# Automatic Speech Recognition: Trials, Tribulations and Triumphs

*Sadaoki Furui*

Professor Emeritus

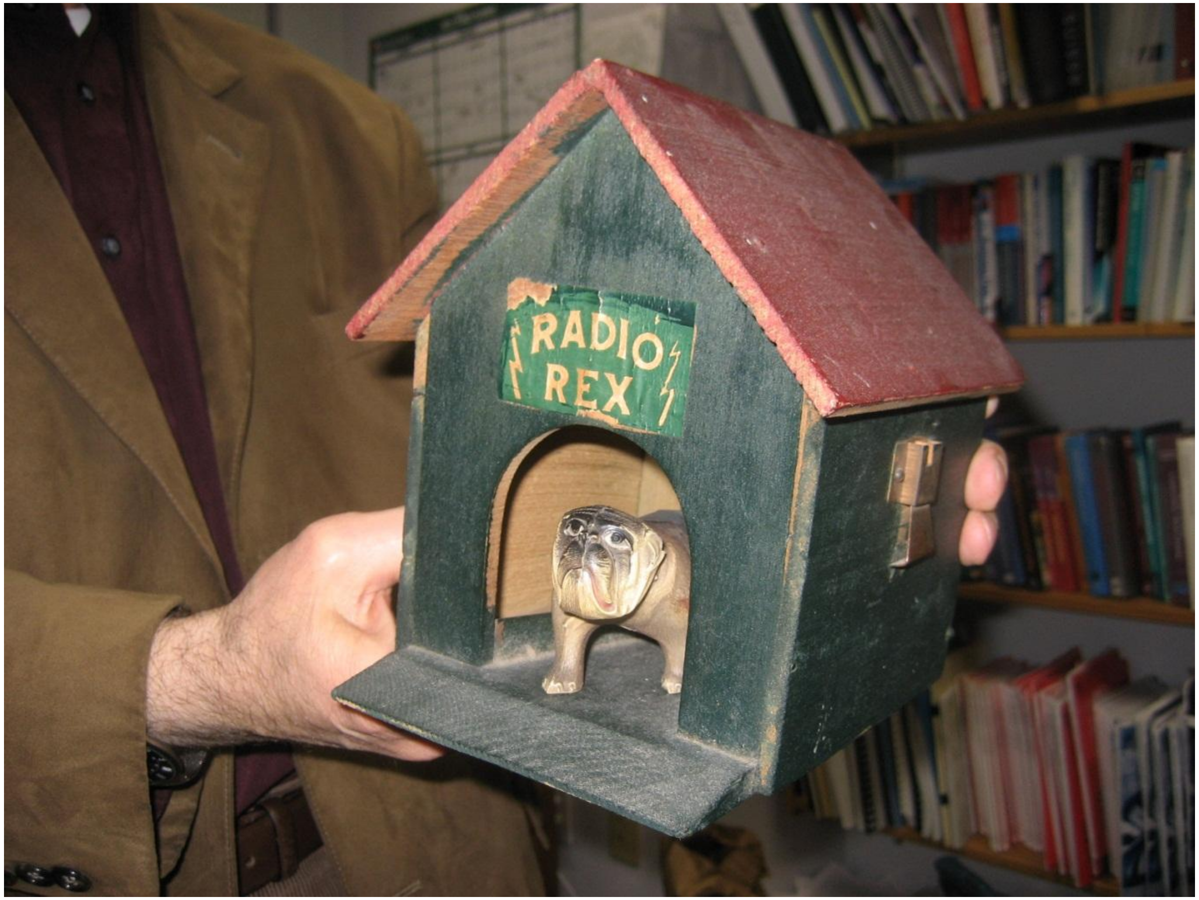
Tokyo Institute of Technology

furui@cs.titech.ac.jp

# Outline

- Progress of automatic speech recognition (ASR) technology
  - 4 generations
  - Synchronization with computer technology
  - Recent activities at Tokyo Tech
  - Major ASR applications
- How to narrow the gap between machine and human speech recognition
- Data-intensive approach for knowledge extraction
- Summary

# Radio Rex – 1920's ASR



A sound-activated toy dog named "Rex" (from Elmwood Button Co.) could be called by name from his doghouse.

# 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**

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

? **4G**

Extended knowledge processing

Prehistory ASR (1925)

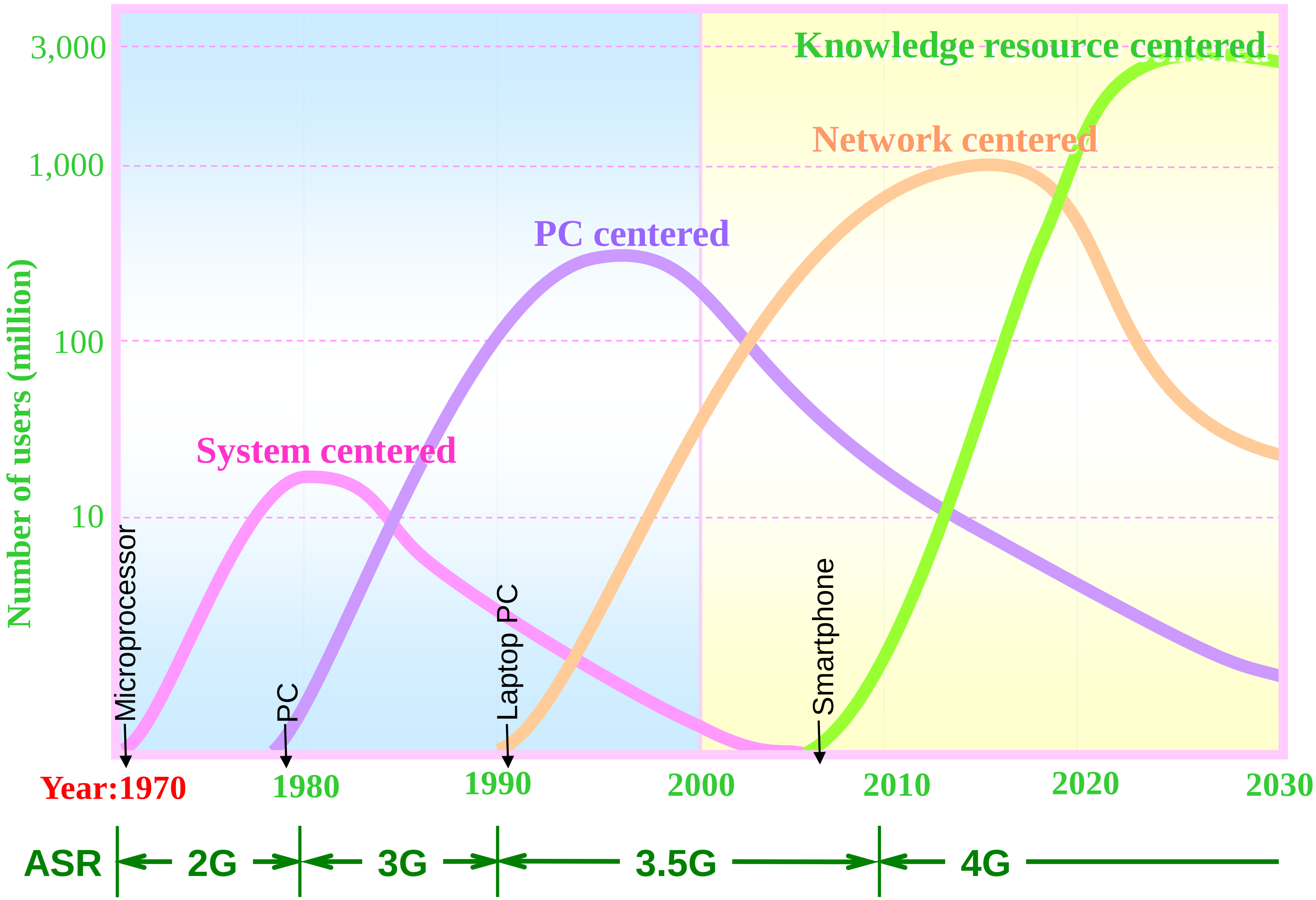
**Our research**
**NTT Labs (+Bell Labs), Tokyo Tech**
**Collaboration with other labs**

# 1st generation technology (1950's-1960's)

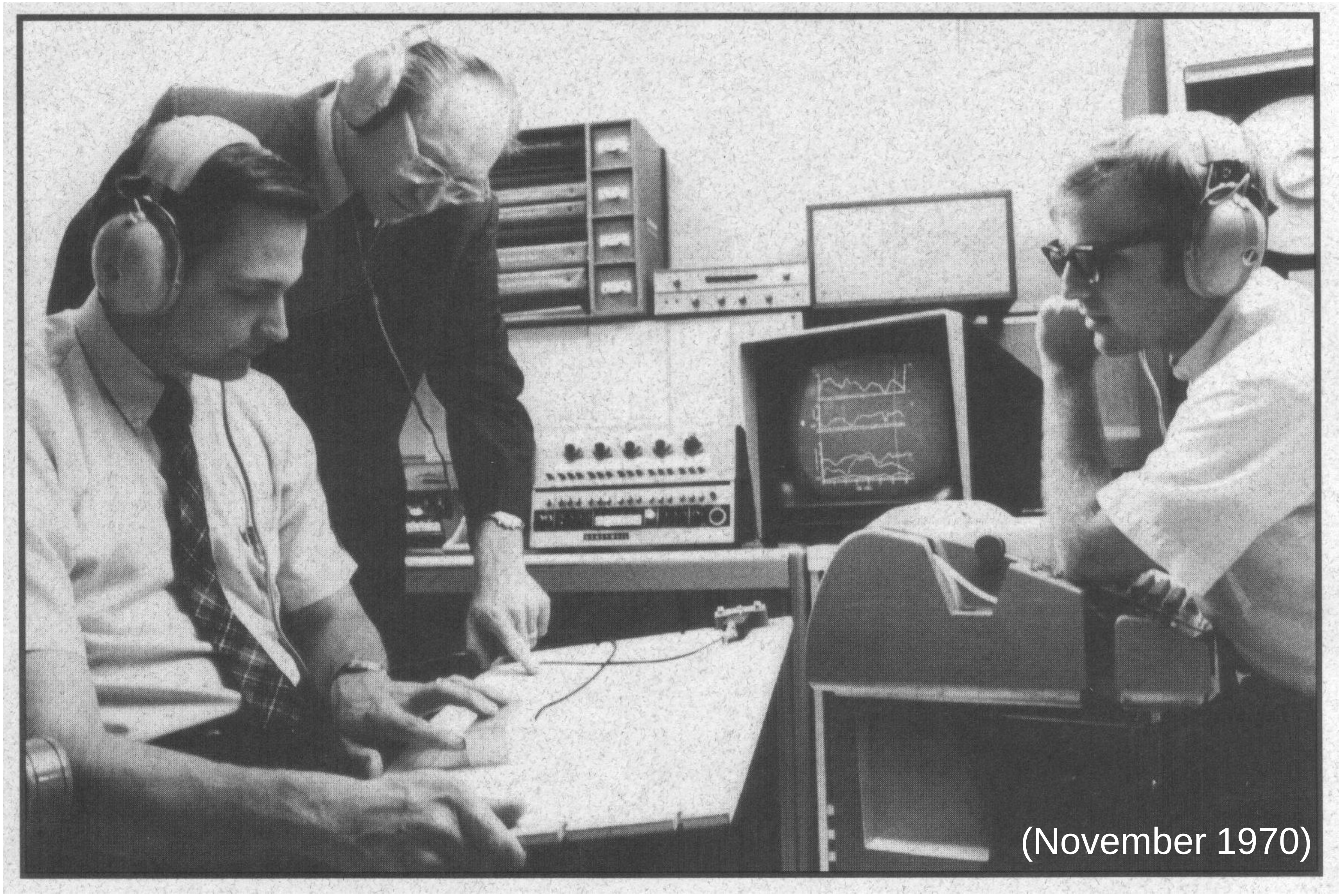
## "Heuristic approaches"

- General
  - The earliest attempts to devise digit/syllable/vowel/phoneme recognition systems
  - Spectral resonances extracted by an analogue filter bank and logic circuits
  - Statistical syntax at the phoneme level
- Early systems
  - Bell labs, RCA Labs, MIT Lincoln Labs (USA)
  - University College London (UK)
  - Radio Research Lab, Kyoto Univ., NEC Labs (Japan)

# IT technology progress



(David C. Moschella: “Waves of Power”)



(November 1970)

Flanagan writes, "… the computer in the background is a Honeywell DDP516. This was the first integrated circuits machine that we had in the laboratory. … with memory of 8K words (one-half of which was occupied by the Fortran II compiler) …"

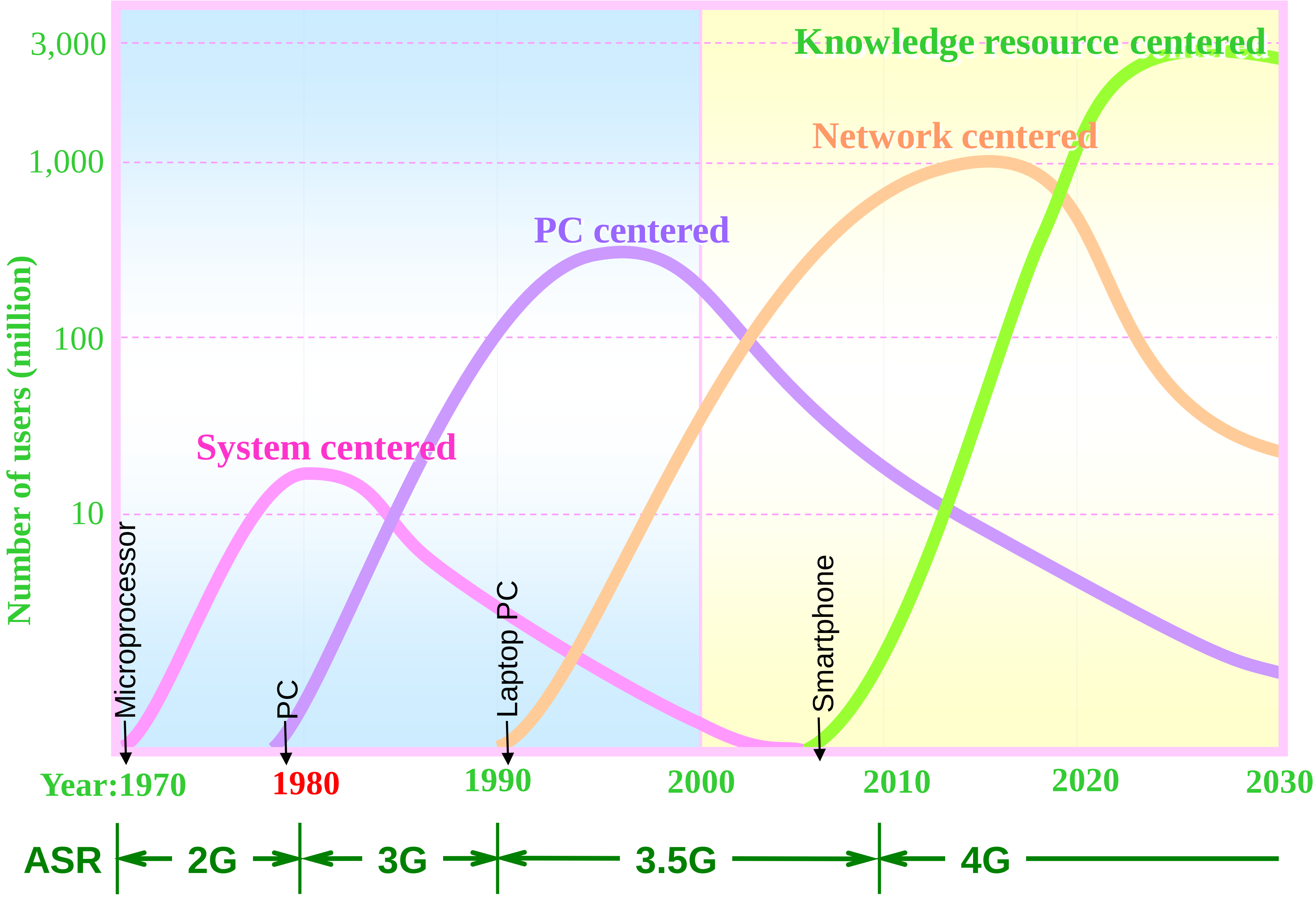
# 2nd generation technology (1970’s) (1)
## “Pattern matching approaches”

- DTW (dynamic time warping)
  - Vintsyuk in Russia (USSR) proposed the use of DP
  - Sakoe & Chiba at NEC labs started to use DP
- Isolated word recognition
  - The area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies in Russia and Japan (Velichko & Zagoruyko, Sakoe & Chiba, and Itakura)
- IBM Labs: large-vocabulary ASR
- Bell Labs: speaker-independent ASR

# 2$^{nd}$ generation technology (1970's) (2)

- Continuous speech recognition
  - Reddy at CMU conducted pioneering research based on dynamic tracking of phonemes
- DARPA program
  - Focus on speech understanding
  - Goal: 1000-word ASR, a few speakers, continuous speech, constrained grammar, less than 10% semantic error
  - Hearsay I & II systems at CMU
  - Harpy system at CMU
  - HWIM system at BBN

