	110.2	7.2	11.3
+RNN	100	121.0	7.5	11.8
	600	108.1	7.0	11.1
+LSTM	100	115.3	7.3	11.7
	600	96.7	6.8	10.8
	2x 100	111.0	7.2	11.4
	2x 600	**92.0**	**6.7**	**10.4**

(Sundermeyer et al., 2015)

Disruptive innovation



Outline

1. Generations of ASR technology
2. Recent success by deep learning (DNN)
3. J. R. Pierce: “Whither speech recognition?”
4. Speech recognition as a *prediction* process
   – Vowel reduction
   – Spectral dynamics and syllable perception
5. Multi-view learning of speech representations
6. Speech recognition by comprehensive knowledge processing
7. Conclusion

Generations of ASR technology

1950 1960 1970 1980 1990 2000 2010

1952 **1G** 1970

Heuristic approaches
(analog filter bank + logic circuits)

1970 **2G** 1980

Pattern matching
(LPC, FFT, DTW)

1980 **3G** 1990

Statistical framework
(HMM, n-gram, neural net)

1990 **3.5G** 2010

Discriminative approaches, machine learning, robust training, adaptation, rich transcription

2010 **4G**

Deep learning (DNN)

Prehistory ASR (1925)

Our research
NTT Labs (+Bell Labs), Tokyo Tech, Toyota Tech. Inst. at Chicago

“Whither speech recognition?”

- Written by J. R. Pierce in 1969
- Stopped Bell Labs from continuing speech recognition research for several years at the beginning of 1970s
- After 45 years, still worthwhile to read

“Speech recognition has glamor. Funds have been available. Results have been less glamorous. … General-purpose speech recognition seems far away. Special-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish.”

J. R. Pierce wrote (1)

William James wrote, in 1899:

“When we listen to a person speaking or read a page of print, much of what we think we see or hear is supplied from our memory. We overlook misprints, imaging the right letters, though we see the wrong ones; and how little we actually hear, when we listen to speech, we realize when we go to a foreign theatre; for there what troubles us is not so much that we cannot understand what the actors say as that we cannot hear their words. The fact is that we hear quite as little under similar conditions at home, only our mind, being fuller of English verbal associations, supplies the requisite material for comprehension upon a much slighter auditory hint”.

. . .

J. R. Pierce wrote (2)

This wisdom is confirmed by various anecdotes. It is said that a native speaker can understand a conversation on a noisy streetcar where a foreigner very fluent in the language cannot. In persistent efforts to understand noisy or indistinct speech, we continually try to guess what the utterance might be, and a conviction as to its content, even a false conviction, is catching. Totally deaf people can understand speech by reading lips, and yet the clues they follow cannot be sufficient for deciphering all phonemes or even all words. A stenotypist can transcribe a stenotype record despite the fact that not all words are represented unambiguously.

. . .

J. R. Pierce wrote (3)

These considerations lead us to believe that a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English.

. . .

Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers.

. . .

A lot of money and time are spent. No simple, clear, sure knowledge is gained. The work has been an experience, not an experiment.

. . .

J. R. Pierce wrote (4)

The arguments given earlier may lead us to believe that performance will continue to be very limited unless the recognizing device understands what is being said with something of the facility of a native speaker (that is, better than a foreigner who is fluent in the language).

… it would seem appropriate that before embarking upon such work, the worker should candidly ask and answer the following questions:

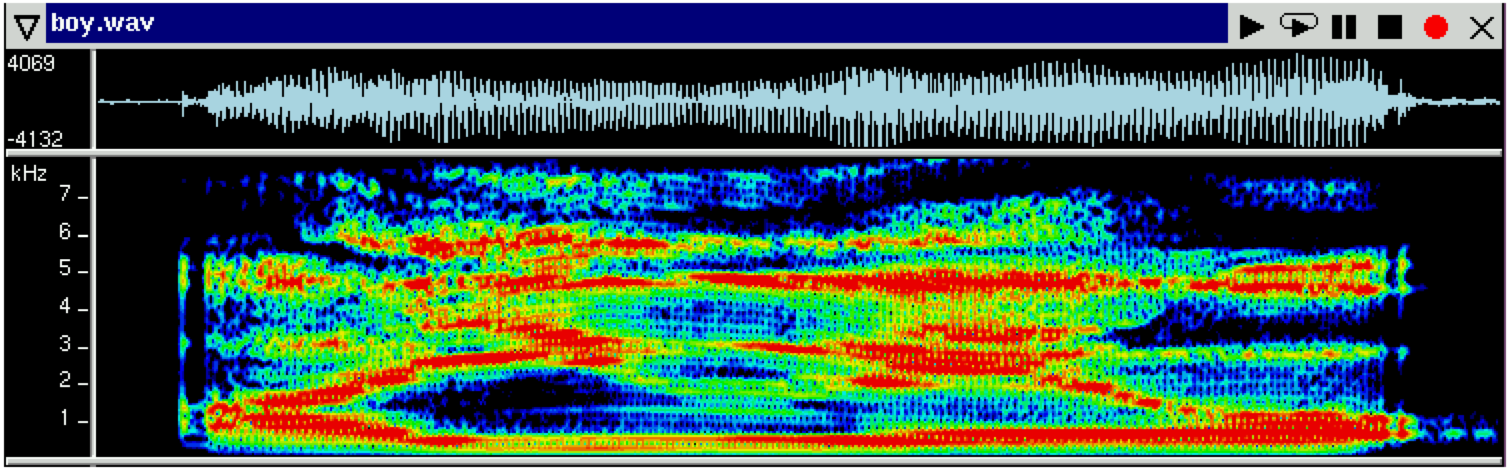
- Why am I working in this field?
- What particular thing do I hope to accomplish?
- Why is it worthwhile?
- Am I likely to succeed?
- How will I know whether or not I have succeeded?
- Where will success take or leave me?

Outline

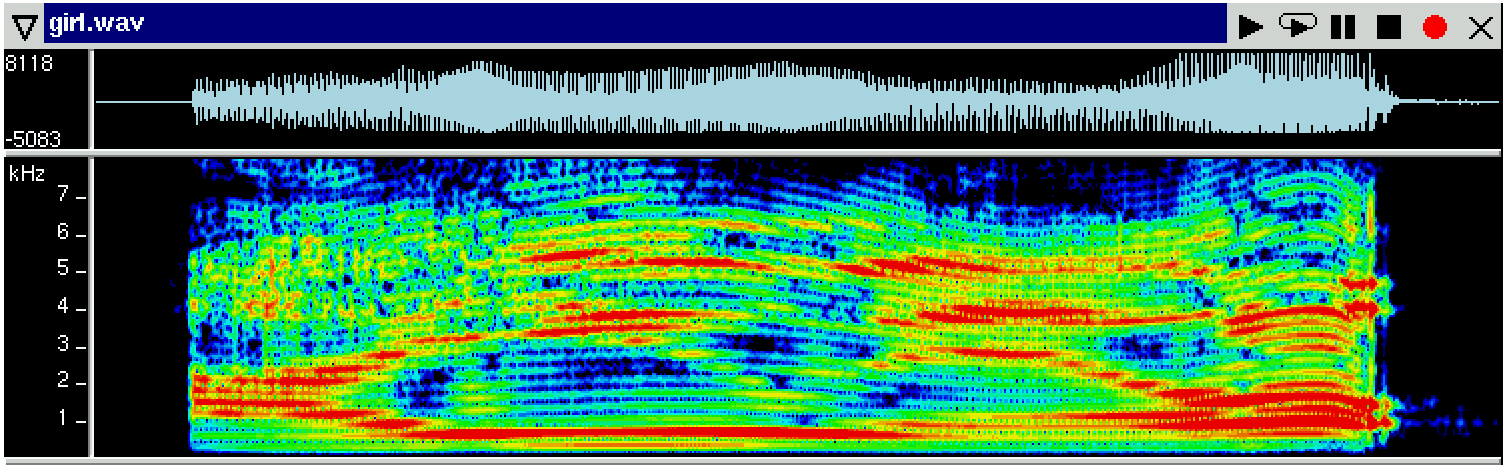
1. Generations of ASR technology
2. Recent success by deep learning (DNN)
3. J. R. Pierce: “Whither speech recognition?”
4. Speech recognition as a *prediction* process
   – Vowel reduction
   – Spectral dynamics and syllable perception
5. Multi-view learning of speech representations
6. Speech recognition by comprehensive knowledge processing
7. Conclusion