# IT technology progress

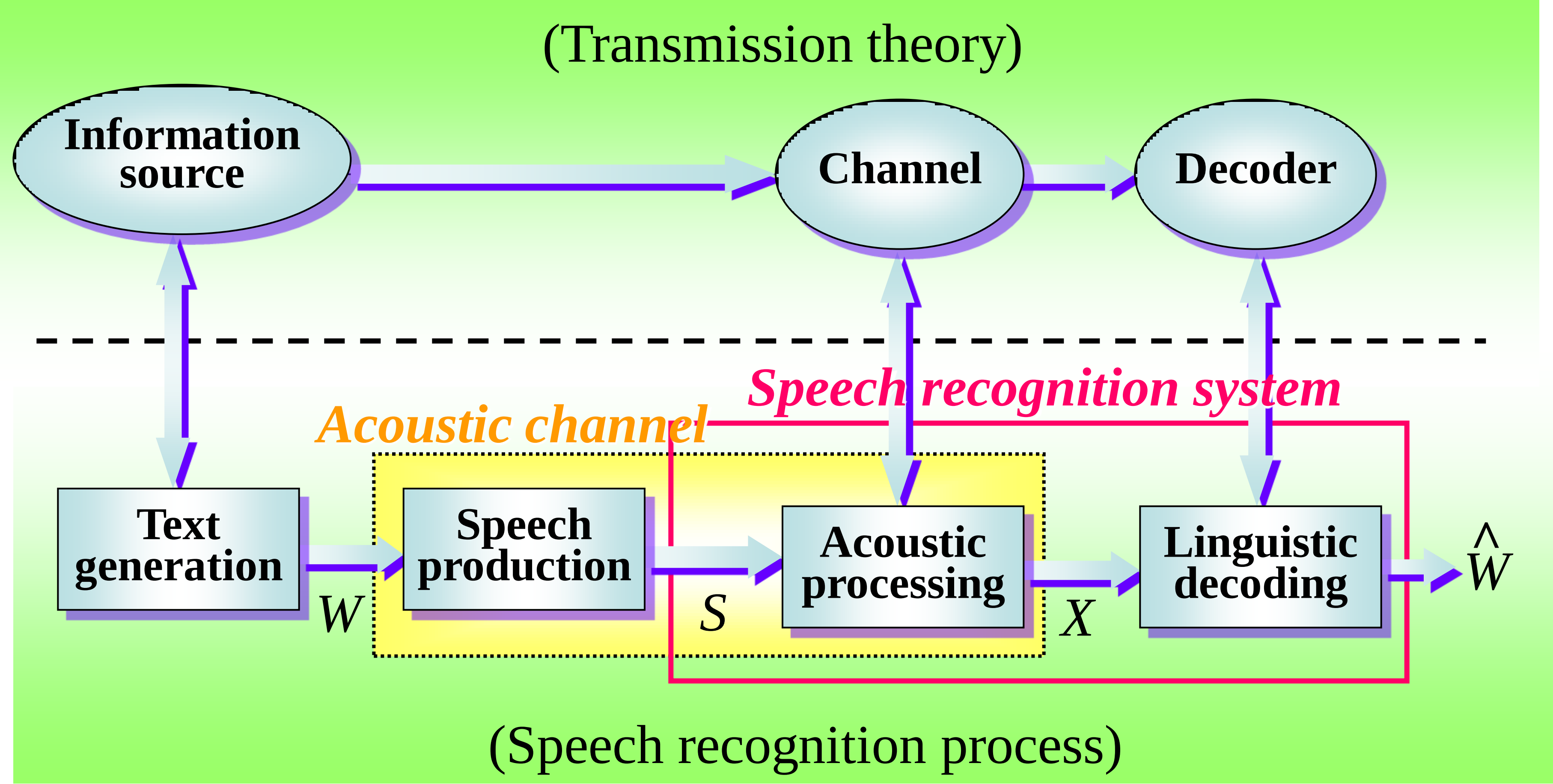


(David C. Moschella: “Waves of Power”)

# 3rd generation technology (1980’s)
# “Statistical framework”

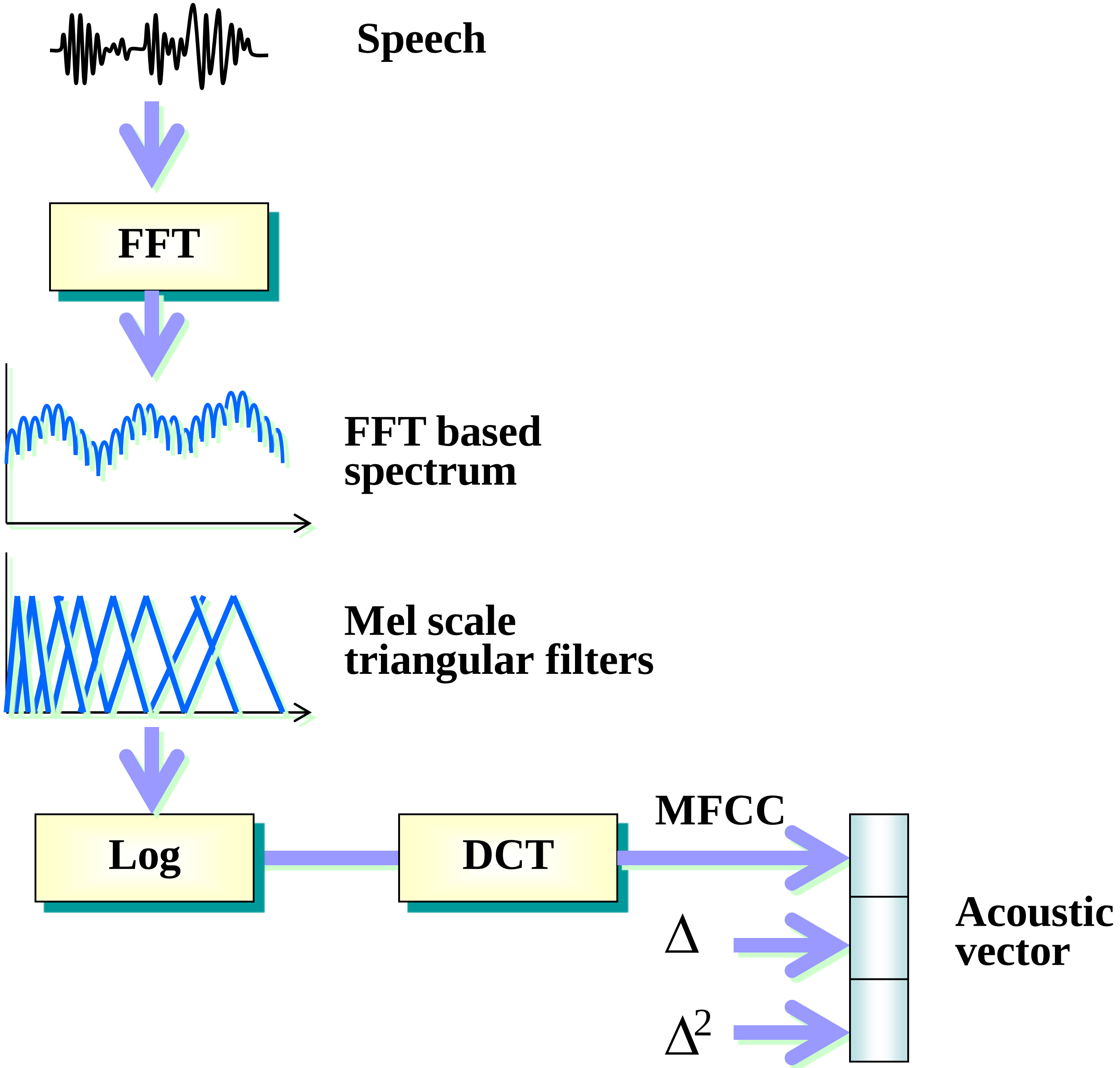
- Connected word recognition
  - Two-level DP, one-pass method, level-building (LB) approach, frame-synchronous LB approach.
- Statistical framework
  - HMM
  - N-gram
  - cepstrum + Δcepstrum
- Neural net
- DARPA program (Resource management task)
  - SPHINX system at CMU
  - BYBLOS system at BBN
  - DECIPHER system at SRI
  - Lincoln Labs, MIT, AT&T Bell Labs

# Structure of speech production and recognition system based on information transmission theory

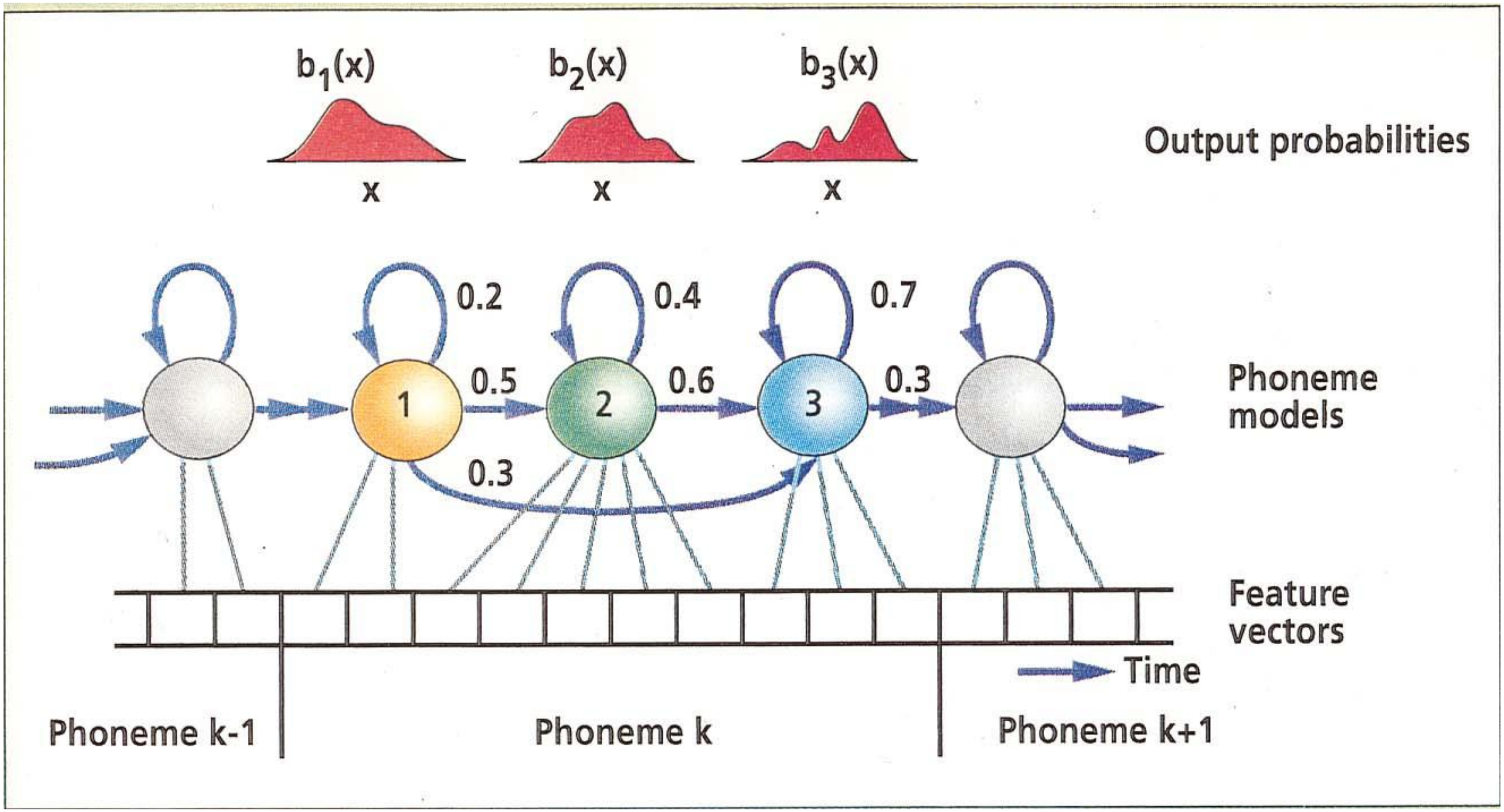


$$\hat{W} = \arg\max_{W} P(W|X) = \arg\max_{W} \frac{P(X|W)P(W)}{P(X)}$$

# MFCC (Mel-frequency cepstral coefficients)-based front-end processor



# Structure of phoneme HMMs



# Statistical language modeling

Probability of the word sequence $w_1{}^k = w_1 w_2 ... w_k$ :

$$P(w_1{}^k) = \prod_{i=1}^{k} P(w_i \,|\, w_1 w_2 \ldots w_{i-1}) = \prod_{i=1}^{k} P(w_i \,|\, w_1{}^{i-1})$$

$$P(w_i \,|\, w_1{}^{i-1}) = N(w_1{}^i) \,/\, N(w_1{}^{i-1})$$

where $N\,(w_1{}^i)$ is the number of occurrences of the string $w_1{}^i$ in the given training corpus.

Approximation by Markov processes:

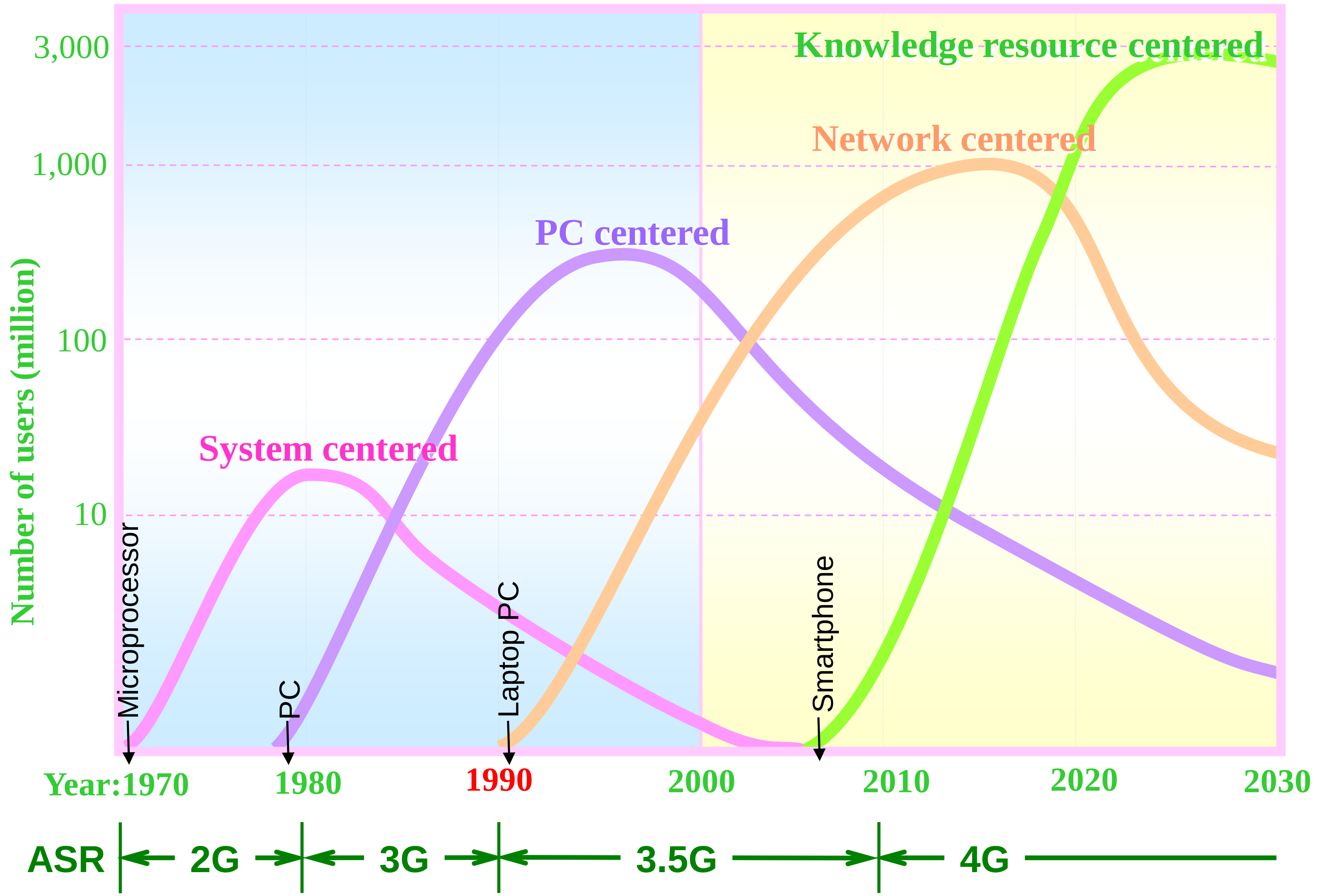
**Bigram** model $\quad P(w_i \,|\, w_1{}^{i-1}) = P(w_i \,|\, w_{i-1})$

**Trigram** model $\quad P(w_i \,|\, w_1{}^{i-1}) = P(w_i \,|\, w_{i-2} w_{i-1})$

Smoothing of trigram by unigram and bigram:

$$\hat{P}(w_i \,|\, w_{i-2} w_{i-1}) = \lambda_1 P(w_i \,|\, w_{i-2} w_{i-1}) + \lambda_2 P(w_i \,|\, w_{i-1}) + \lambda_3 P(w_i)$$

# IT technology progress

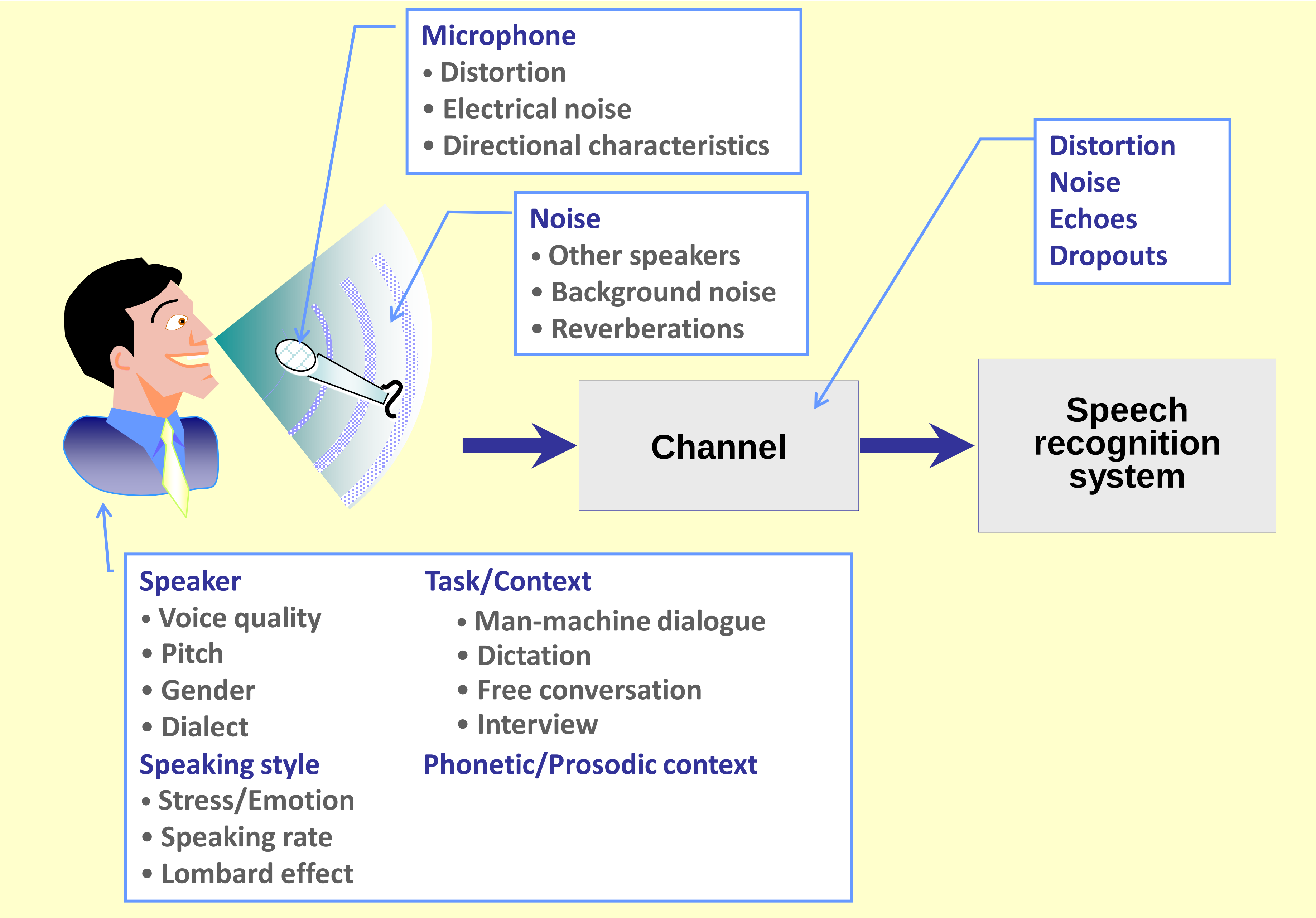


(David C. Moschella: "Waves of Power")