Spectrograms of /aiueo/ in Japanese

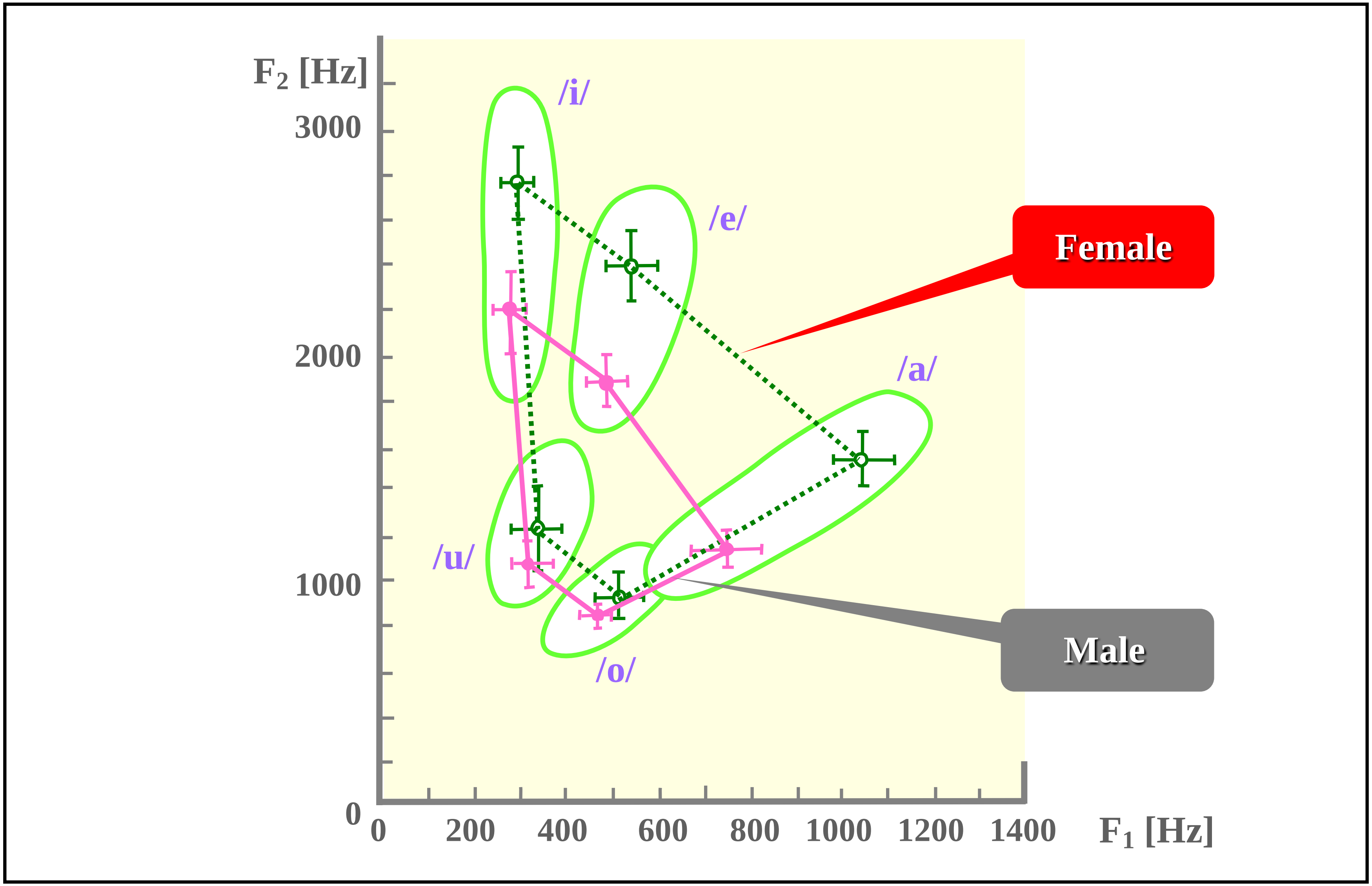
Boy



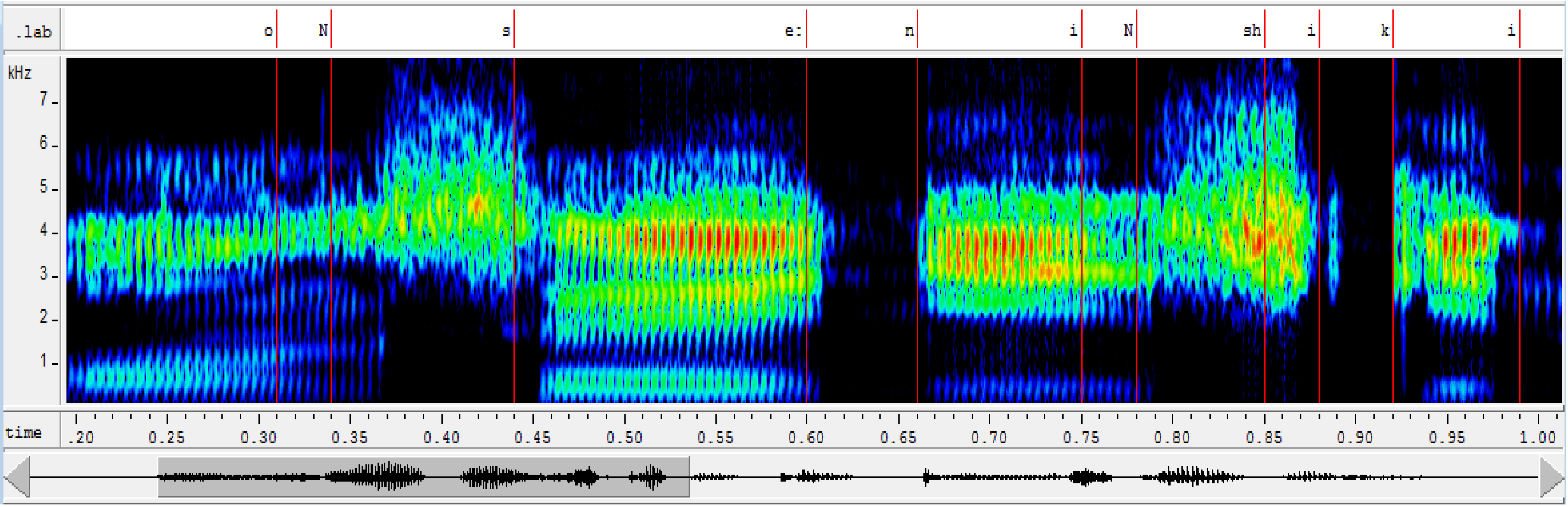
Girl



Scatter diagram of formant frequencies of five Japanese vowels uttered by 60 speakers (30 males and 30 females) in the F_1-F_2 plane



Automatic segmentation of /oNse:niNsh(i)ki/ (*speech recognition* in Japanese) by triphone HMMs

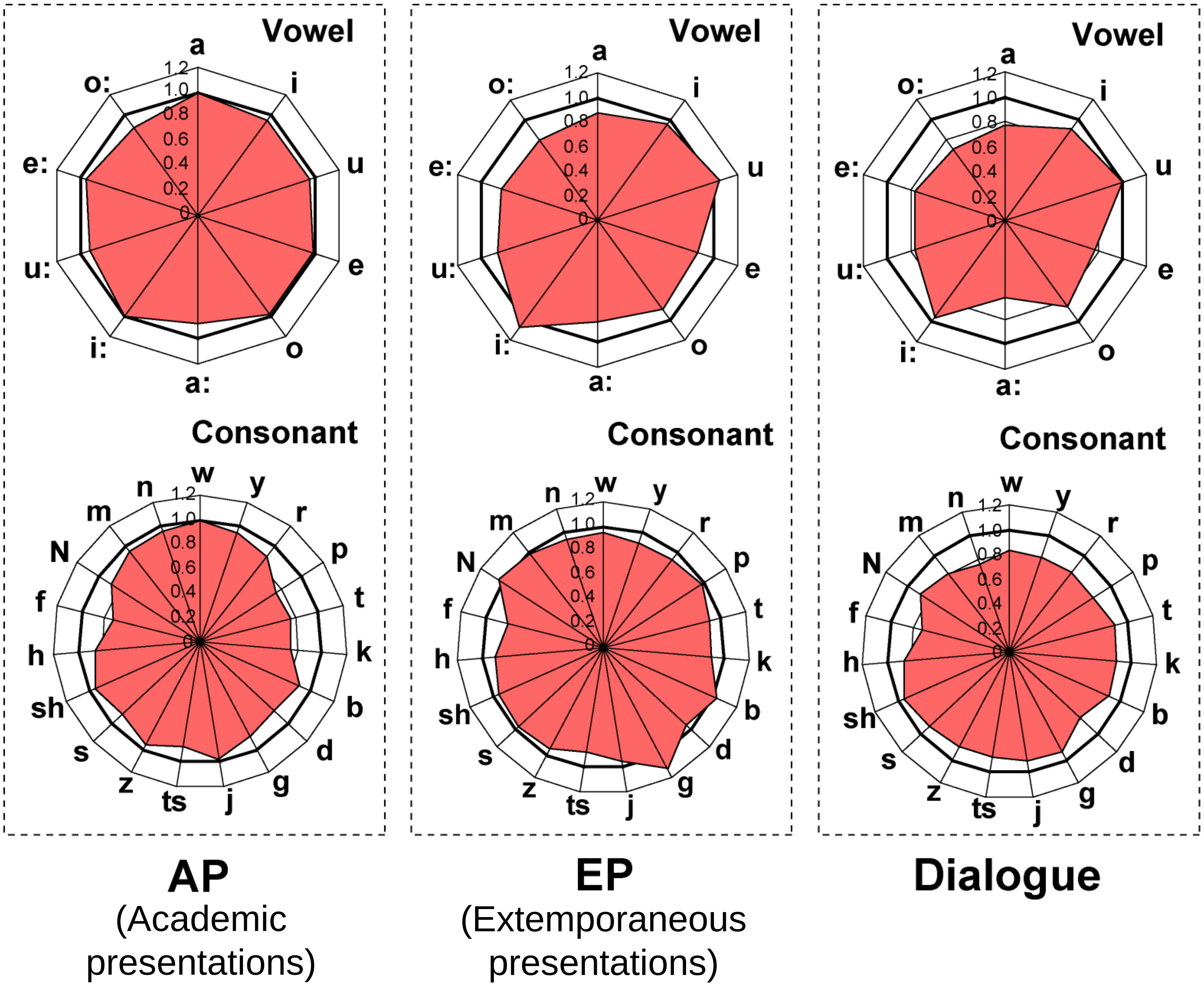


Speech recognition is a prediction process.