# 3.5th generation technology (1990's)

- Error minimization (discriminative) approach
  - MCE (Minimum Classification Error) approach
  - MMI (Maximum Mutual Information) criterion
  - MPE (Minimum Phone Error) criterion
- Robust ASR
  - Background noise, voice individuality, microphones, transmission channel, room reverberation, etc.
  - VTLN, MLLR, HLDA, fMPE, PMC, etc.
- DARPA program
  - ATIS task
  - Broadcast news (BN) transcription integrated with information extraction and retrieval technology
  - Switchboard task
- Applications
  - Automate and enhance operator services

# Main causes of acoustic variation in speech



**Microphone**
- Distortion
- Electrical noise
- Directional characteristics

**Noise**
- Other speakers
- Background noise
- Reverberations

**Distortion**
**Noise**
**Echoes**
**Dropouts**

**Channel**

**Speech recognition system**

**Speaker**
- Voice quality
- Pitch
- Gender
- Dialect

**Speaking style**
- Stress/Emotion
- Speaking rate
- Lombard effect

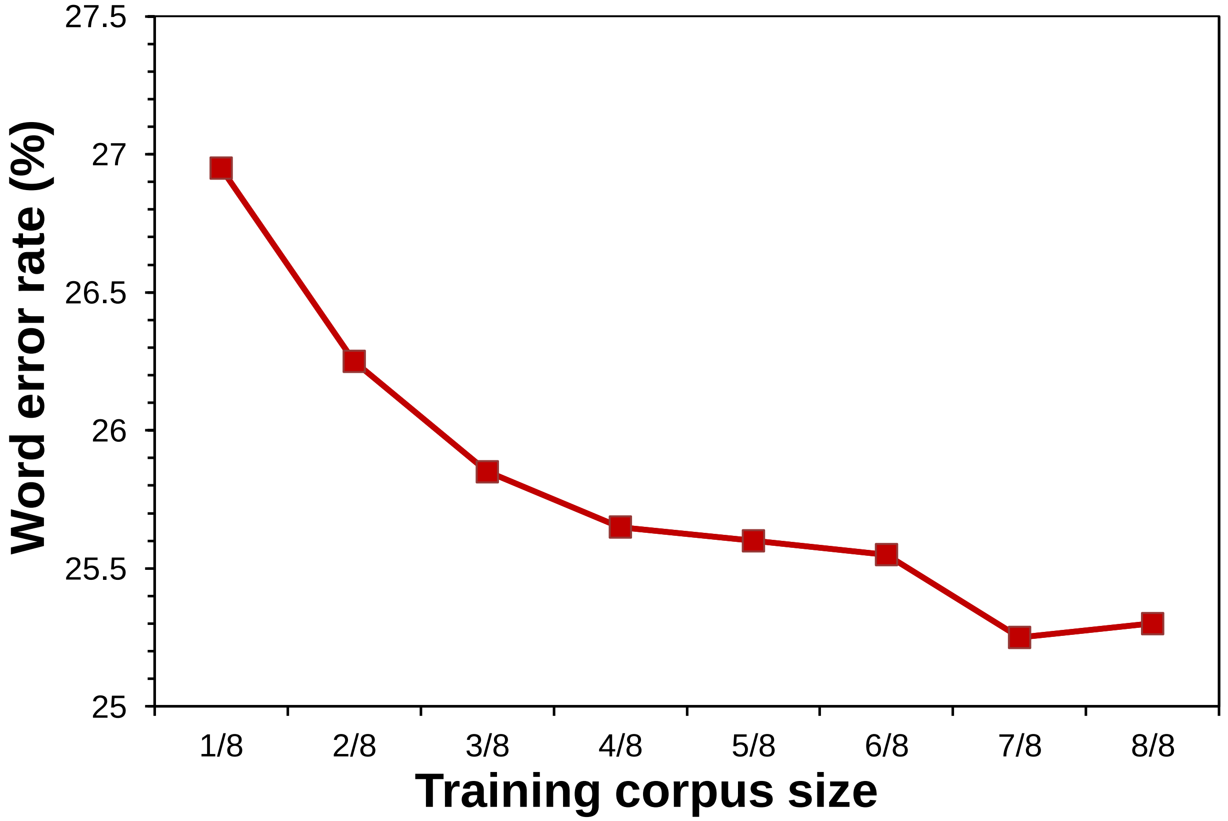
**Task/Context**
- Man-machine dialogue
- Dictation
- Free conversation
- Interview

**Phonetic/Prosodic context**

# 3.5th generation technology (2000's)

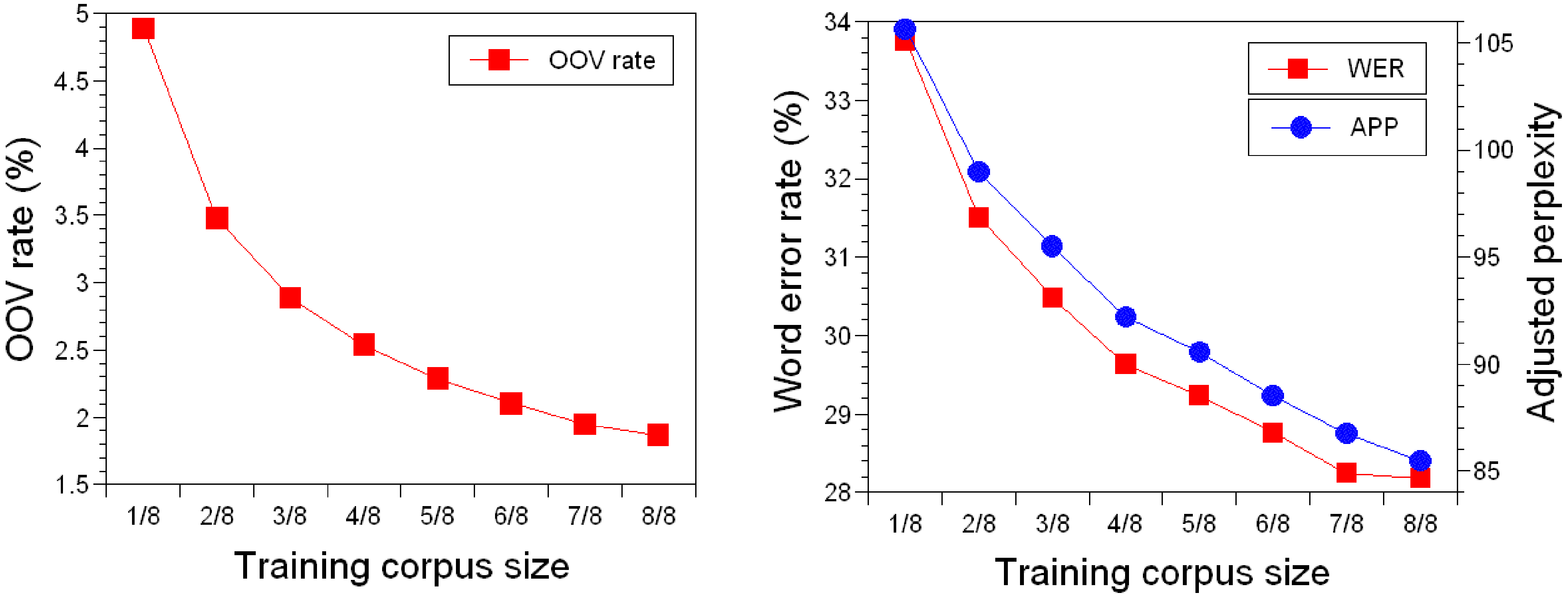
- DARPA program
  - EARS (Effective Affordable Reusable Speech-to-Text) program for rich transcription, GALE
  - Detecting, extracting, summarizing, and translating important information
- Spontaneous speech recognition
  - CSJ lecture project in Japan
  - Meeting projects in US and Europe
- Robust ASR
  - Utterance verification and confidence measures
  - Combining systems or subsystems
- Machine learning
  - Graphical models (DBN)
  - Deep neural network (DNN)

# Word error rate (WER) as a function of the size of acoustic model training data (8/8 = 510 hours)



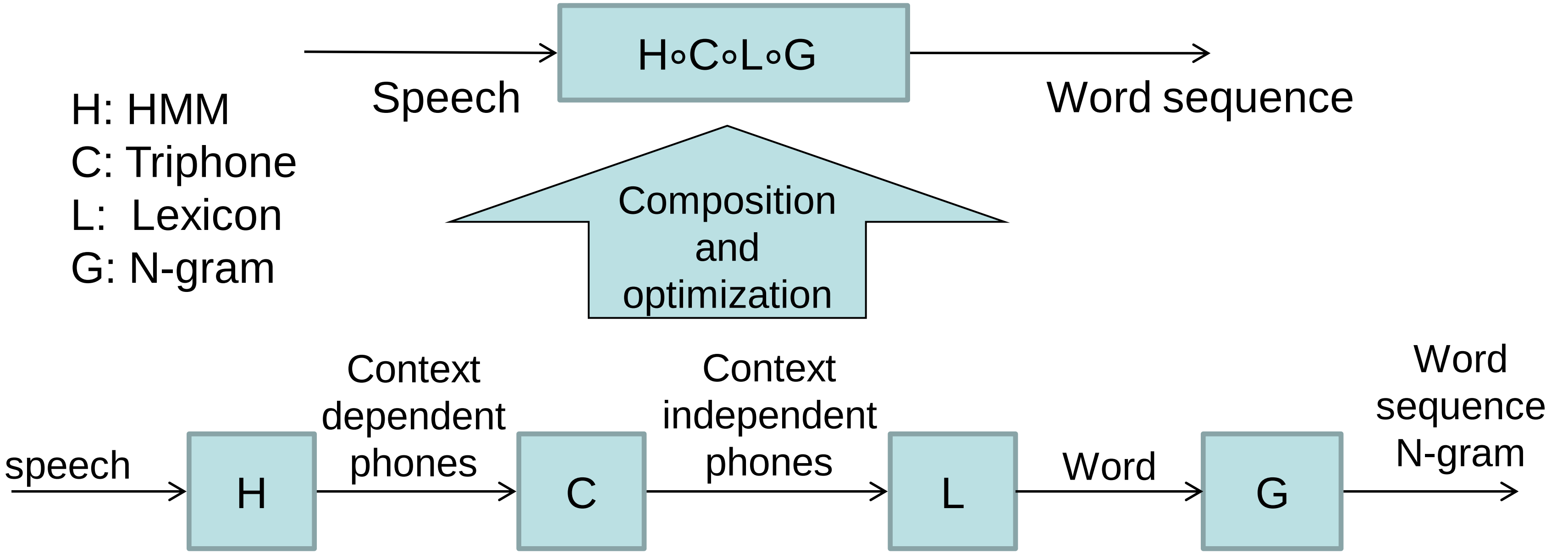
**Task: CSJ (Corpus of Spontaneous Japanese)**

# Out-of-vocabulary (OOV) rate, word error rate (WER) and adjusted test-set perplexity (APP) as a function of the size of language model training data (8/8 = 6.84M words)



**Task: CSJ (Corpus of Spontaneous Japanese)**

# WFST (Weighted Finite State Transducer)-based “T$^3$ decoder”



H: HMM
C: Triphone
L:  Lexicon
G: N-gram

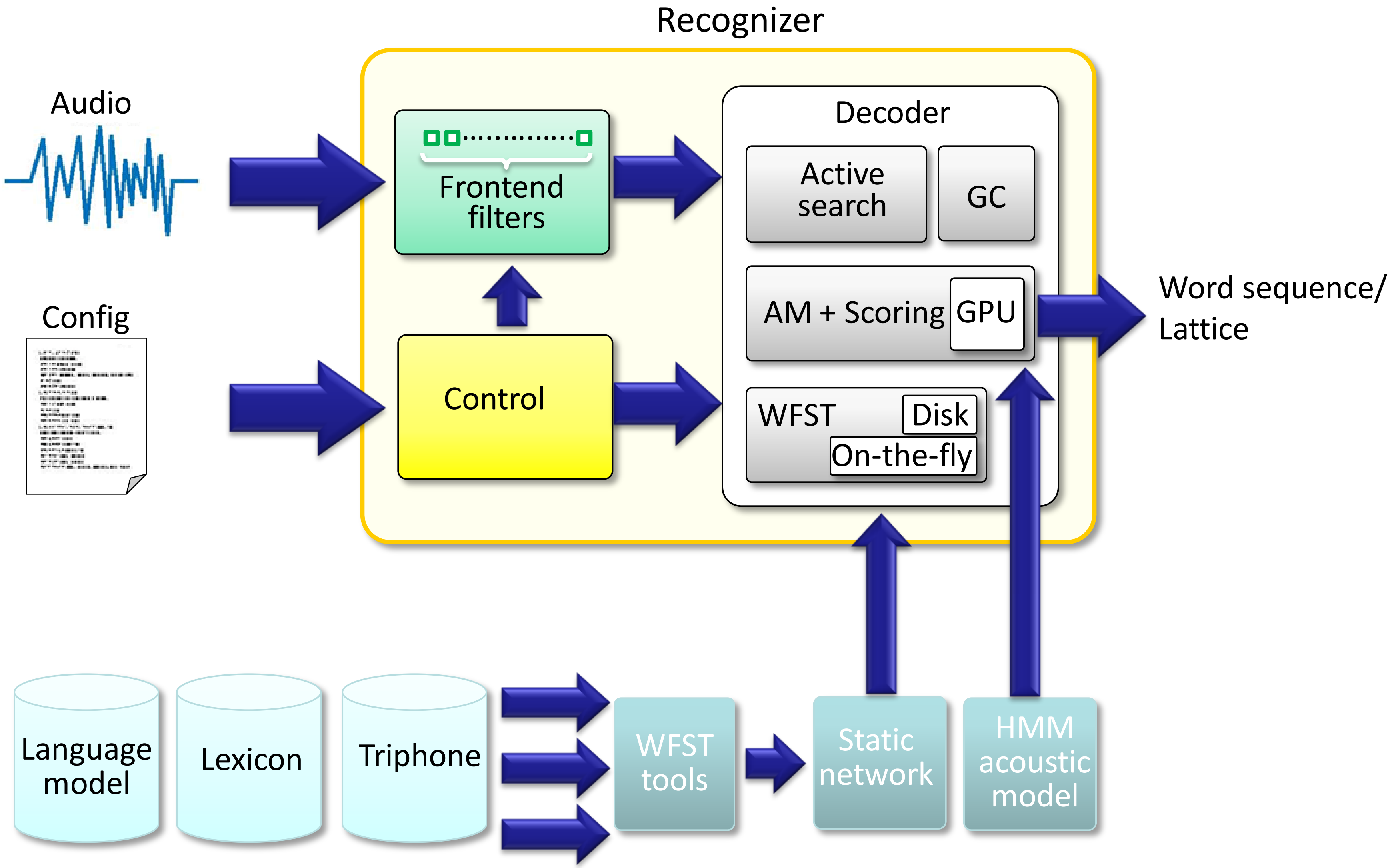
Problems of WFST-based decoder:

- – Large memory requirement
- – Small flexibility
  - Difficult to change partial models

On-the-fly composition

Parallel decoding

# Structure of the T$^3$ decoder

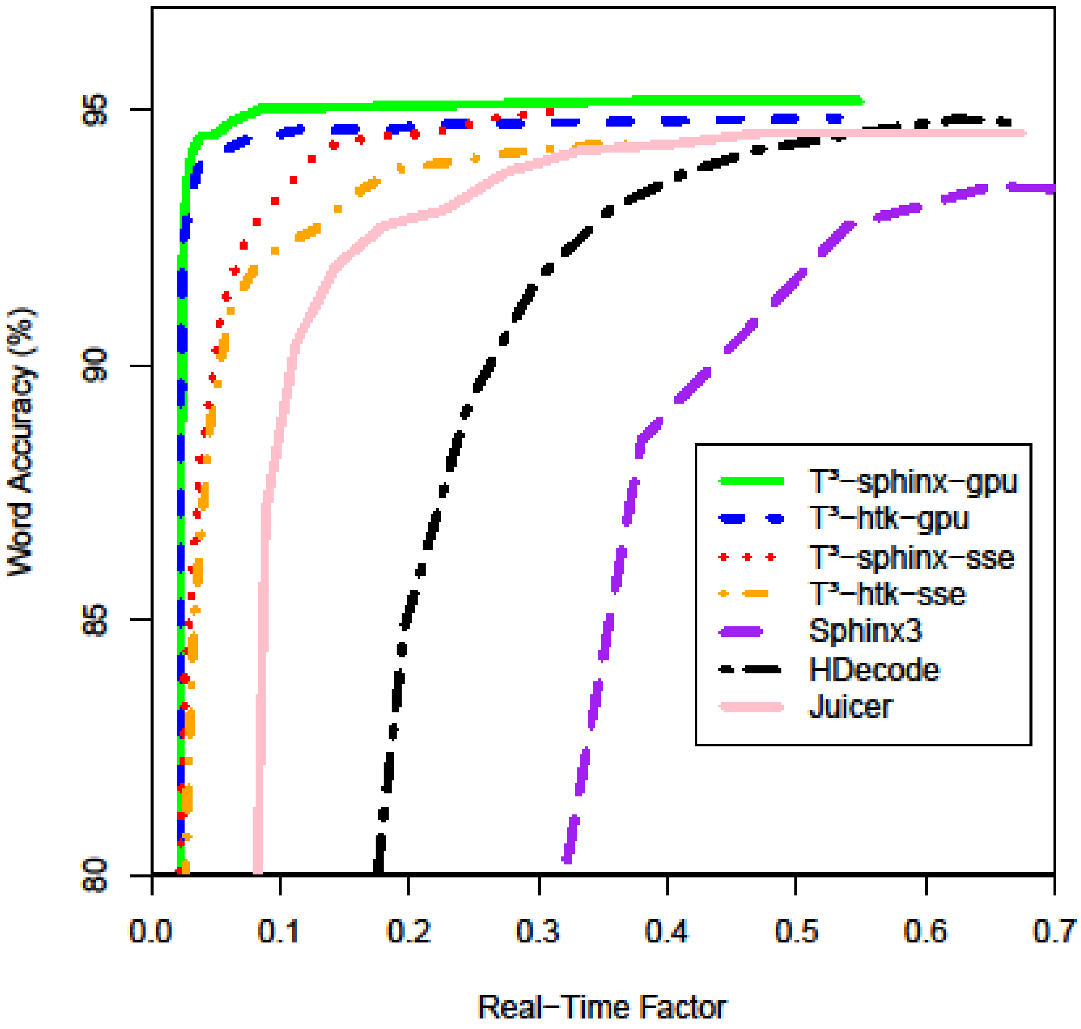


# RTF vs. word accuracy for the various decoders and acoustic models (WSJ 5k task)

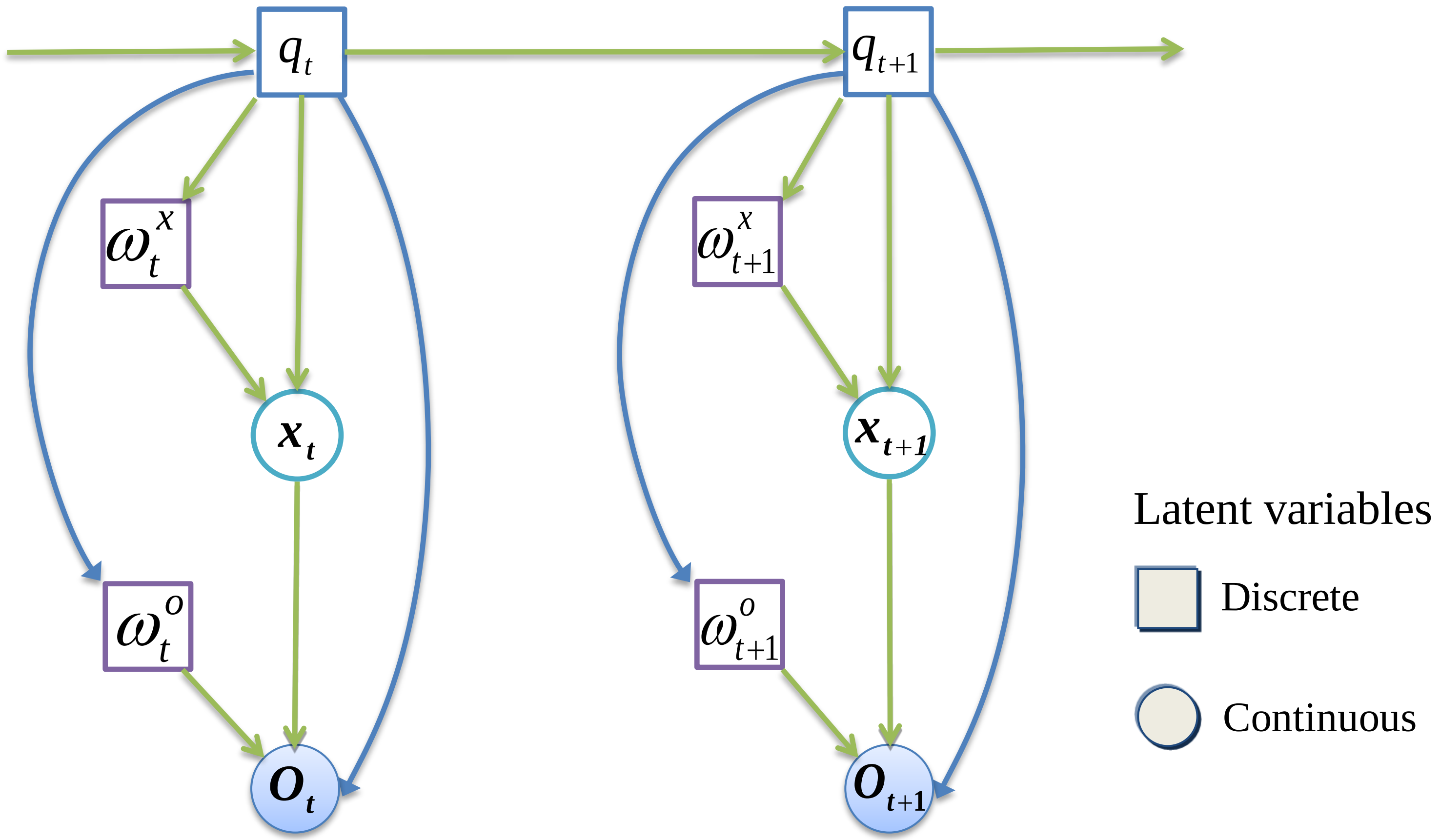
- **Decoder:**
  **T$^3$, Sphinx3, HDecode, Juicer**

- **AM:**
  **Sphinx, HDecode**

- **Cascade construction:**
  $$\pi(\det(C \circ \det(L \circ G)))$$



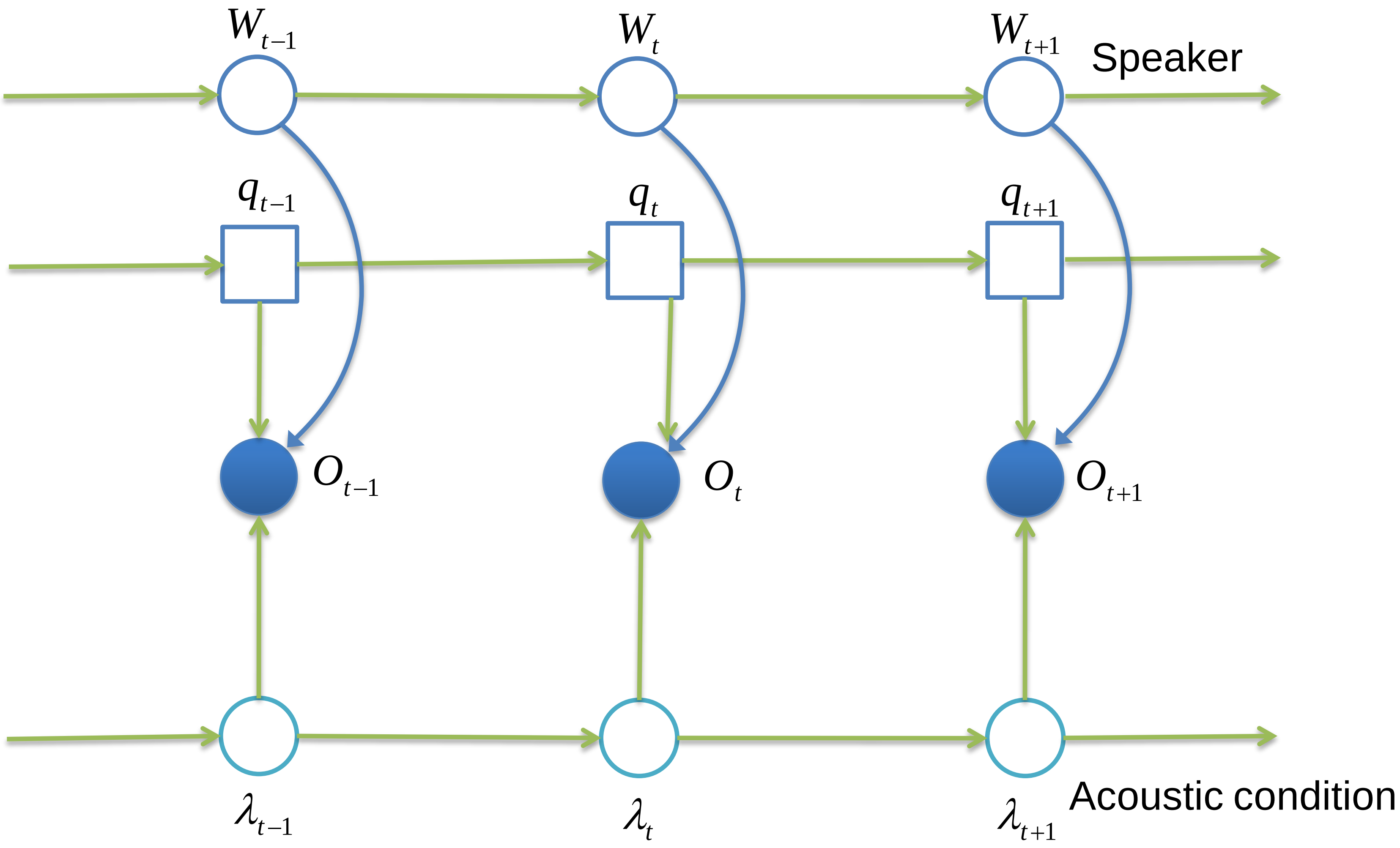
# DBN representing a factor analyzed HMM



$x_t$ : state vector, $O_t$ : observation vector, $q_t$ : HMM state,
$\omega_t^x$, $\omega_t^o$ : mixture indicator

(A.-V. I. Rosti et al., 2002)

# DBN for acoustic factorization



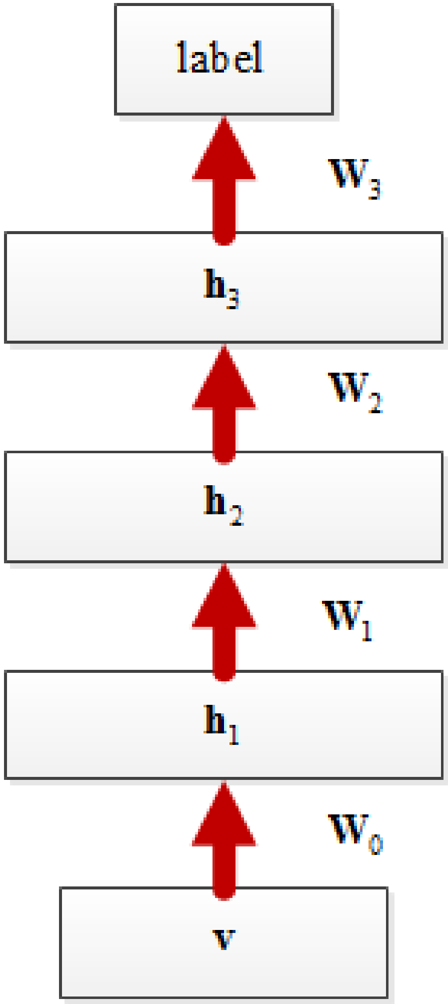
(M. J. F. Gales, 2001)

# Deep neural network

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

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

- 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

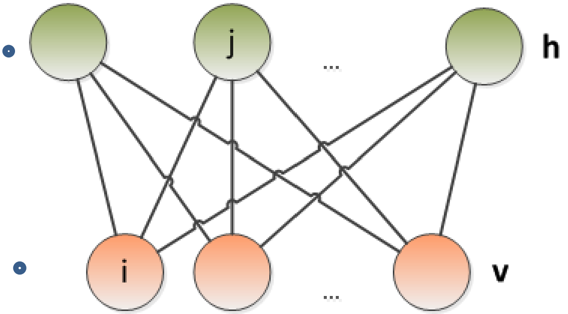


(Dong Yu, 2012)

# Restricted Boltzmann machine

Hidden Layer
No within layer connection

Visible Layer
No within layer connection



- 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

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

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

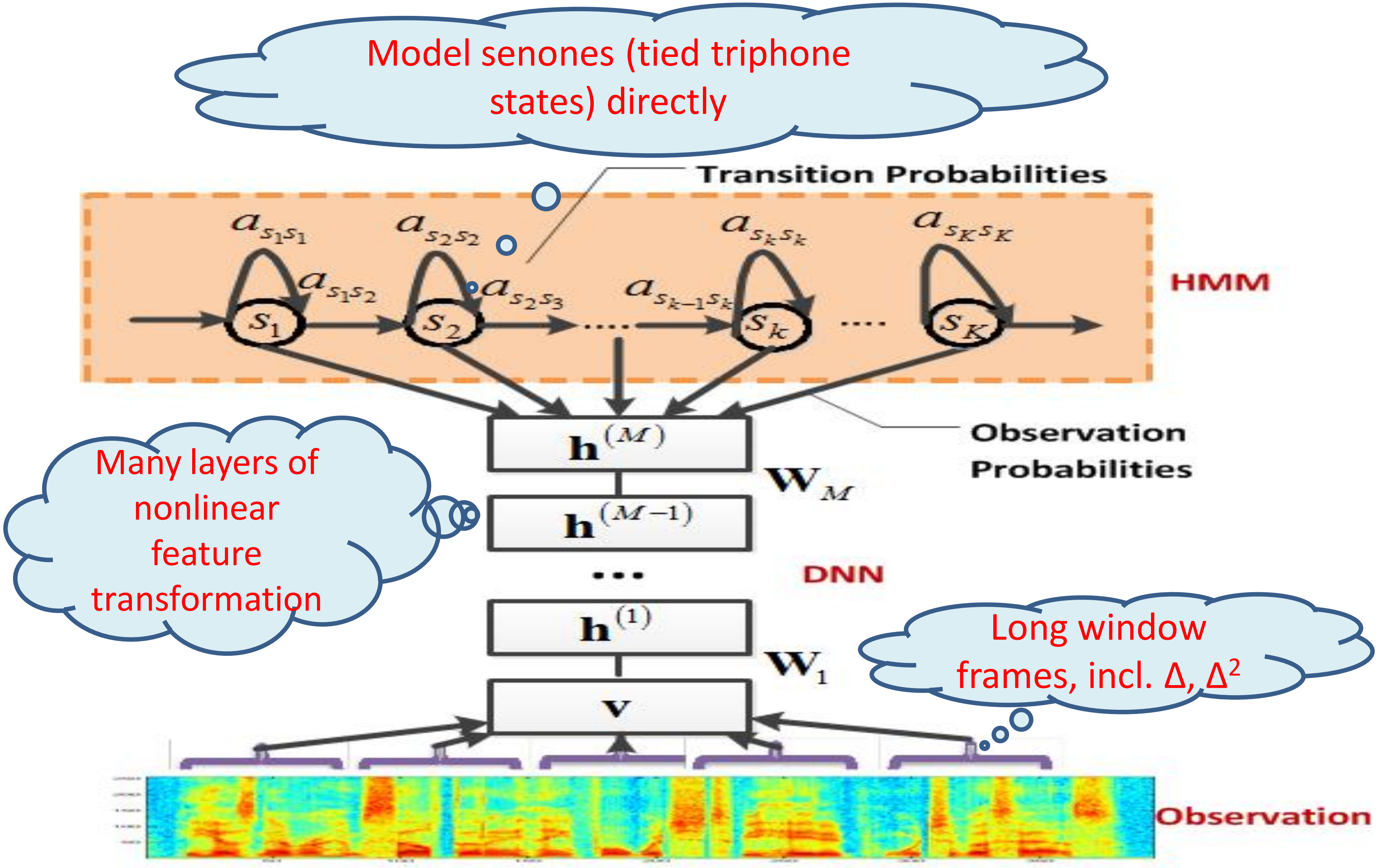
(Dong Yu, 2012)

# 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


Model senones (tied triphone states) directly

Many layers of nonlinear feature transformation

Long window frames, incl. Δ, Δ²

(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?

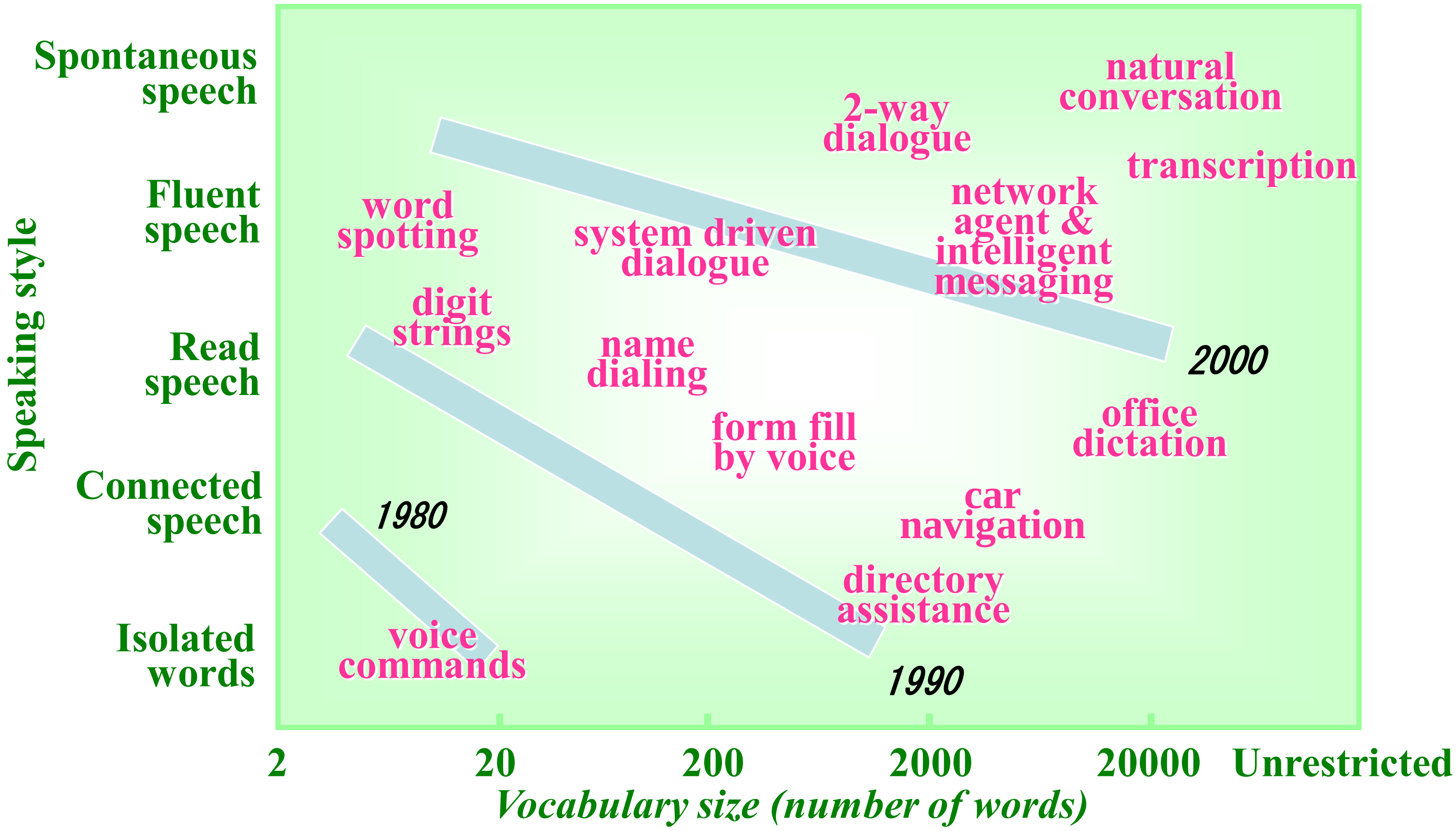
(Seide, Li, Yu 2011, Seide, Li, Chen, Yu 2011)

| 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)