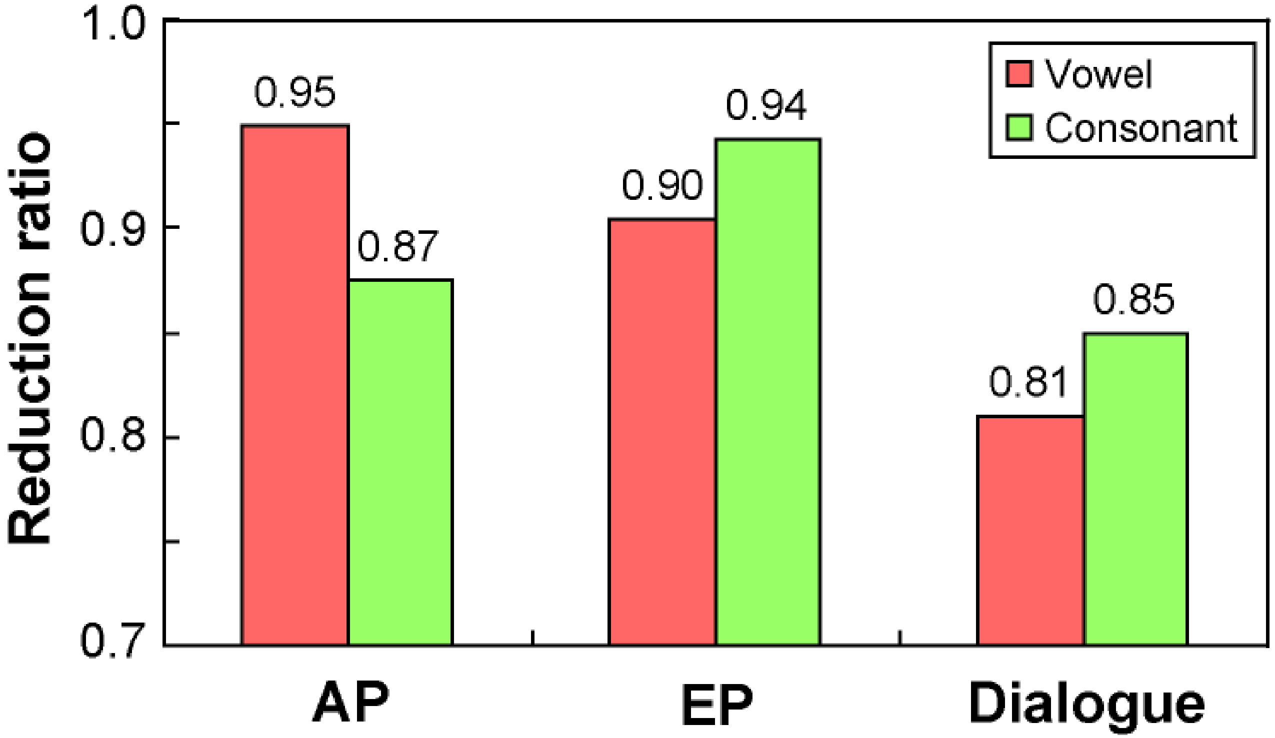
Challenges

- Speech is not a process of reading written text
- Analysis of human speech perception
- Understanding and extracting meanings and intensions
- *Prediction* by spectral dynamics
- *Prediction* by context (topics, speakers, noises, etc.)
- Higher-order language models (prosody and long-distance dependency)
- Unknown (new) words
- Big data for spontaneous speech
- Machine learning for combining various knowledge sources using big data and unsupervised/semi-supervised/lightly-supervised training

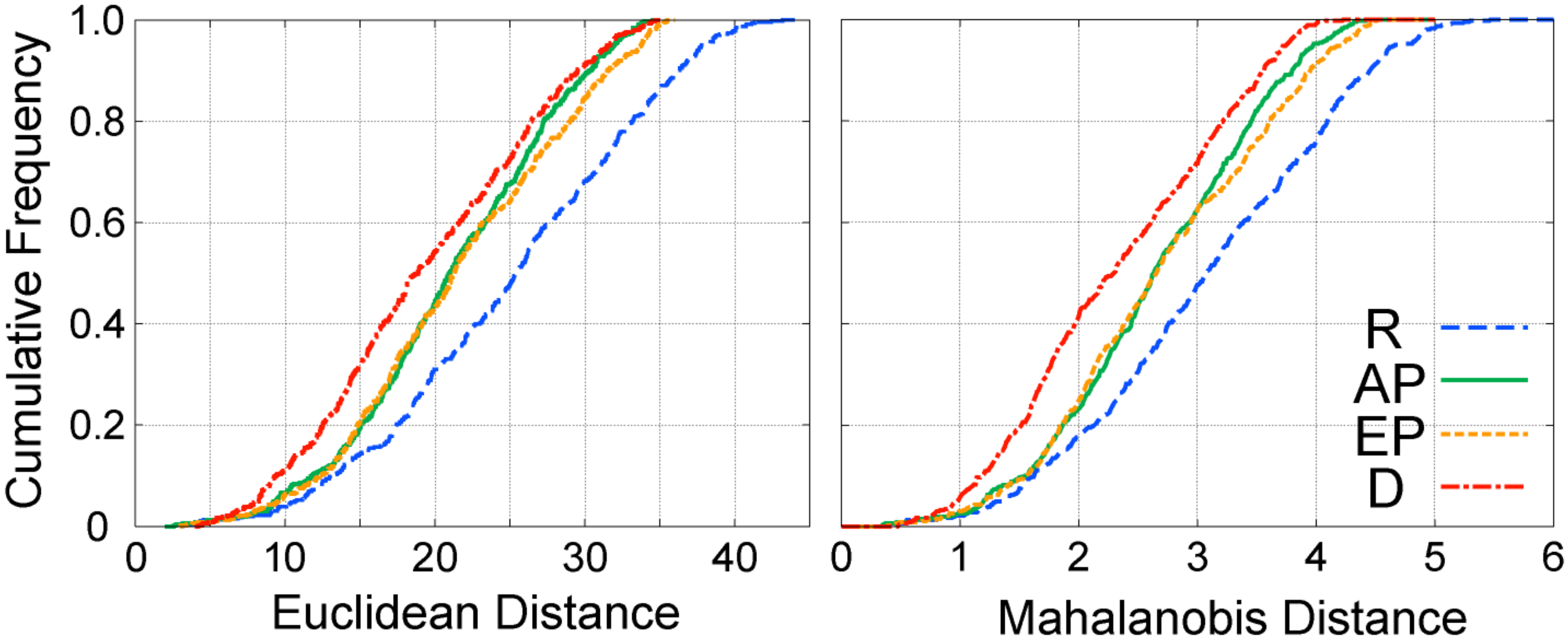
Reduction ratio of the vector norm from phoneme center to each phoneme in spontaneous speech to that in read speech



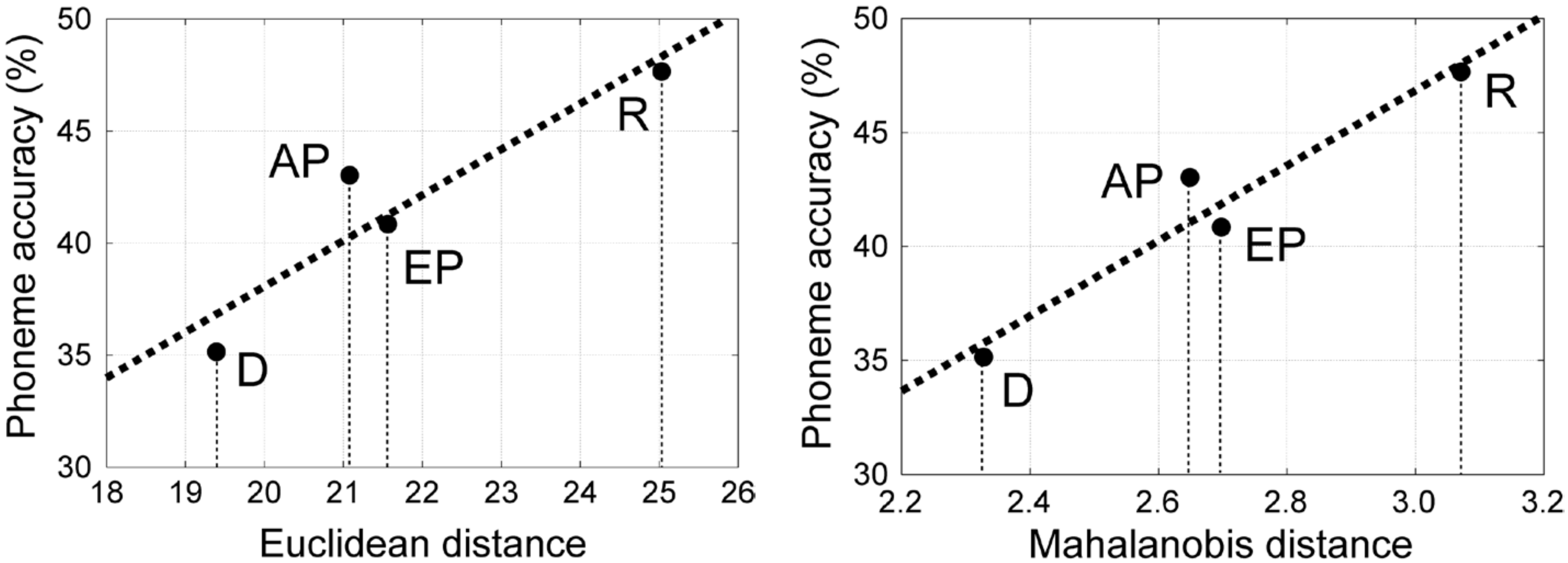
Mean reduction ratio of vowels and consonants for each speaking style