# Progress of speech recognition technology since 1980



# Many applications now

# Major speech recognition applications

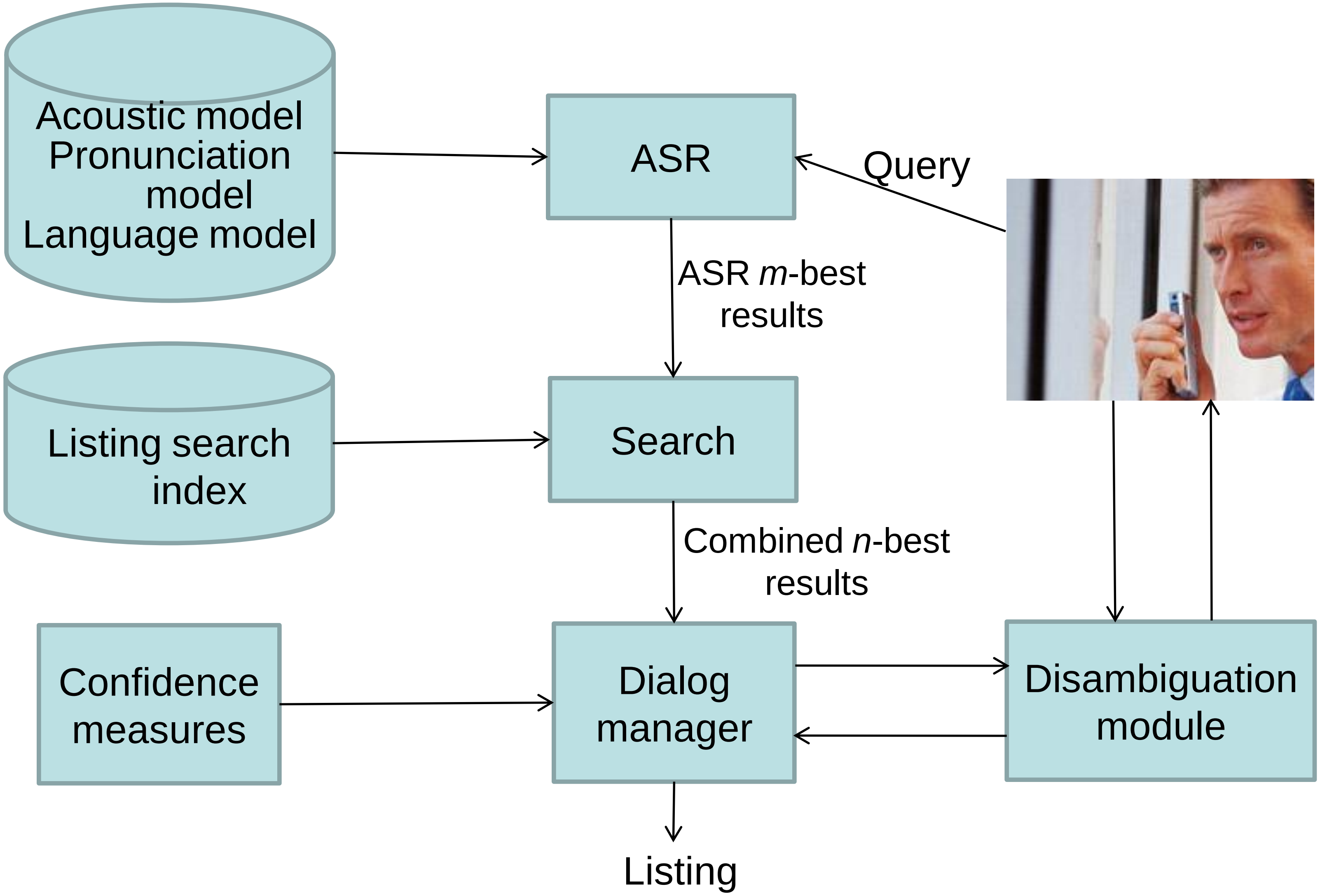
- **Spoken dialog systems** for accessing information services and voice search

  (e.g. directory assistance, flight information systems, stock price quotes, virtual agents, How May I Help You?, GOOG-411, livesearch411, Vlingo, and Google VS)

- Systems for **transcription, summarization and information extraction** from speech documents

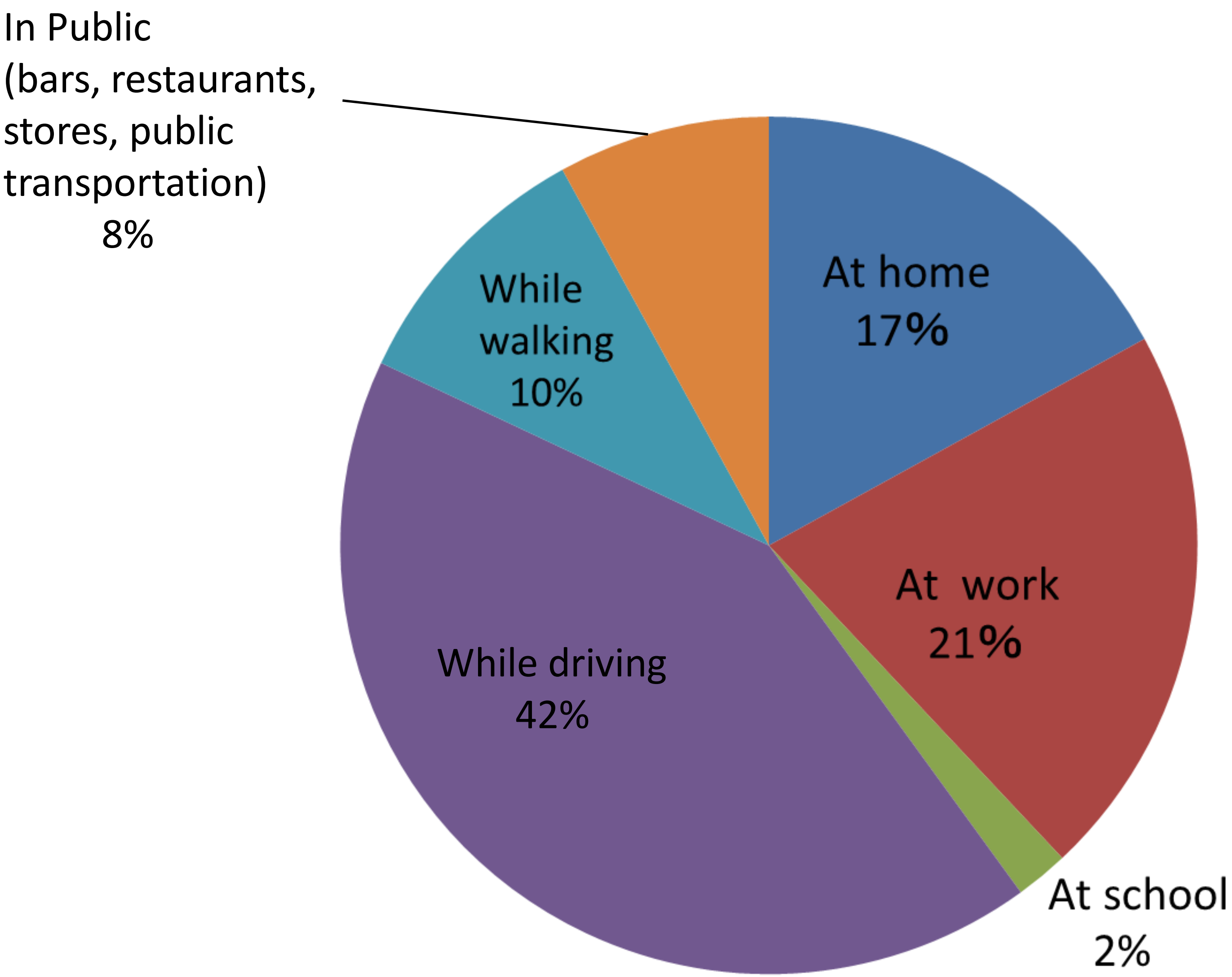
  (e.g. broadcast news, meetings, lectures, presentations, congressional records, court records, and voicemails)

# Voice search system architecture

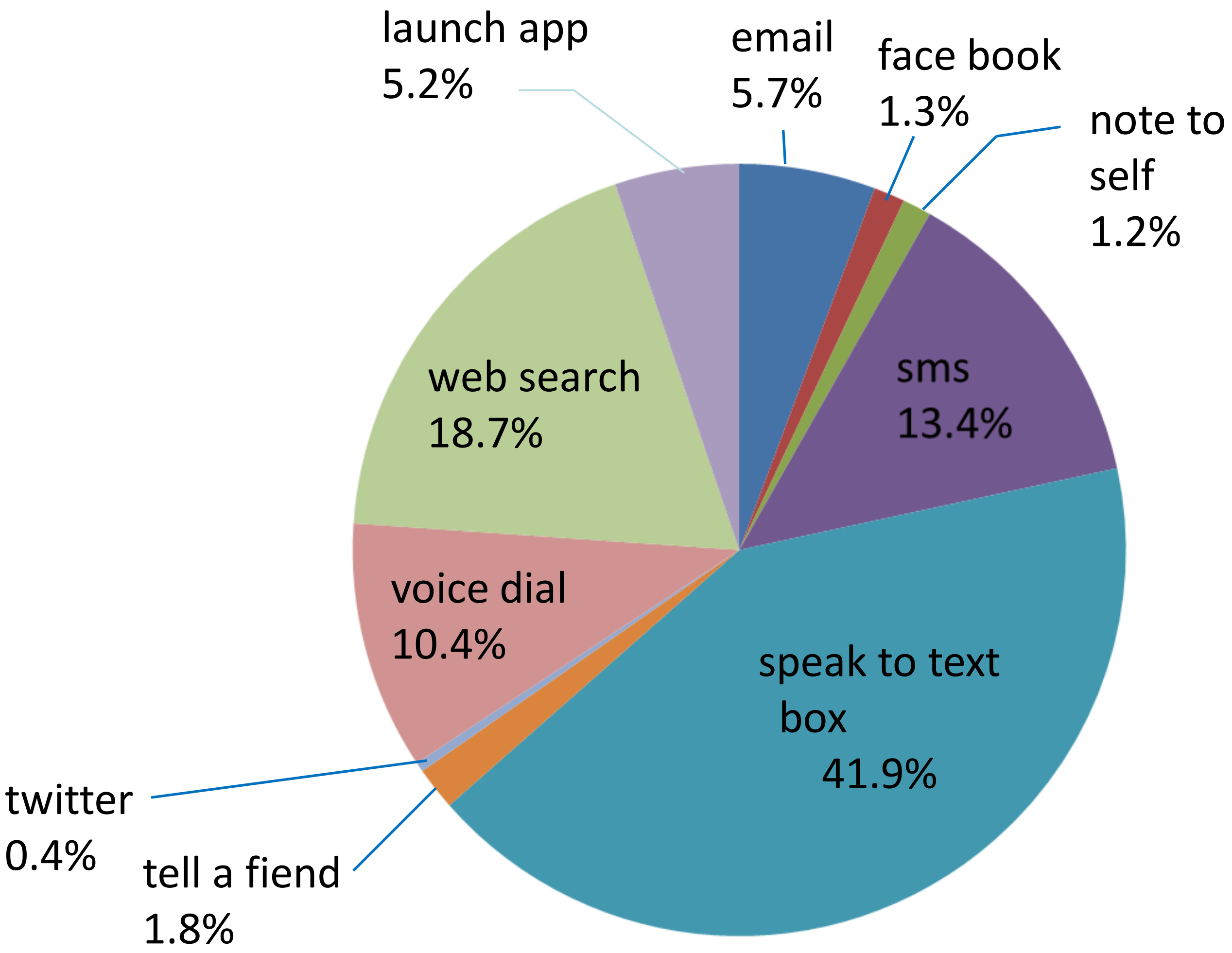


(Y.-Y. Wang et al., 2008)

# Context of *Vlingo* mobile interface usage (Self-reported)

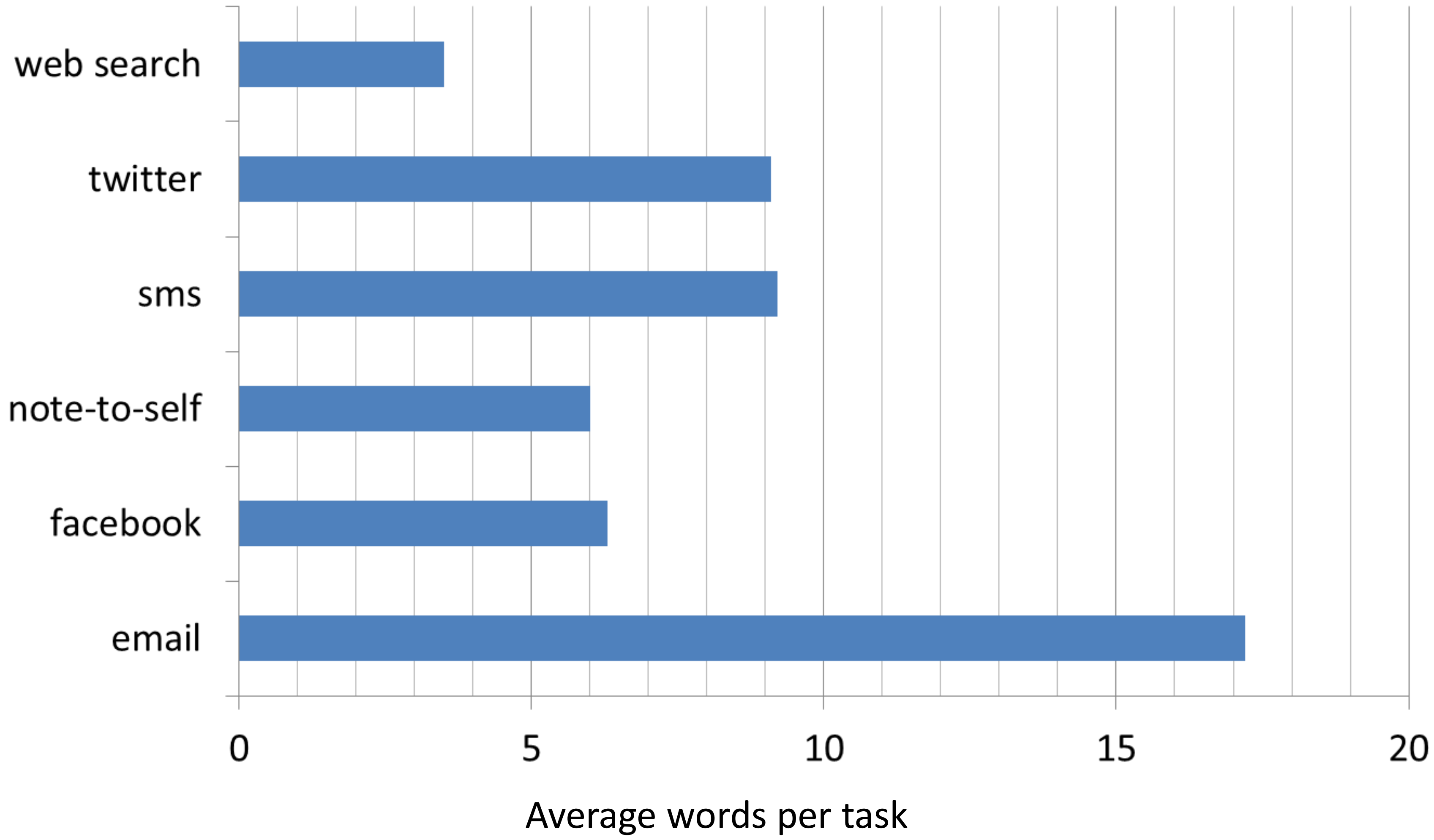


# *Vlingo* Blackberry usage function



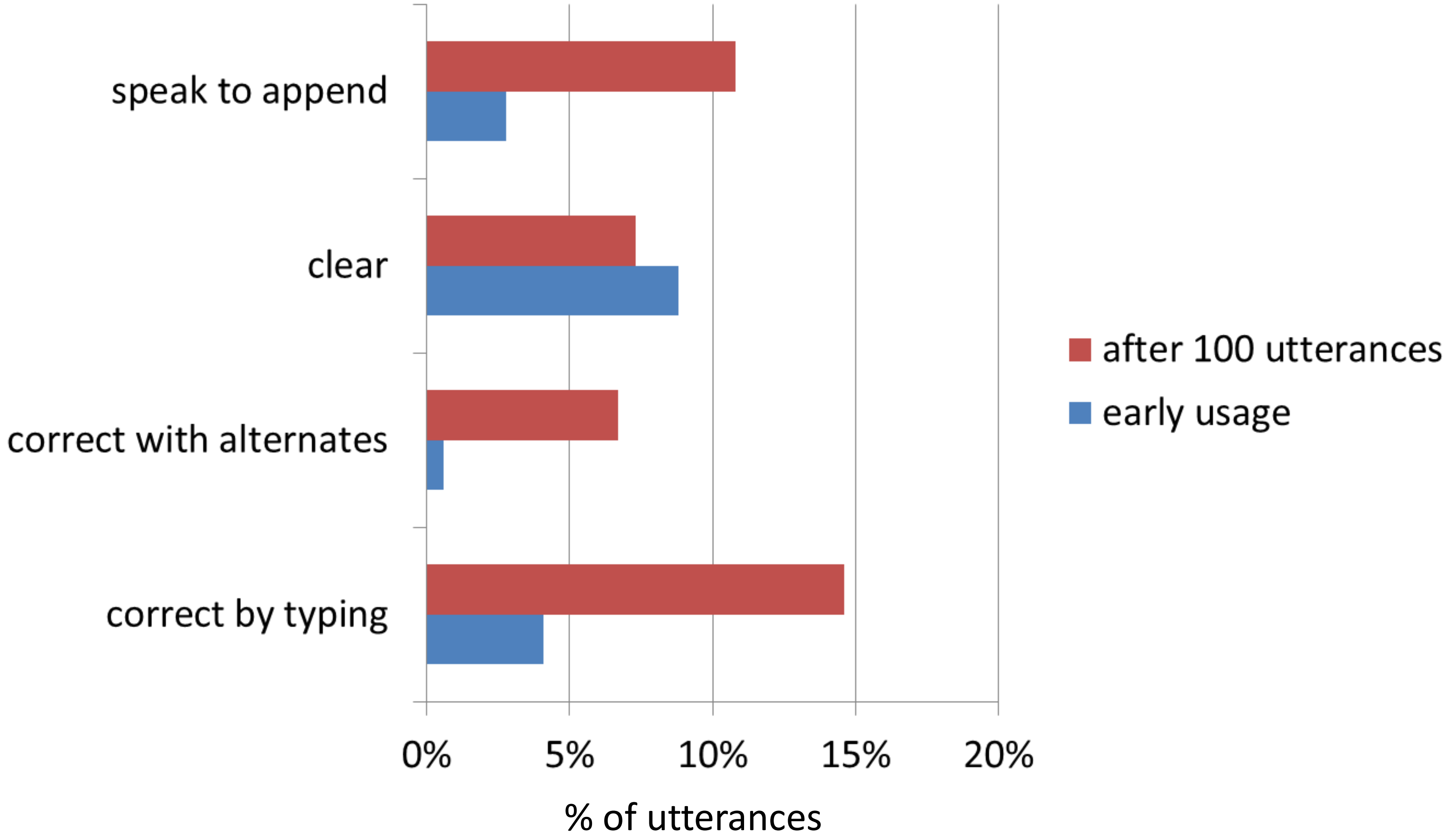
“Speak to text box” usage is the case where users speak into an existing application (hence, using the speech interface as a keyboard).

# Average spoken words per *Vlingo* task



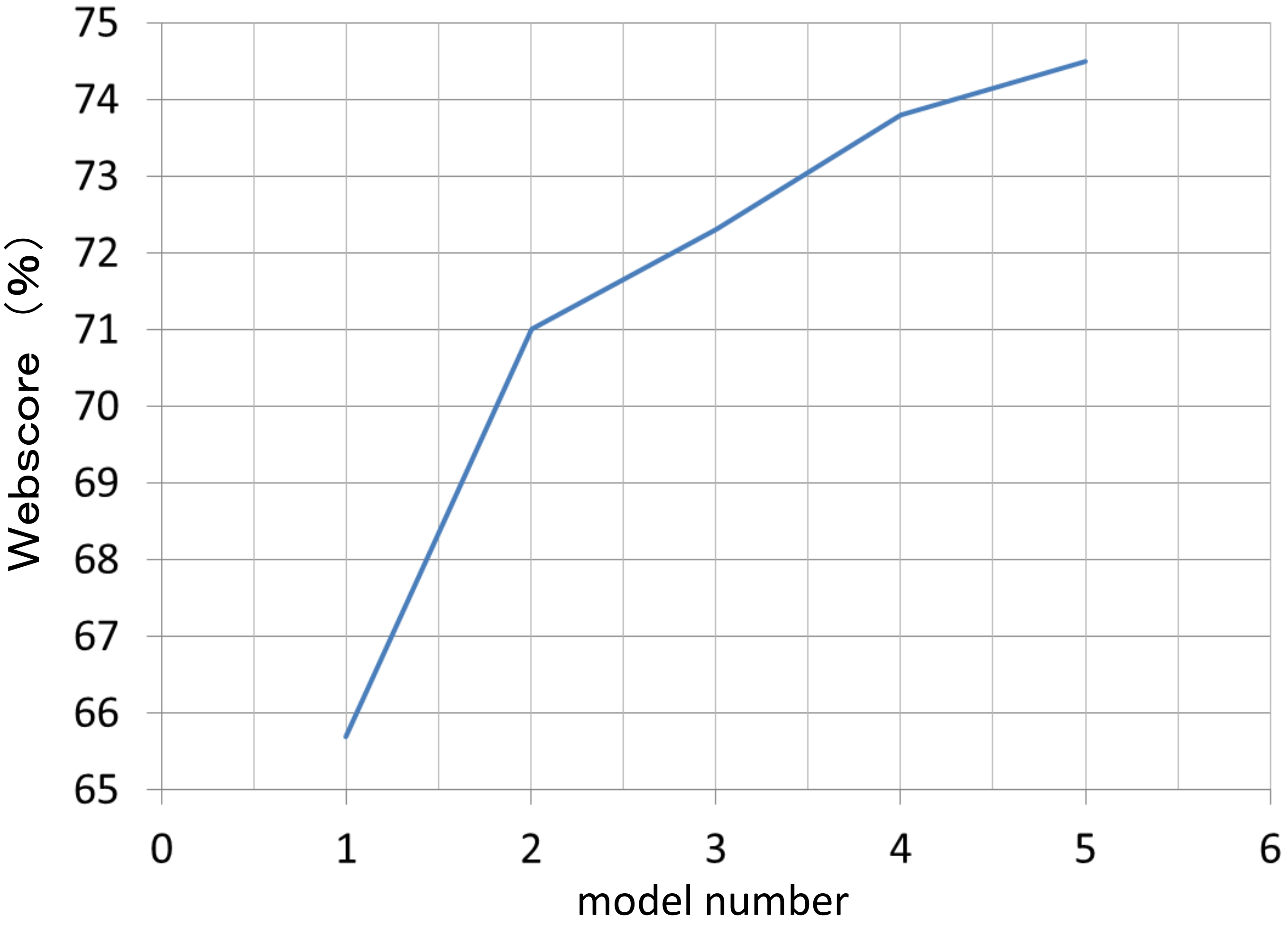
When speaking to Vlingo, users accomplish most tasks with only a small number of words.

# *Vlingo* user behavior by expertise



Initial users tend to clear results when faced with recognition issues, whereas expert users are much more likely to correct either by typing or selecting from alternate results.

# Google Voice Search accuracy evolution over time

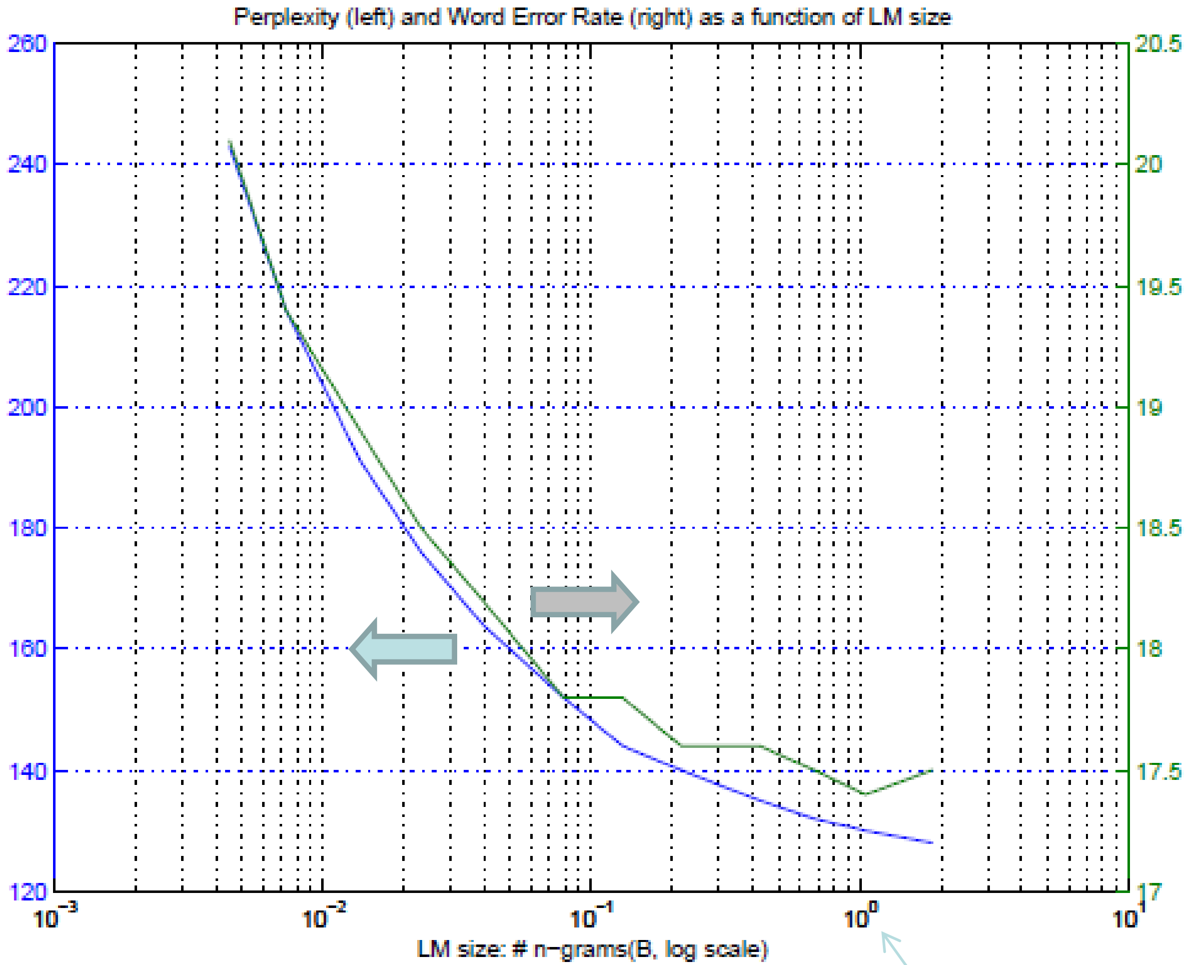


Vocabulary size: 46K

**Size of transcribed data: AM1: mismatched, AM2: 1K hours, AM3: 2K hours, AM4: 7K hours, AM5: more.   Model structure and training methods are also improved.**

(J. Schalkwyk at al., 2010)

# Perplexity and WER as a function of 3-gram LM size



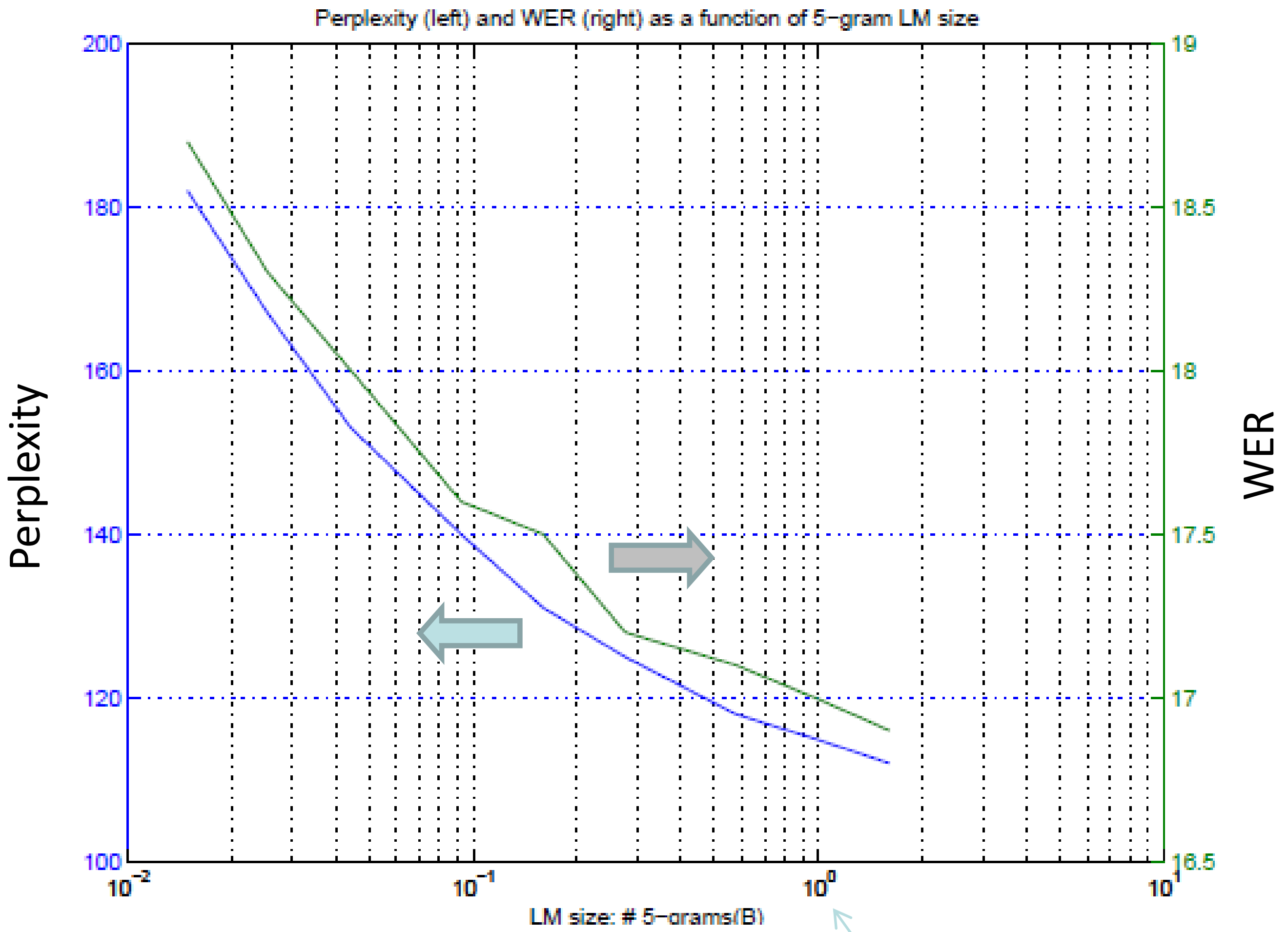
Perplexity

WER

3-gram LM size

1 billion

# Perplexity and WER as a function of 5-gram LM size



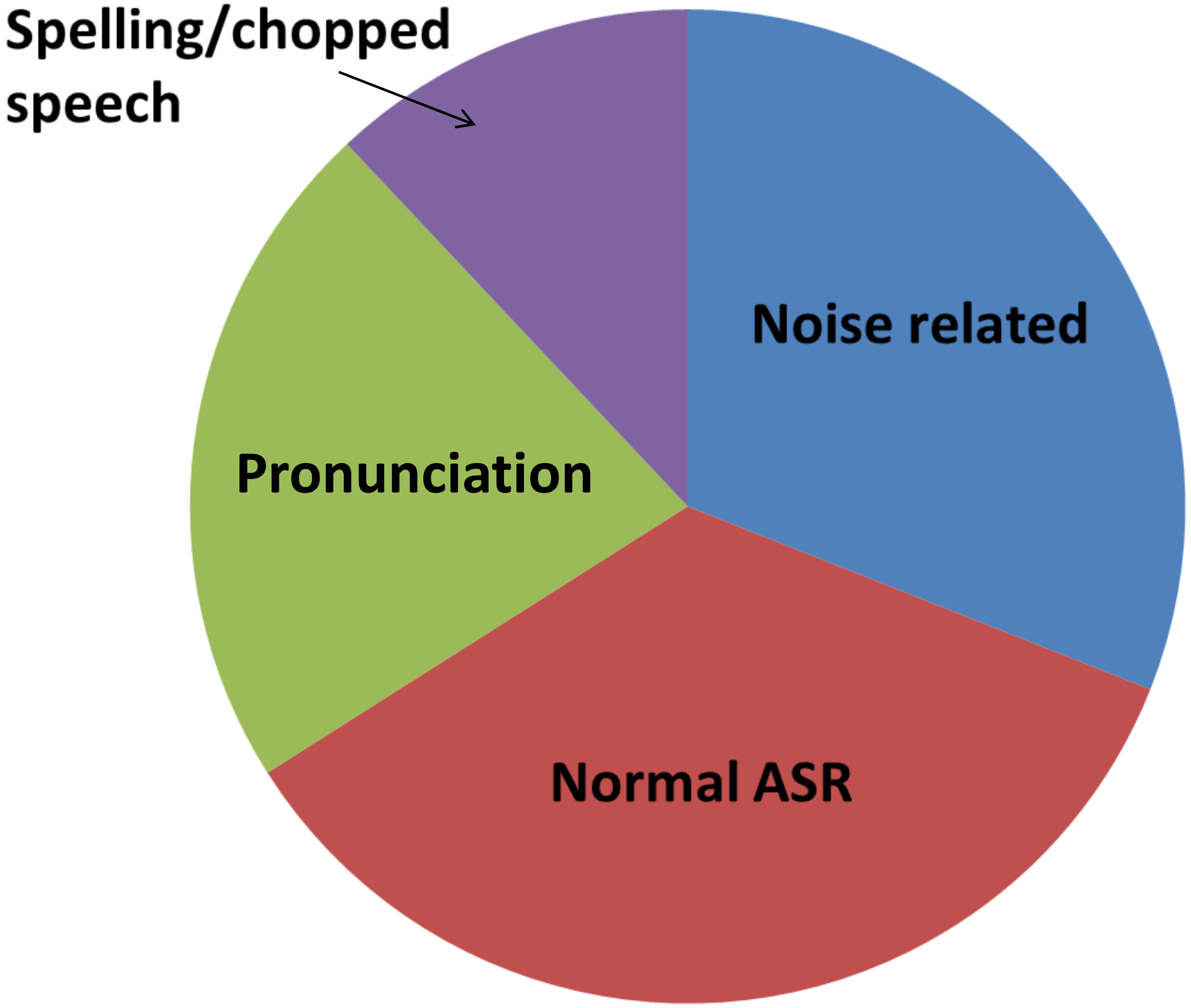
5-gram LM size

1 billion

# Human vs. machine recognition word error rates across a variety of tasks (Lippmann, 1997)

| Task | Machine performance % | Human performance % |
|---|---|---|
| Connected digits | 0.72 | 0.009 |
| Letters | 5 | 1.6 |
| Resource management | 3.6 | 0.1 |
| WSJ (Wall Street Journal) | 7.2 | 0.9 |
| Switchboard | 43 | 4 |

# ASR error analysis for a voice search application



(Y.-Y. Wang et al., 2008)

# Major ASR problems

- Robustness against various speech variations

  - Speakers, noise, language, topics, etc.