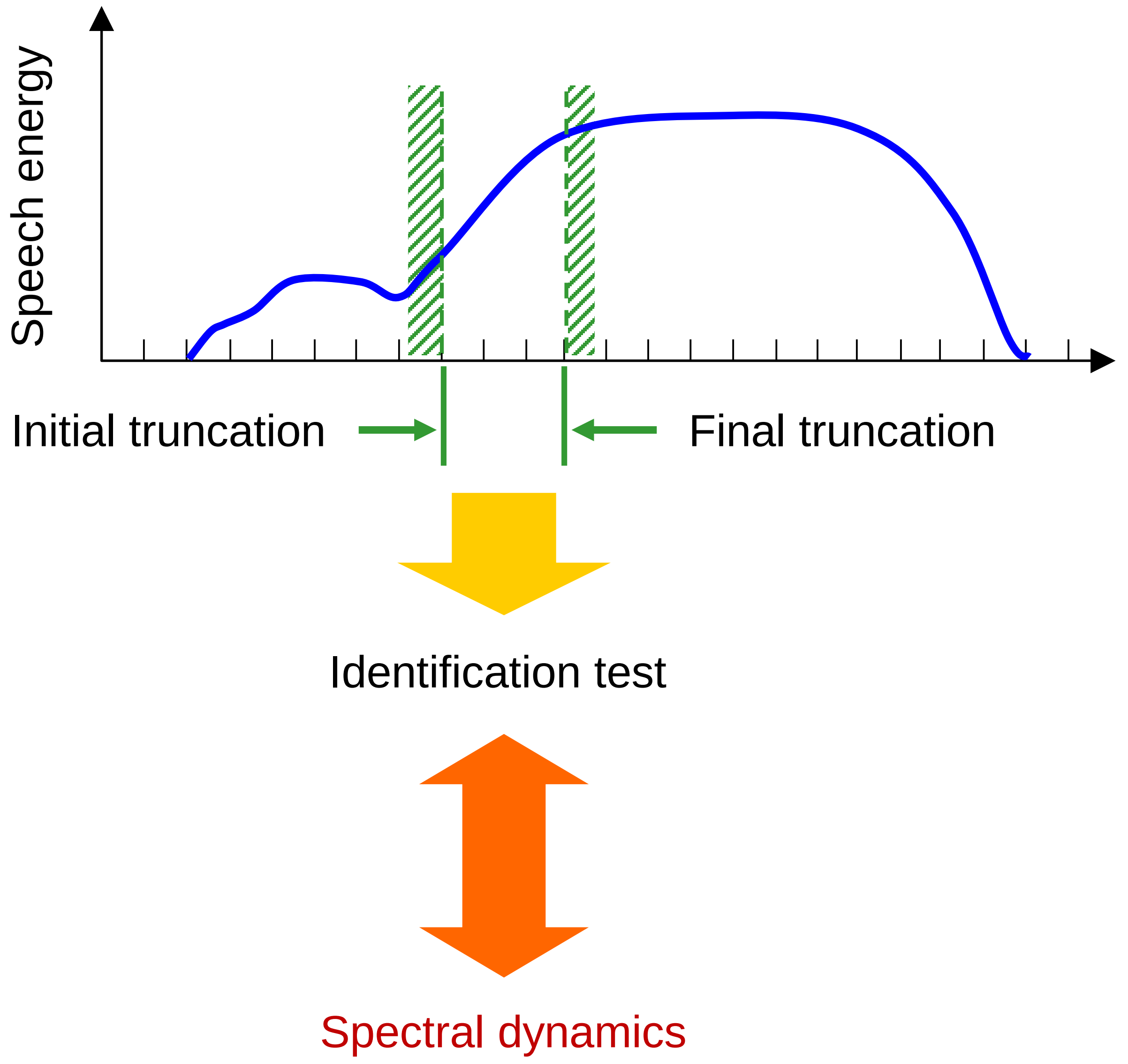
Distribution of distances between phonemes
(R: read speech, AP: academic presentations, EP: extemporaneous presentations, D: dialogue)



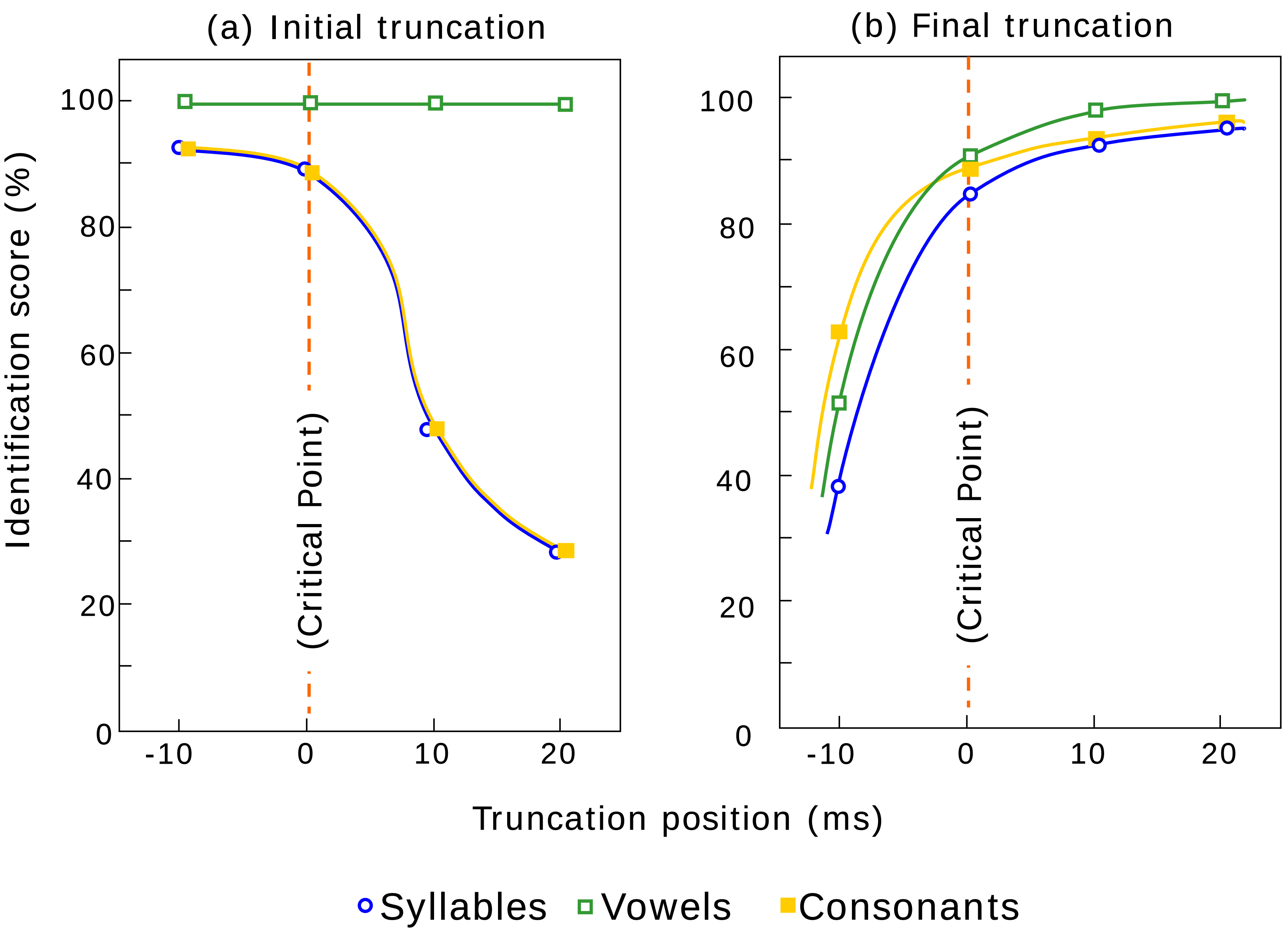
Relationship between mean phoneme distance and phoneme recognition accuracy



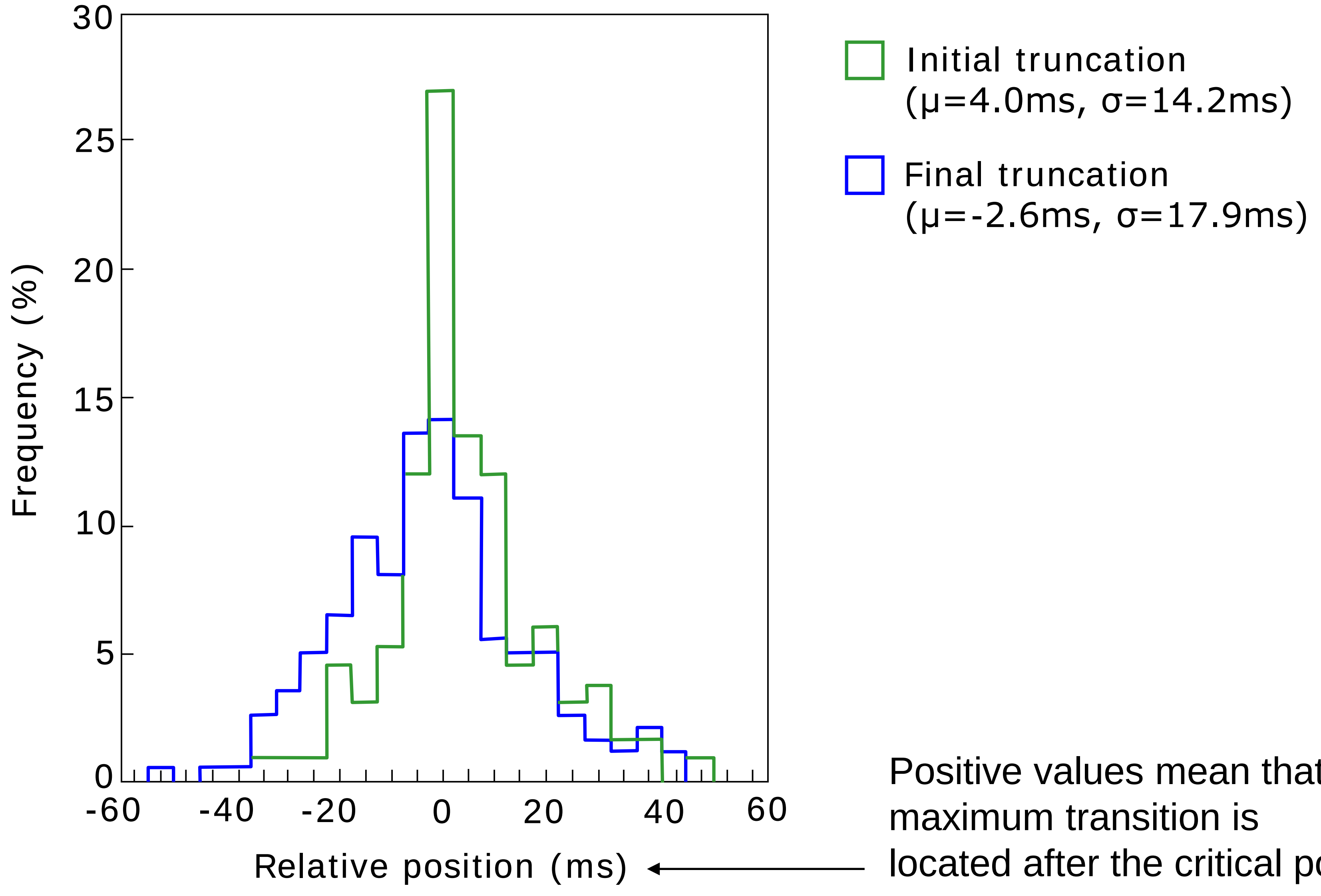
Analysis of relationships between spectral dynamics and syllable perception



Relationship between truncated CV syllable identification scores and truncation position relative to the perceptual critical point

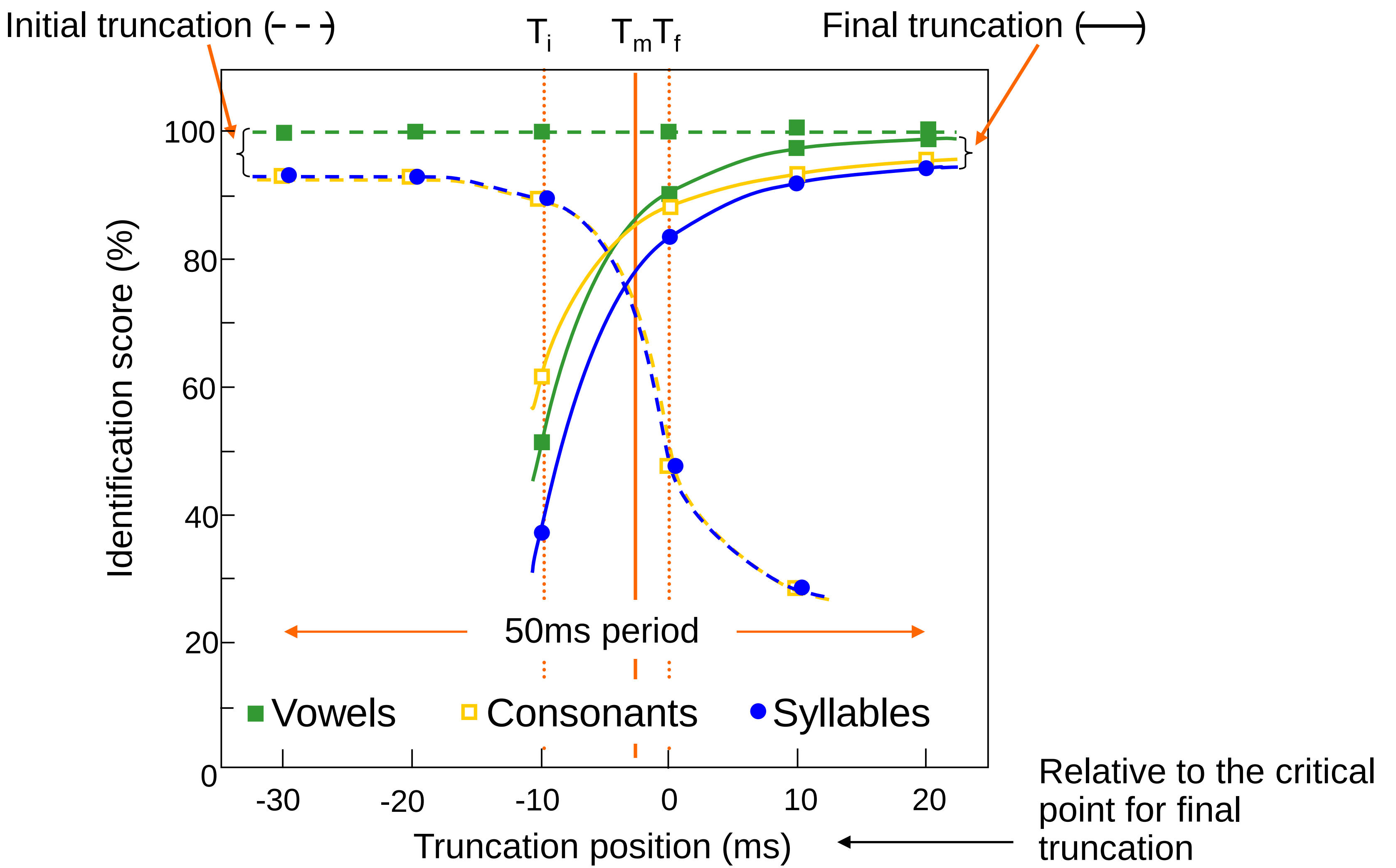


Distribution of the difference between the perceptual critical point and the maximum spectral transition position for all 100 syllables



Positive values mean that the maximum transition is located after the critical point

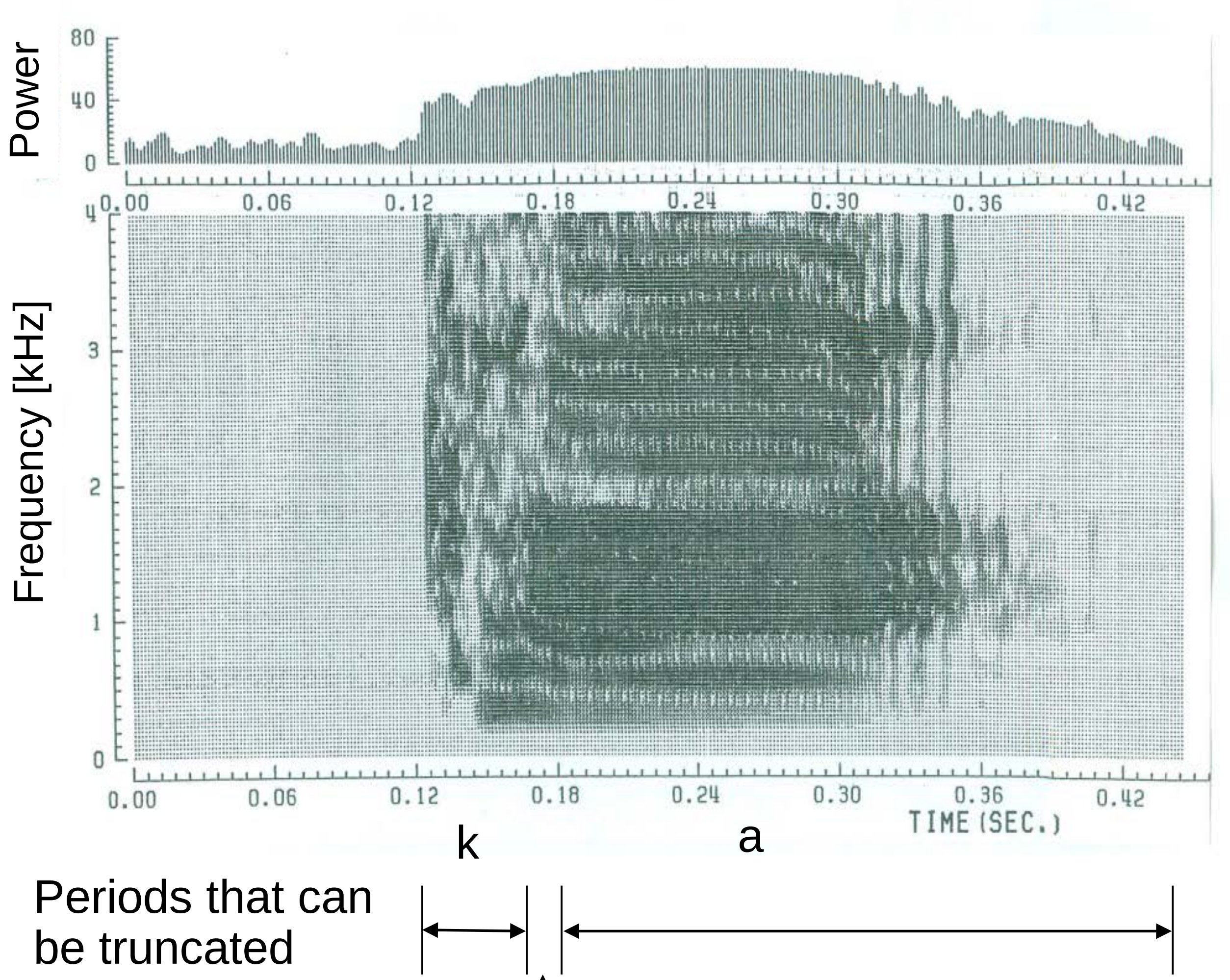
Relationship between truncation position and identification scores for the truncated CV syllables