- Out-of-vocabulary (OOV) problem

  - Don’t know when we know

- Few advances in basic understanding
- It takes a long time to build a system for a new language; requires a large amount of expensive speech databases

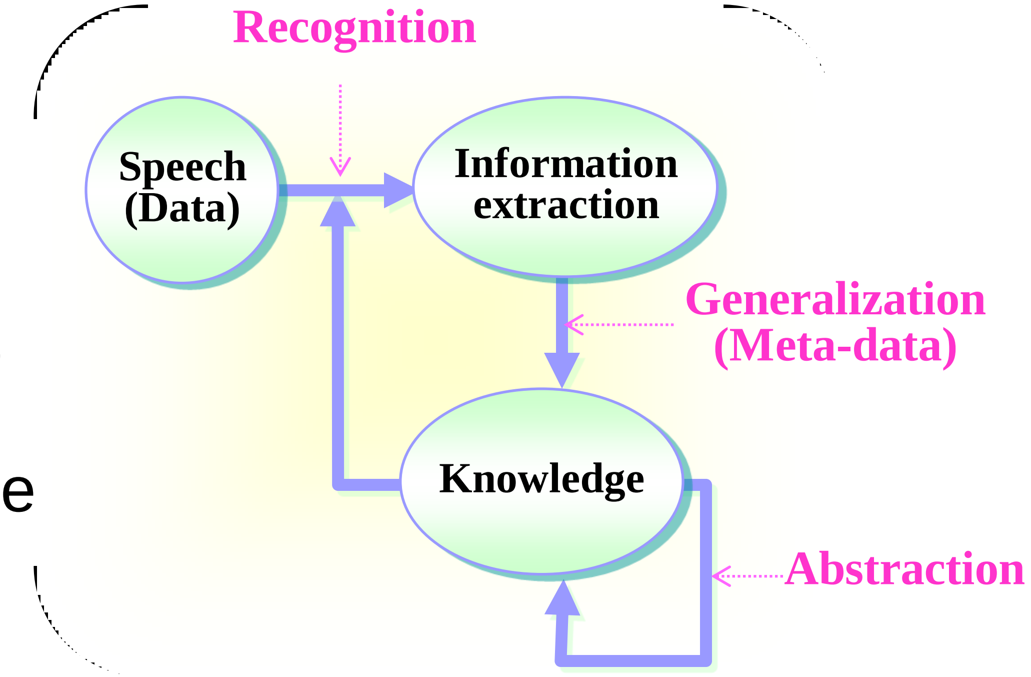
# Knowledge sources for speech recognition

Human speech recognition is a matching process whereby an audio signal is matched to existing knowledge (comprehension)

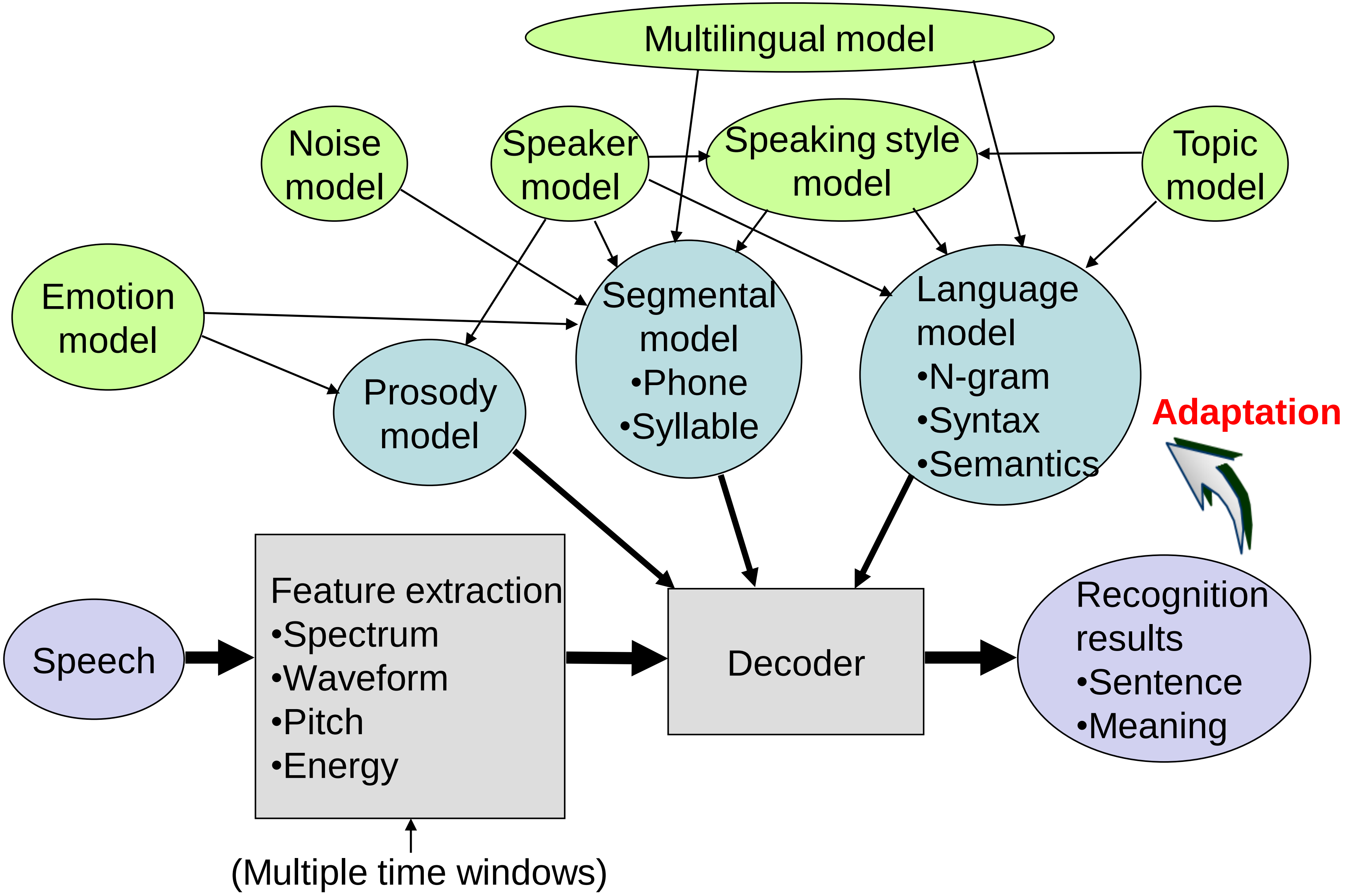
- Knowledge (Meta-data)
  - Domain and topics
  - Context
  - Semantics
  - Speakers
  - Environment, etc.

- How to incorporate knowledge sources into the statistical ASR framework is a key issue

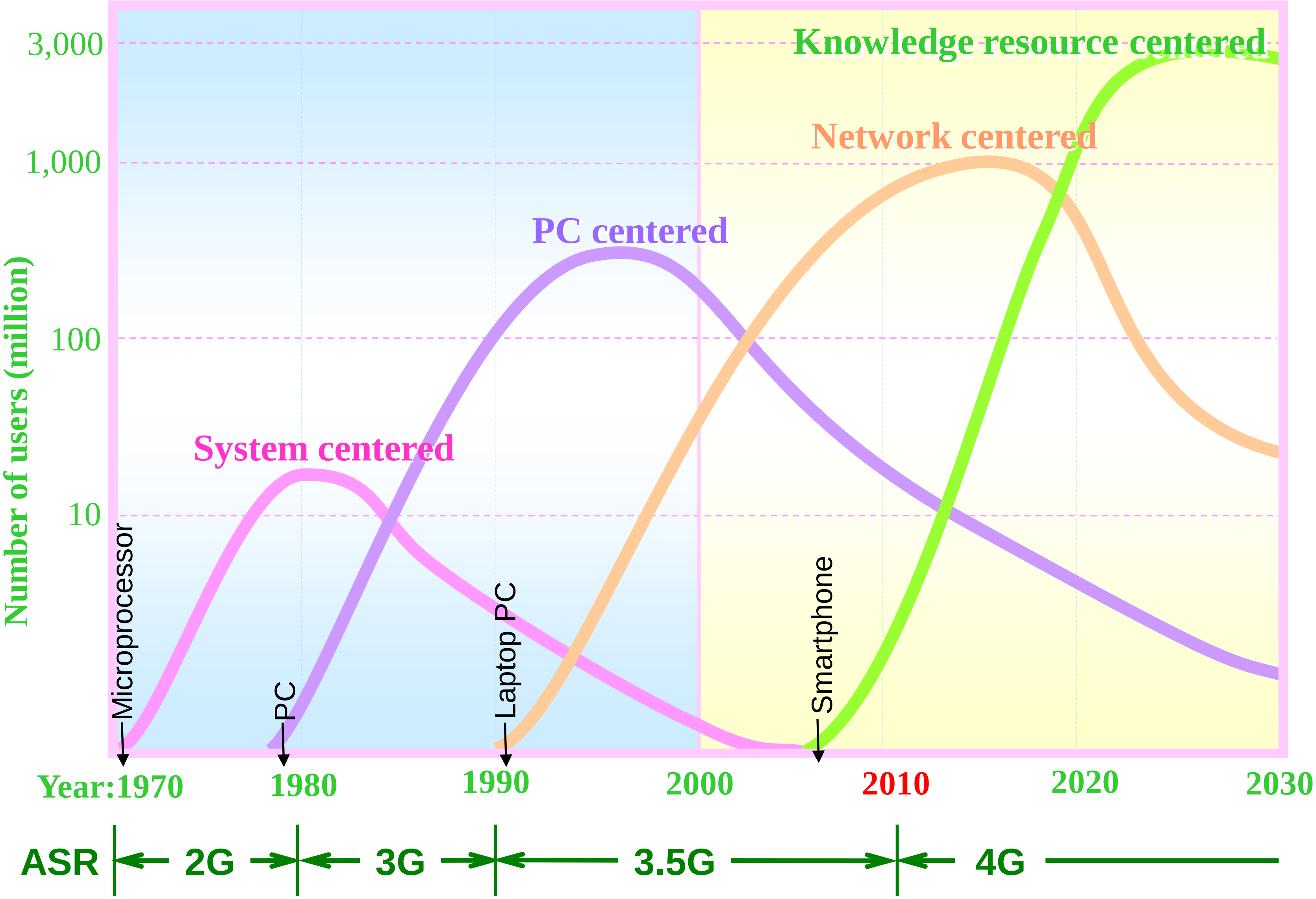
- Unsupervised or lightly supervised training is crucial



# Next-generation ASR using comprehensive knowledge sources

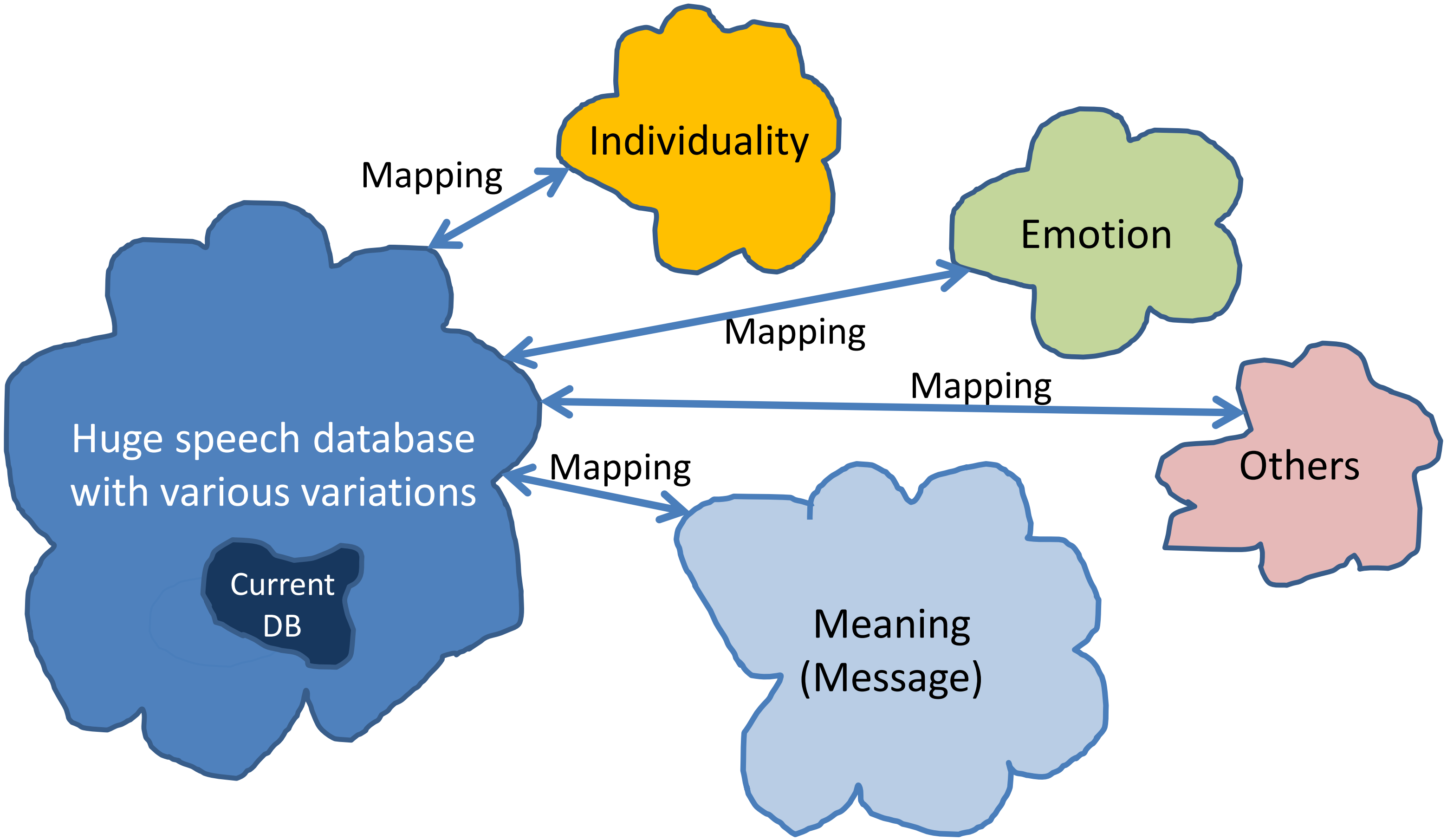


# IT technology progress



(David C. Moschella: “Waves of Power”)

# Data-intensive ("Big data") ASR



"We have actually made fair progress on simulation tools, but not very much on data analysis tools." (Jim Gray)

"Quantity decides quality."  "Resources are better spent collecting more data."

# Technical issues

- **How to collect rich and orders of magnitude larger speech DBs, covering all possible variations**
  - Current typical speech DB (for a single task)
    - 1000-10,000 h speech ~ 100 M - 1 Gbyte
  - Current model complexity (for a single task)
    - HMM : several million parameters
    - Number of Trigrams : 100 million - 1 billion
  - Many sources of variations
- **How to build and utilize huge DBs (Scalable approach)**
  - Cheap, fast and good enough metadata/transcription
  - Well structured model (machine learning)
  - Efficient data selection for annotation (active learning)
  - Unsupervised, semi-supervised or lightly-supervised training/adaptation
  - High-performance computing

# Speaker and noise factorization

In the Mel-cepstral domain, mismatch function relating clean speech $x$ and noisy speech $y$ is given by:

$$y = x + h + \mathrm{C} \log (1 + \exp (\mathrm{C}^{-1}(n - x - h)))$$

where

$n$ and $h$ : additive and convolutional noise, respectively

$\mathrm{C}$ : DCT matrix

The model is first adapted to the target speaker by MLLR, and then adapted to the target noise environment via model-based vector Taylor series (VTS) compensation. These transforms can be jointly estimated using ML.

(Y.-Q. Wang and M. J. F. Gales, 2011)

# Subspace mixture model

$$p(x \mid j)^{(s)} = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(x; \mu_{jmi}^{(s)}, \Sigma_i \ \ )$$

$$\mu_{jmi}^{(s)} = \mathrm{M}_i \mathrm{v}_{jm} + \mathrm{N}_i \mathrm{v}^{(s)}$$

$j$ : phonetic state index

$m$ : substate index

$c_{jm}$ : mixture weight

$i$ : Gaussian index

$w_{jmi}$ : mixture weight

$s$ : speaker

$\mathbf{M}_i$ : mean projection matrix

$\mathbf{v}_{jm}$ : state-substate-specific vector

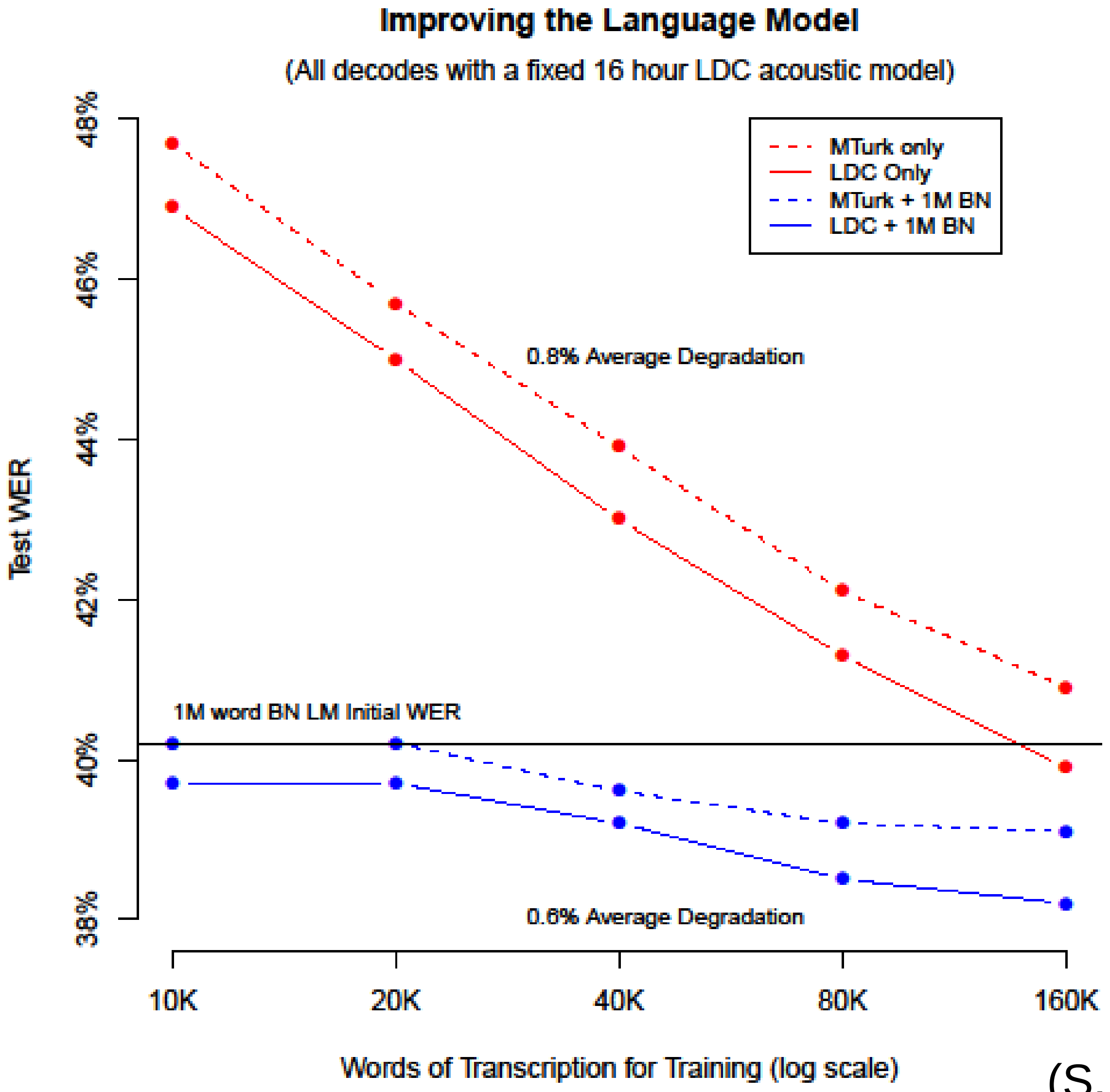
$\mathbf{N}_i\mathbf{v}^{(s)}$ : speaker-specific offset

$\mathbf{v}^{(s)}$ : speaker vector

Similar to Joint Factor Analysis, Eigenvoice, and Cluster AdaptiveTraining.