T_i, T_f : Perceptual critical point for initial & final truncation

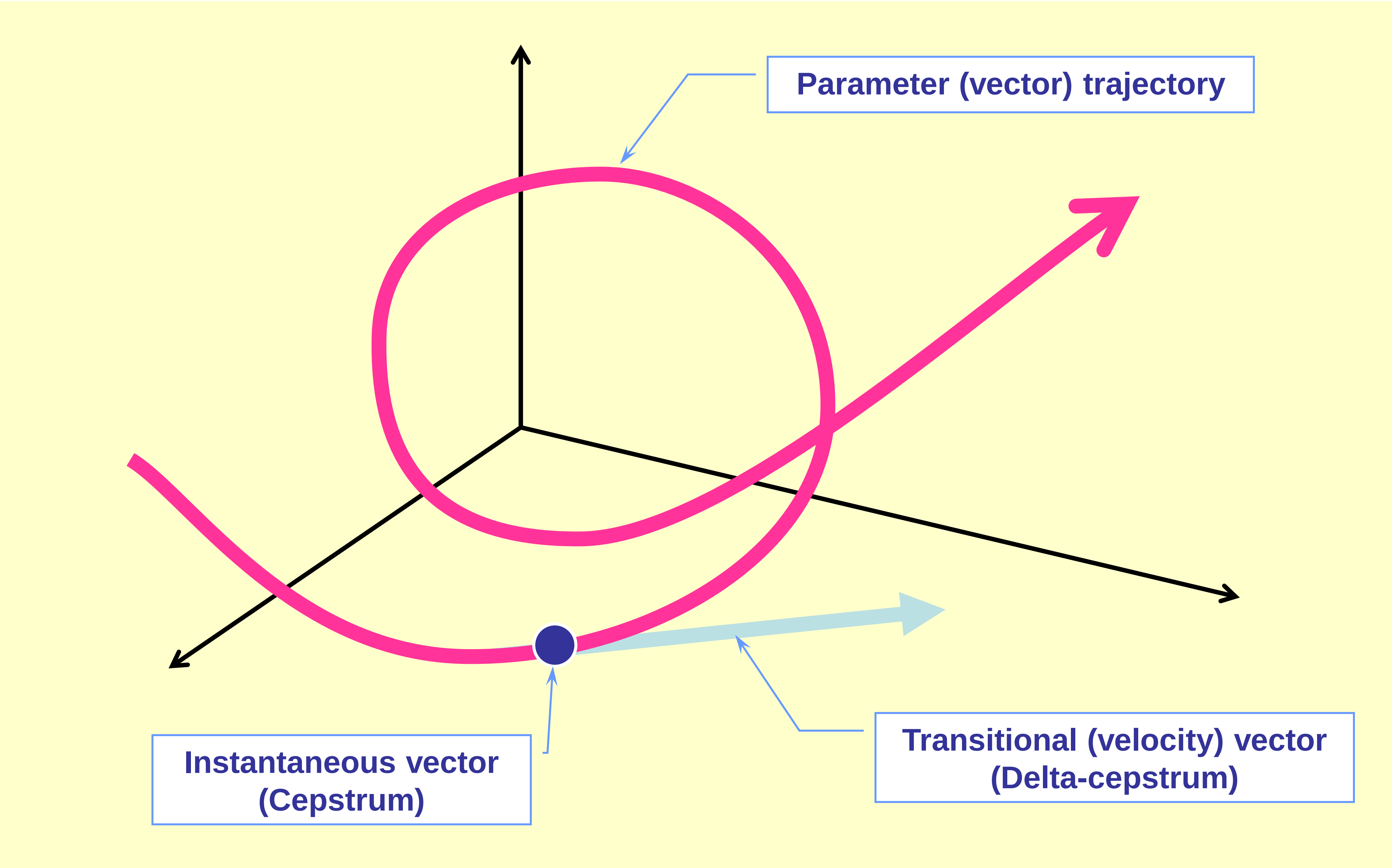
T_m : Maximum spectral transition position

Role of spectral transition for speech perception

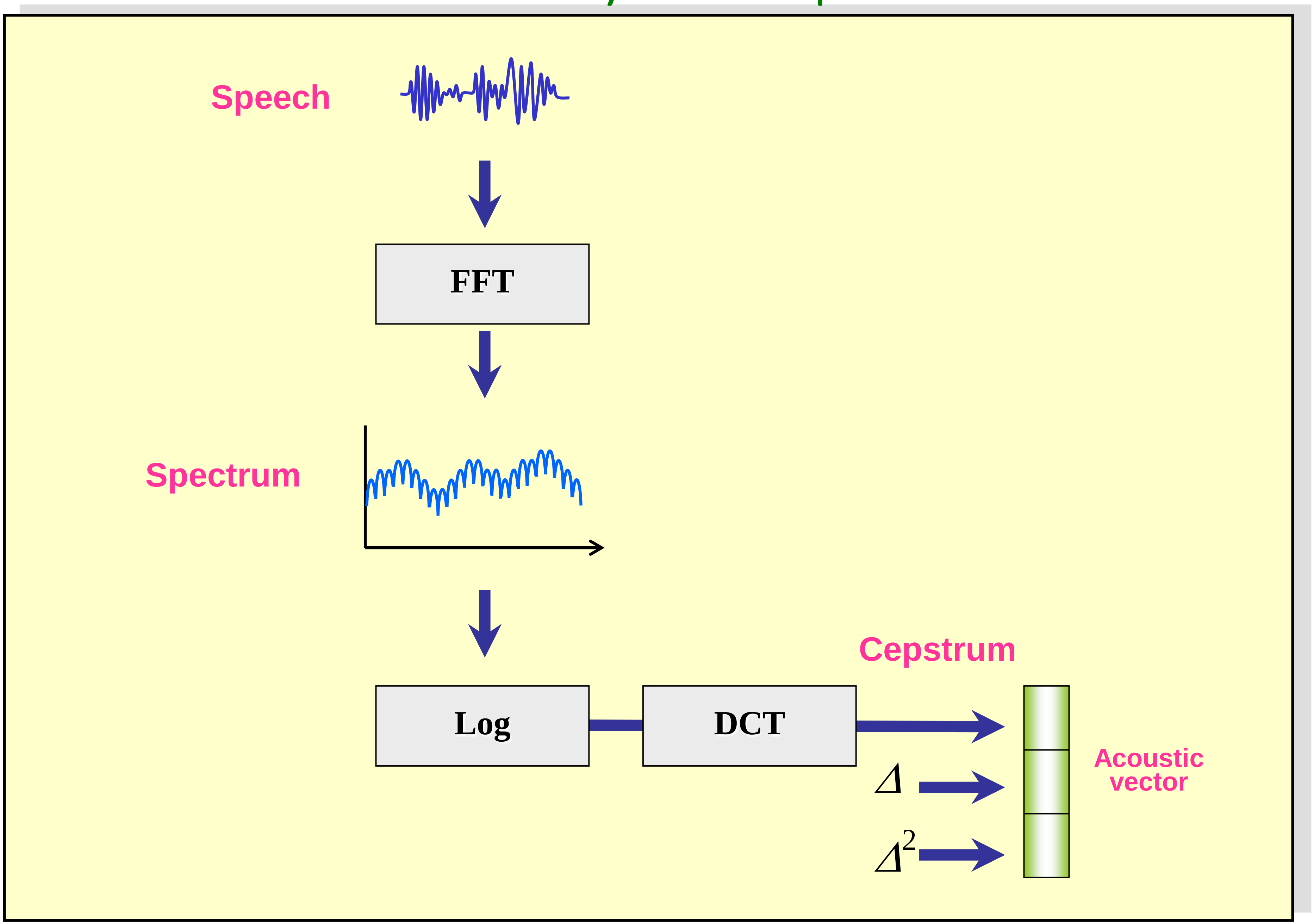


Maximum spectral change period: essential for syllable perception

Cepstrum and delta-cepstrum coefficients



Instantaneous and dynamic cepstrum features



Outline

1. Generations of ASR technology
2. Recent success by deep learning (DNN)
3. J. R. Pierce: “Whither speech recognition?”
4. Speech recognition as a *prediction* process
   – Vowel reduction
   – Spectral dynamics and syllable perception
5. Multi-view learning of speech representations
6. Speech recognition by comprehensive knowledge processing
7. Conclusion

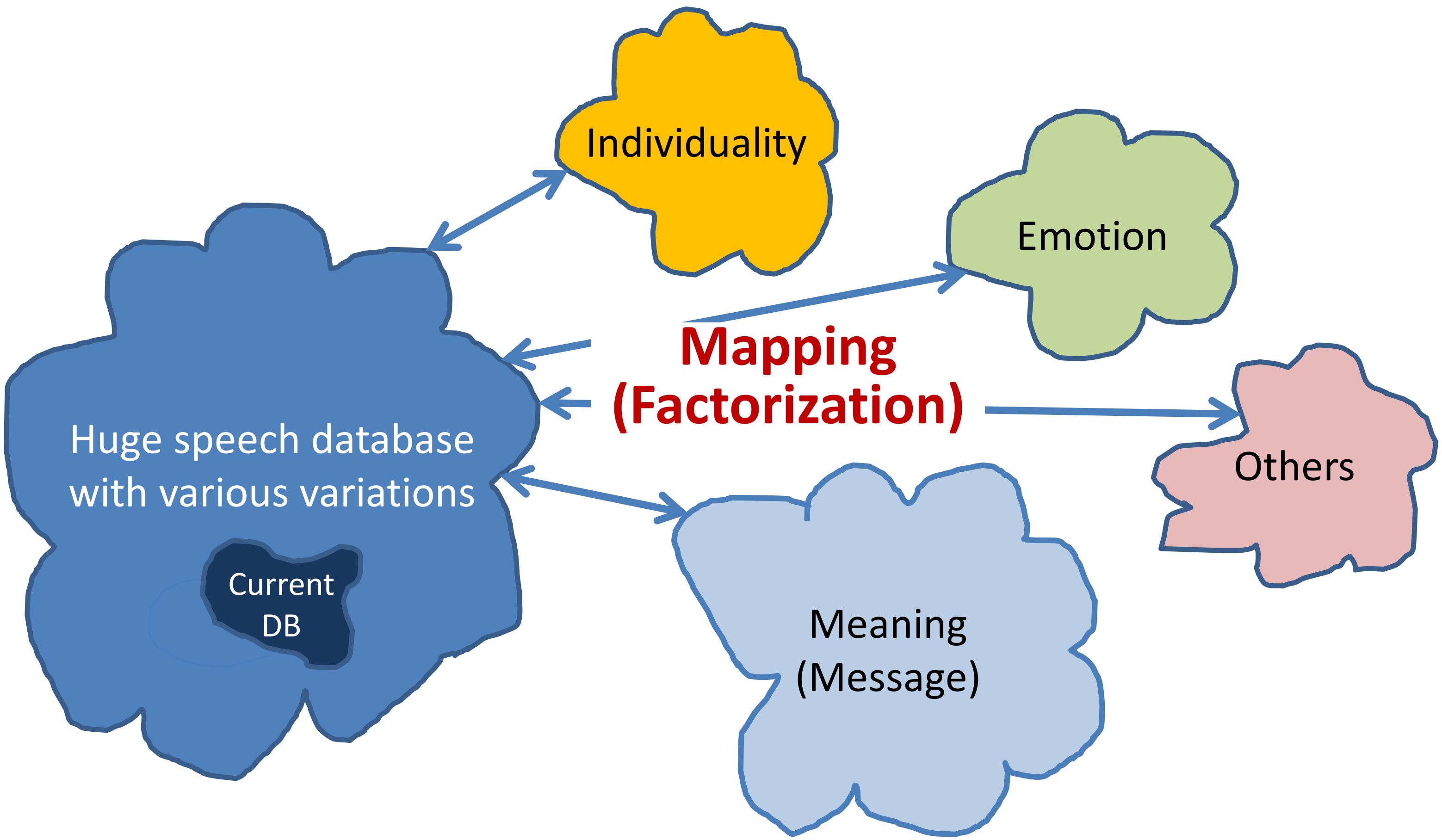
Multi-view learning of speech representations (Karen Livescu, et al.)

1. Multi-view data
2. Multi-view representation learning
3. Canonical correlation analysis (CCA)
4. Kernel canonical correlation analysis (KCCA)
5. Deep canonical correlation analysis (DCCA)
6. Speech recognition experiments with XRMB data

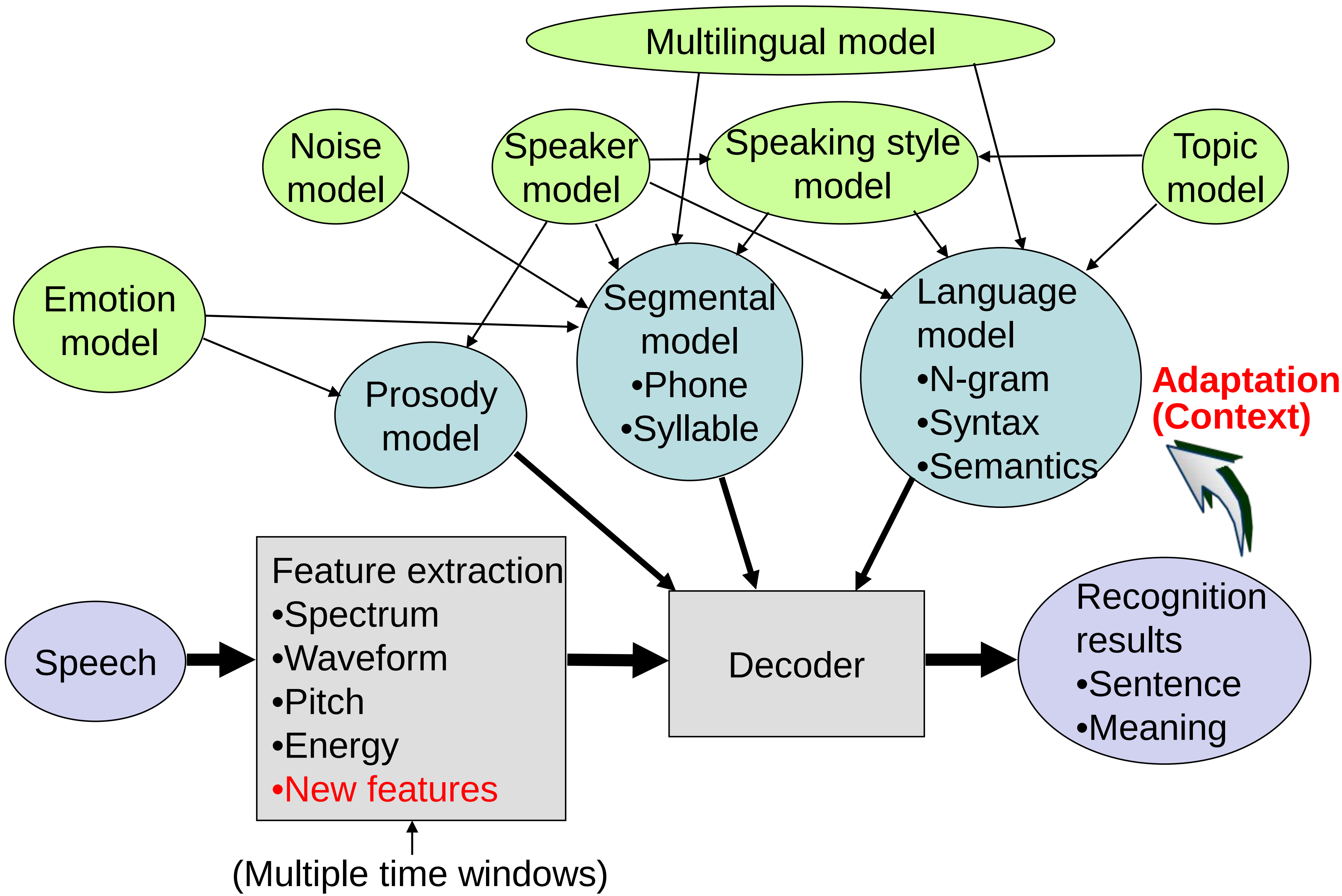
Outline

1. Generations of ASR technology
2. Recent success by deep learning (DNN)
3. J. R. Pierce: “Whither speech recognition?”
4. Speech recognition as a *prediction* process
   – Vowel reduction
   – Spectral dynamics and syllable perception
5. Multi-view learning of speech representations
6. Speech recognition by comprehensive knowledge processing
7. Conclusion

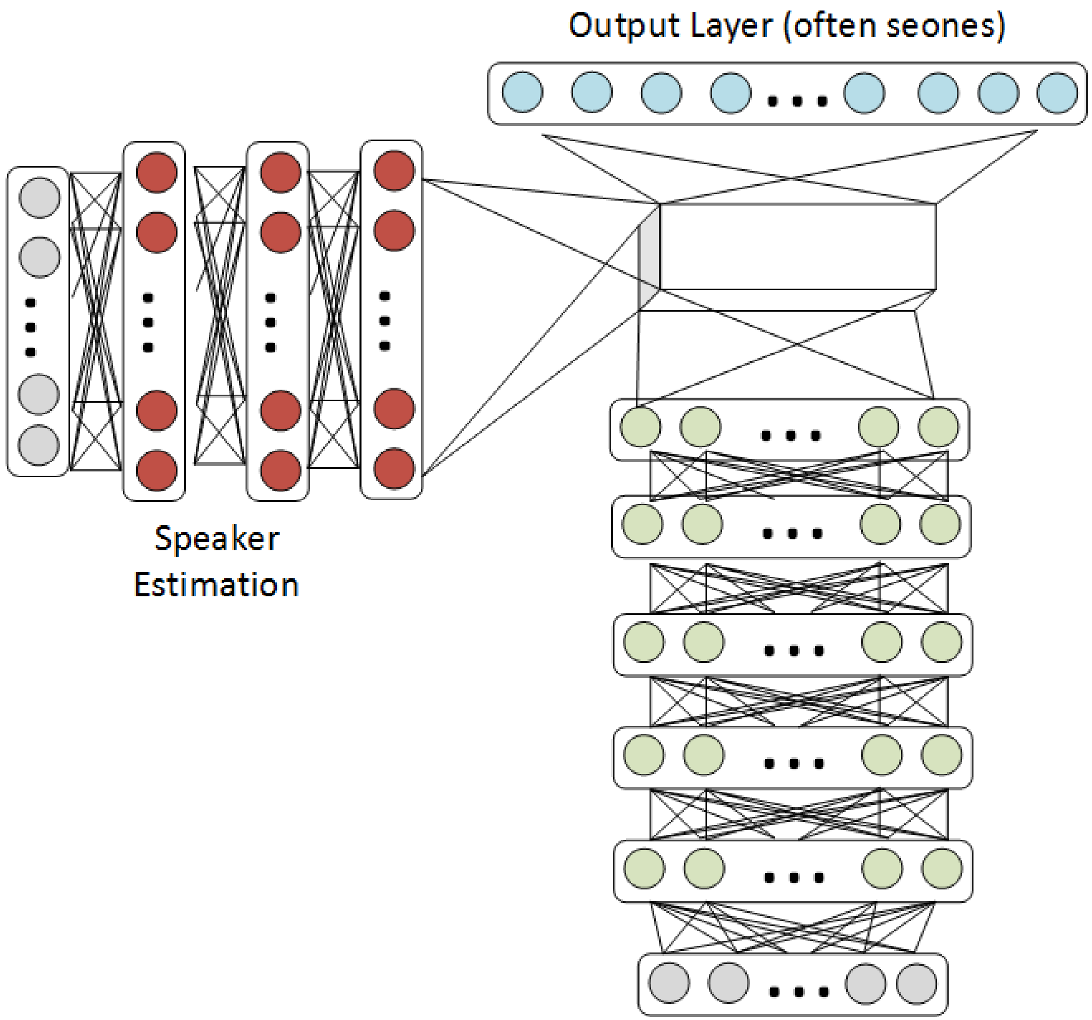
Data-intensive ("Big data") ASR



Next-generation ASR by comprehensive knowledge processing

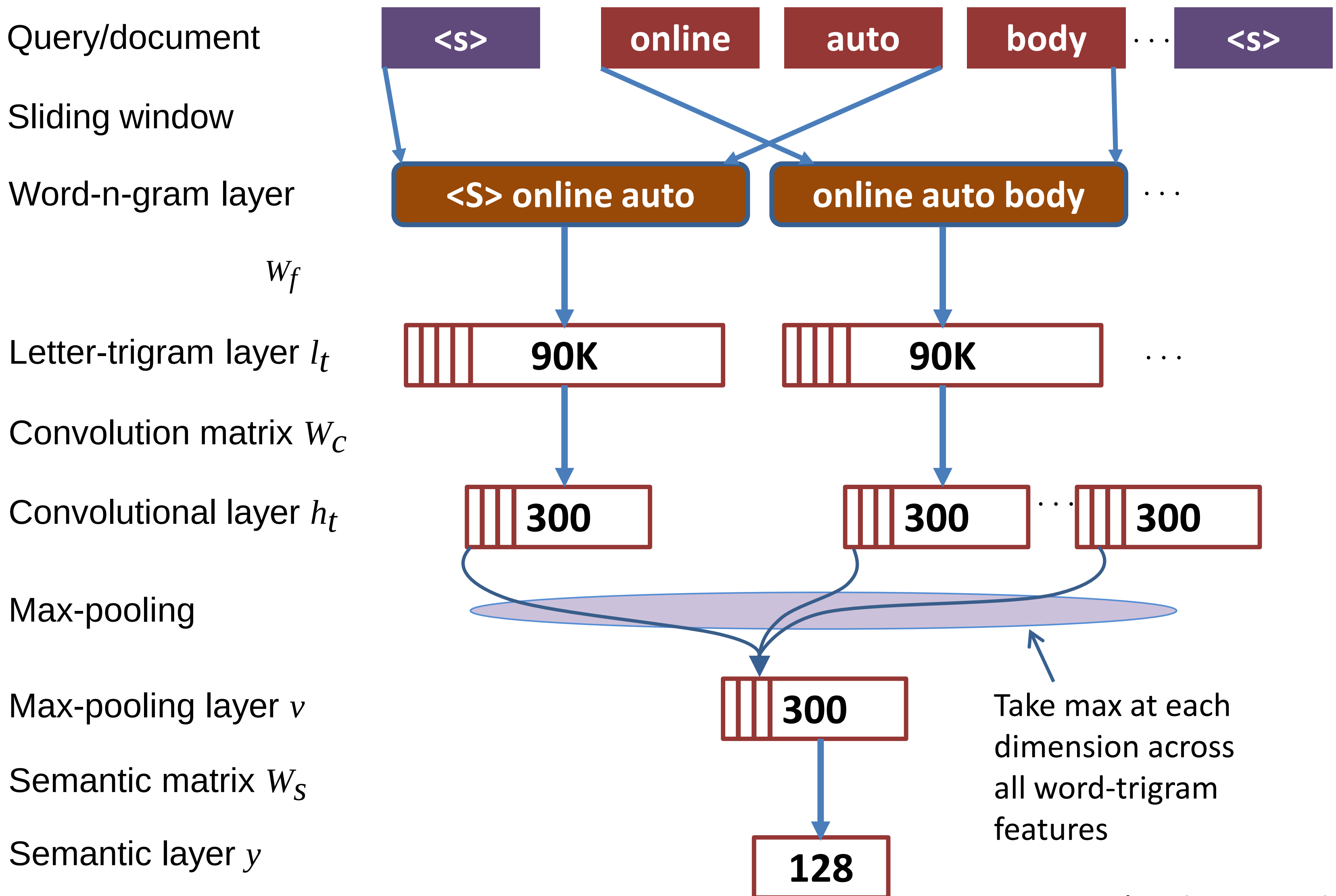


Disjoint factorized DNN model for speaker adaptation



(D. Yu and L. Deng, 2014)

CLSM (Convolutional Latent Semantic Model) for topic extraction



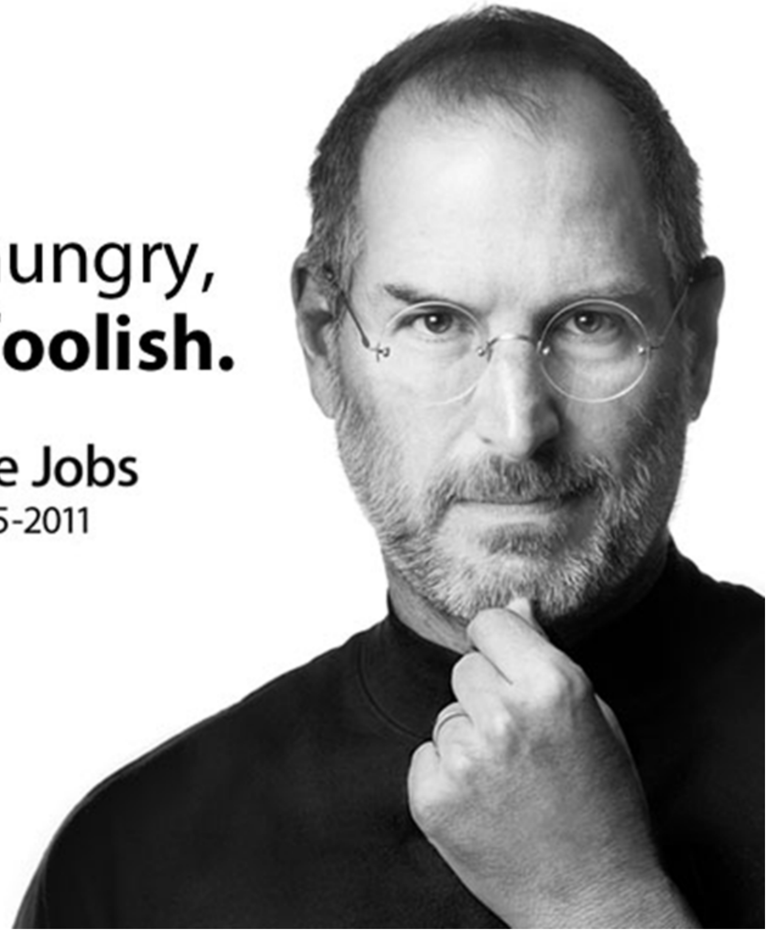
(Y. Shen, et al., 2014)

Word embedding

- Most neural LMs for ASR start with a “one-hot” embedding for each word and learn a embedding as part of the LM training.
- The LM training can start with inputs that are themselves semantic embeddings learned on an external corpus.
- Perplexity of a LSTM language model was reduced by using continuous distributed representations of words trained with a skip-gram method on a big corpora as the input instead of traditional “one-hot” coding. This method has potential to learn new words. (D. Soutner, et al., 2014)

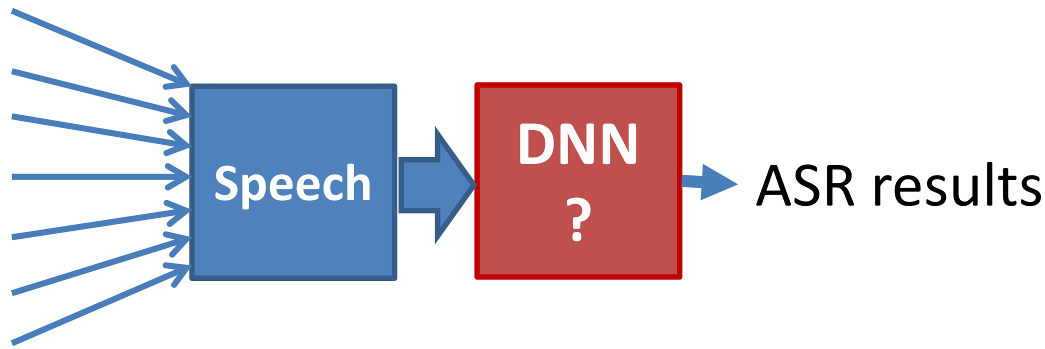
Stay hungry,
Stay foolish.

Steve Jobs
1955-2011



Deep learning to deep thinking

- Don’t be foolish
  - Combinatorial explosion
    - Speaker
    - Dialect
    - Speaking style (task)
    - Emotion
    - Microphone
    - Background noise
    - Reverberation, etc.



  - Time sequence processing
- Think deeply
  - *Prediction* and *knowledge processing* (top-down and bottom-up process)
  - AGI (Artificial General Intelligence): “strong AI”
  - BICA (Biologically Inspired Cognitive Architecture)

Outline

1. Generations of ASR technology
2. Recent success by deep learning (DNN)
3. J. R. Pierce: "Whither speech recognition?"
4. Speech recognition as a *prediction* process
   – Vowel reduction
   – Spectral dynamics and syllable perception
5. Multi-view learning of speech representations
6. Speech recognition by comprehensive knowledge processing
7. Conclusion

Conclusion

- Automatic speech recognition (ASR) technology has made significant progress with the help of ML (machine learning) and computer technology.
- DNNs using "deep learning" has significantly raised the performance.
- We still have many challenges that cannot be solved simply by relying on the current technology.
- We need to deeply think about and model how human beings are *predicting* speech by implementing various knowledge sources in ASR systems using advanced ML techniques (top-down and bottom-up processes).
- How to create and use big speech data, and utilize various knowledge sources in a flexible way?
- How to model and process meanings/semantic understanding?
- Active learning, and unsupervised, semi-supervised or lightly-supervised training/adaptation technologies are crucial.