(D. Povey et al., 2010)

# Cheap, fast and good enough: non-expert transcription through crowdsourcing



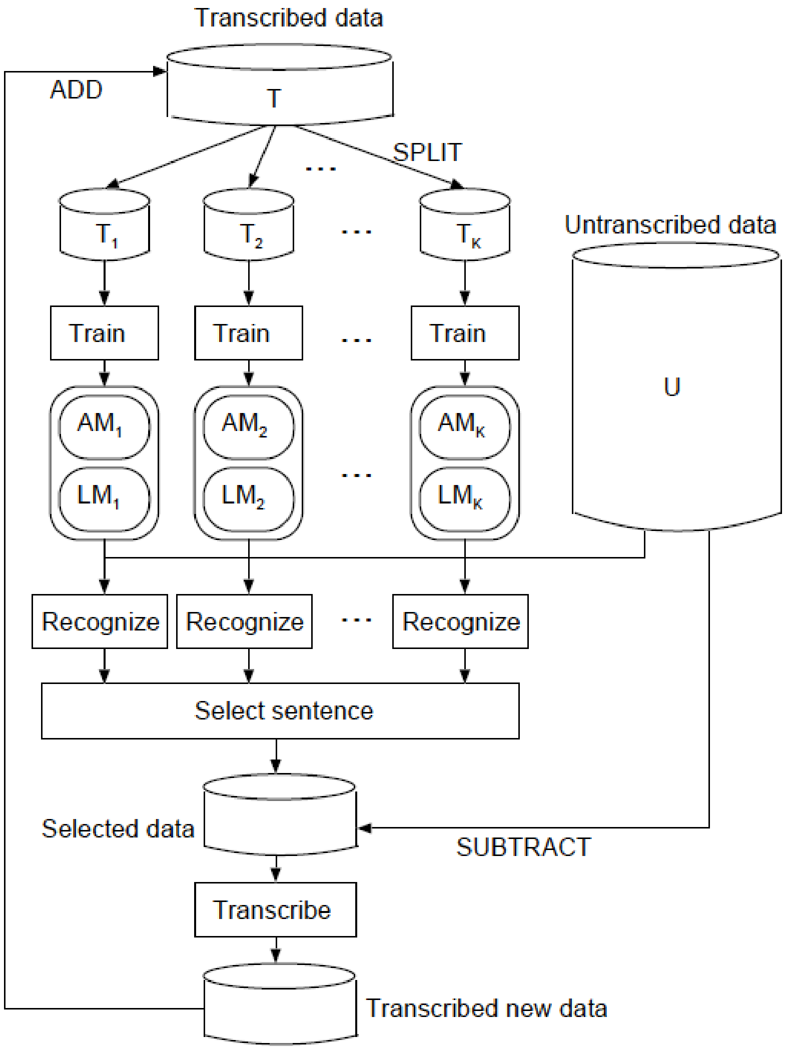
(S. Novotney et al., 2010)

# Efficient data selection for annotation (Active learning)

- Select most informative utterances, having
  - Low certainty (word posterior probability, confidence score, relative likelihood, margin) (Hakkani-Tur et al., 2002)
  - Large word lattice entropy (Varadarajan et al., 2009)
  - Large classification disagreement among committees (query by committee) (Seung et al., 1992; Hamanaka et al., 2010)
- Manually transcribe selected utterances and use them for model training
- Problem: how to remove outliers
  - Density-based approaches
- Combine semi-supervised training for remaining utterances (Riccardi et al., 2003)

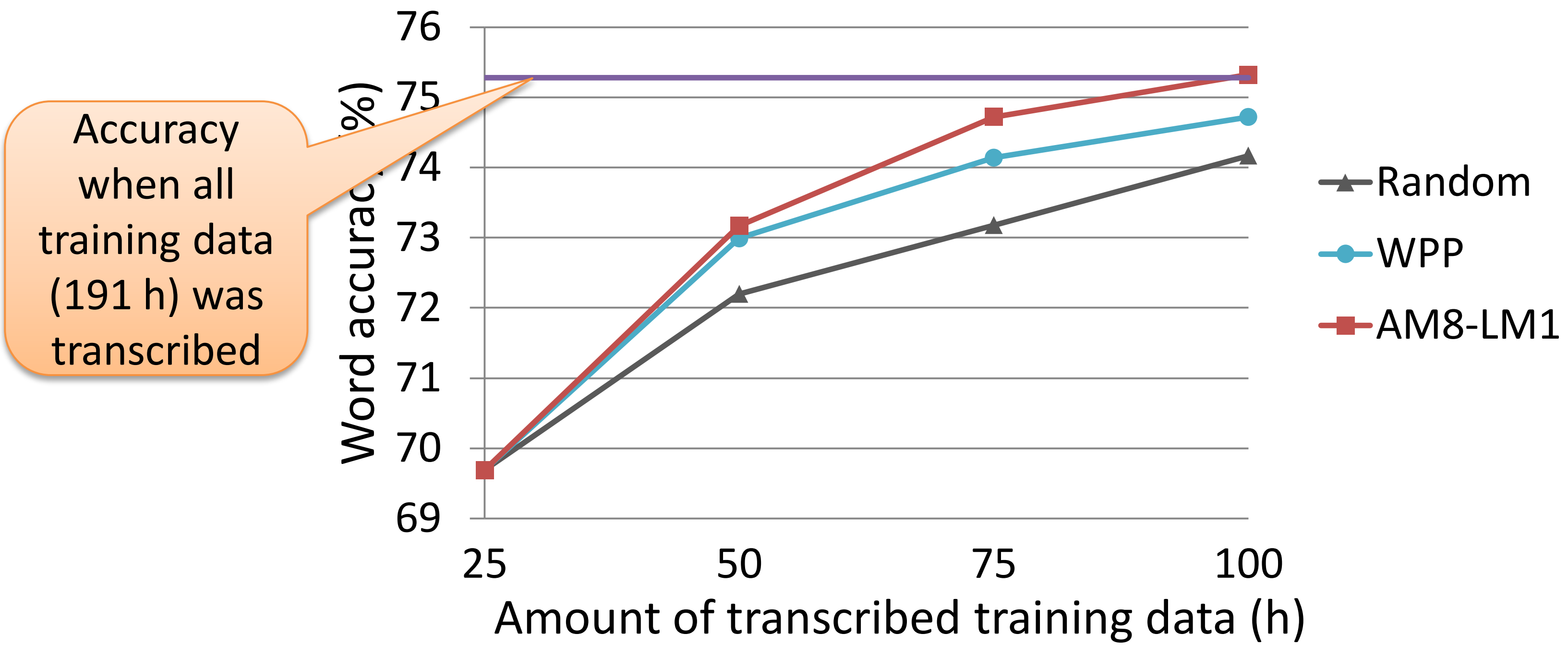
# Committee-based utterance selection

1. Randomly and equally divide transcribed data $T$ into $K$ data sets
2. Train $K$ recognizers (committees) using the $K$ data sets
3. Recognize all the utterances in the un-transcribed data $U$ by each recognizer
4. Select $N$ hour utterances with relatively high degree of disagreement (vote entropy) among $K$ recognizers
5. Transcribe the selected utterances and add them to $T$, and go back to Step 1.



# Comparison with other methods

- Random selection: randomly select utterances to be transcribed
- WPP (word posterior probability)-based method
- Committee-based method (8 and 1 committees for AM and LM, respectively)

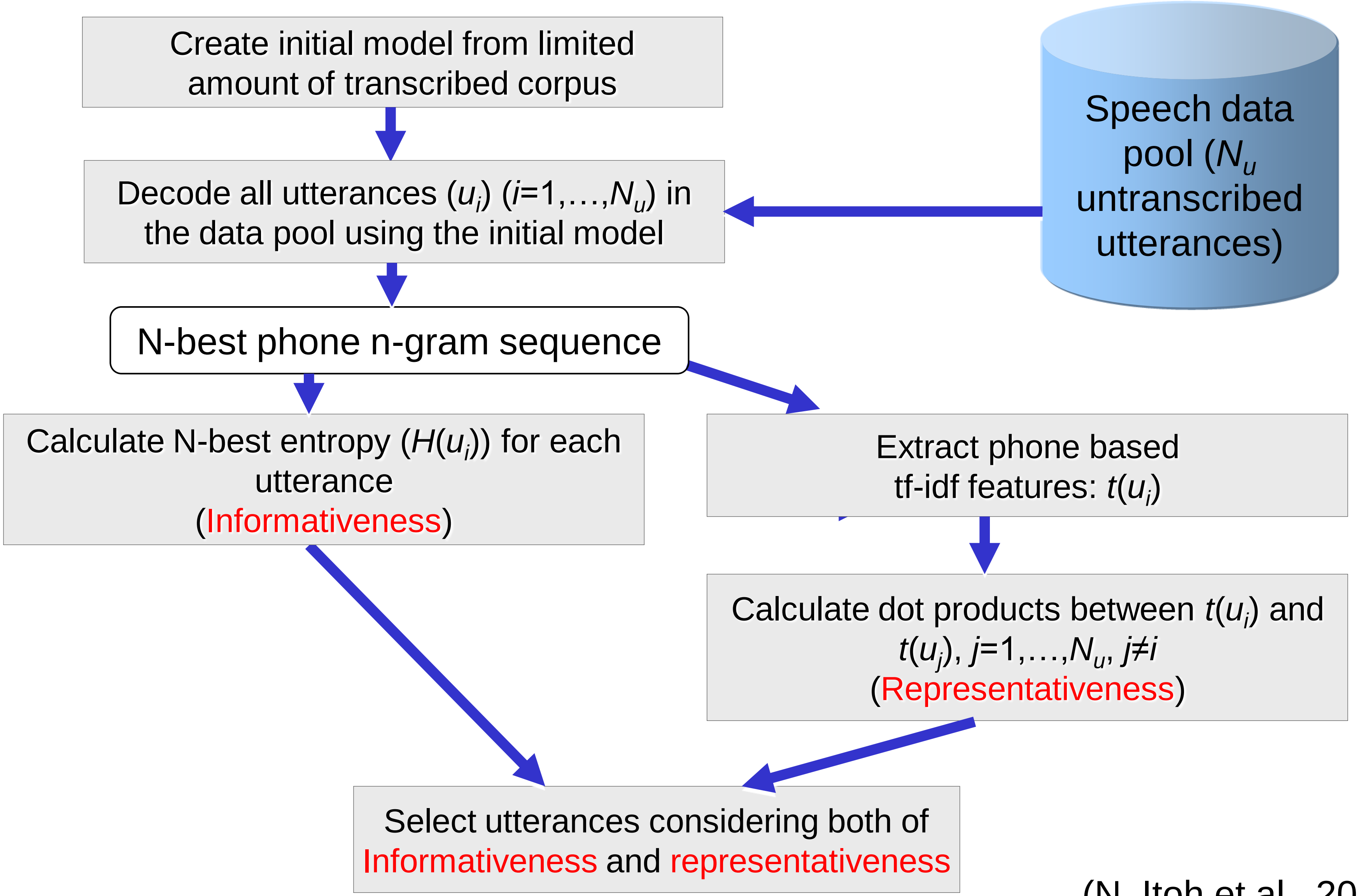


Accuracy when all training data (191 h) was transcribed

The amount of data $T$ to achieve a word accuracy of 74%:
Random: 95 h, WPP-based: 72 h, Committee-based: 63 h

(Database: Corpus of Spontaneous Japanese)

# Combination of informativeness and representativeness



Create initial model from limited amount of transcribed corpus

Speech data pool ($N_u$ untranscribed utterances)

Decode all utterances ($u_i$) ($i$=1,…,$N_u$) in the data pool using the initial model

N-best phone n-gram sequence

Calculate N-best entropy ($H(u_i)$) for each utterance (Informativeness)

Extract phone based tf-idf features: $t(u_i)$

Calculate dot products between $t(u_i)$ and $t(u_j)$, $j$=1,…,$N_u$, $j \neq i$ (Representativeness)

Select utterances considering both of Informativeness and representativeness

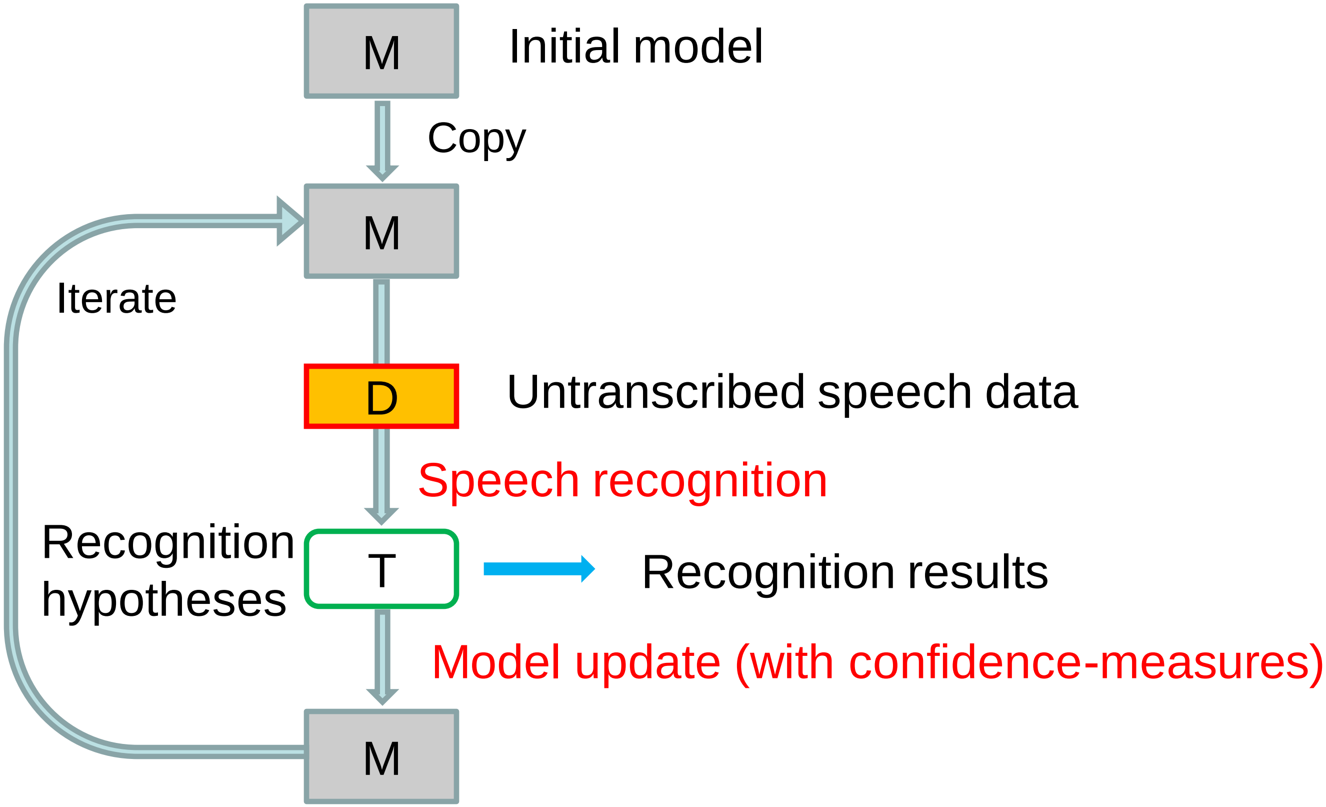
(N. Itoh et al., 2012)

# Lightly-supervised AM training for broadcast news ASR

- BN has closed-captions which are close, but not exact transcription, and only coarsely time-aligned with the audio signal.
- A speech recognizer is used to automatically transcribe unannotated BN data. The closed-captions are aligned with transcriptions and only segments that agree are selected for training.
- Using the closed captions to provide supervision via the LM is sufficient, and there is only a small advantage in using them to filter the “incorrect” hypotheses/words.

(L. Lamel et al, 2000)

# Semi-supervised batch-mode adaptation



- Run a decoder to obtain recognition hypotheses
- Use the hypotheses as reference for adaptation (eg. MLLR)

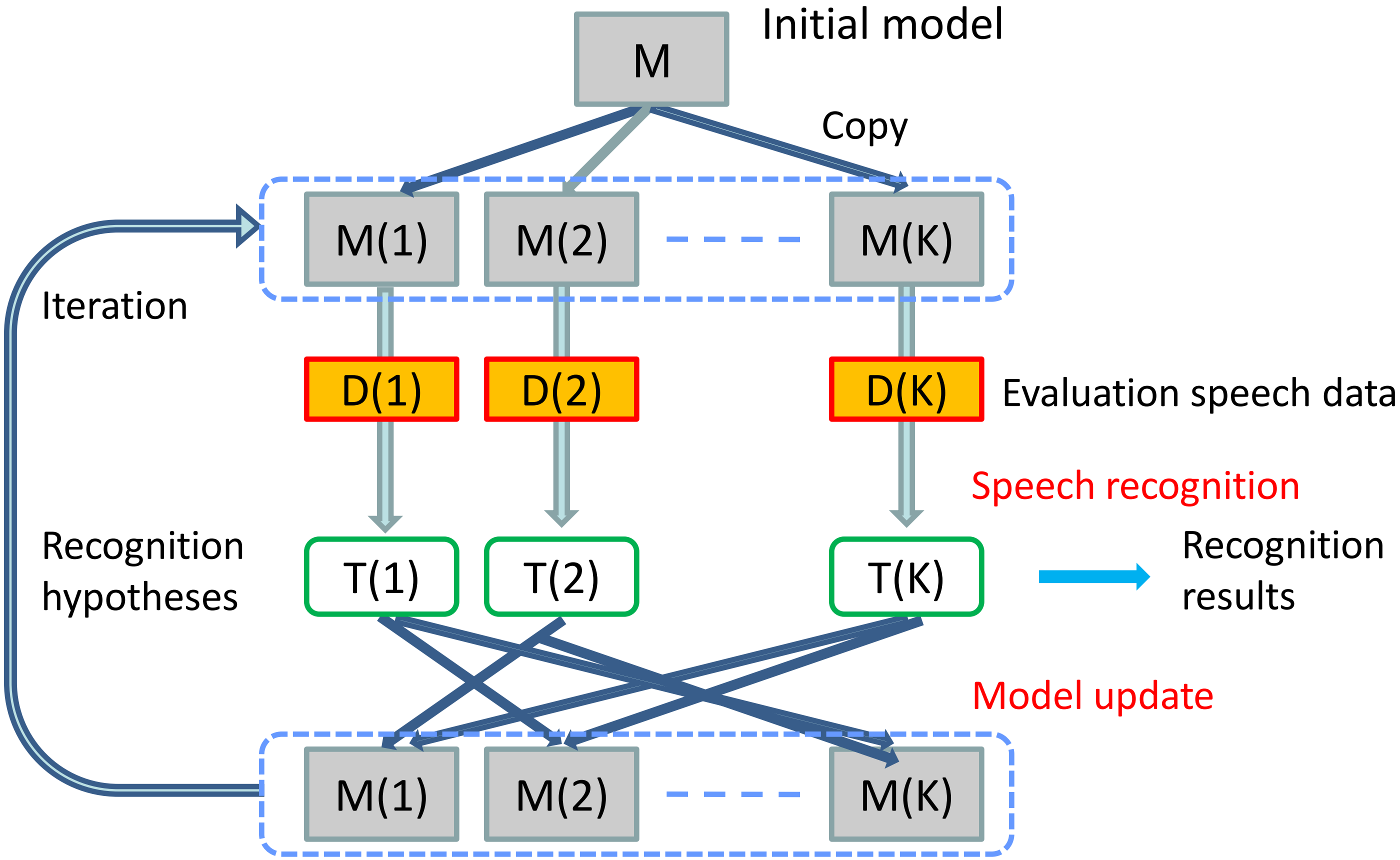
# Problems of the batch-mode adaptation

- Errors are unavoidable in the recognition
- Model parameters are estimated using the hypotheses including errors
- In the next decoding step, the adapted model tends to make the same errors
- During the iterations, errors are reinforced

↓

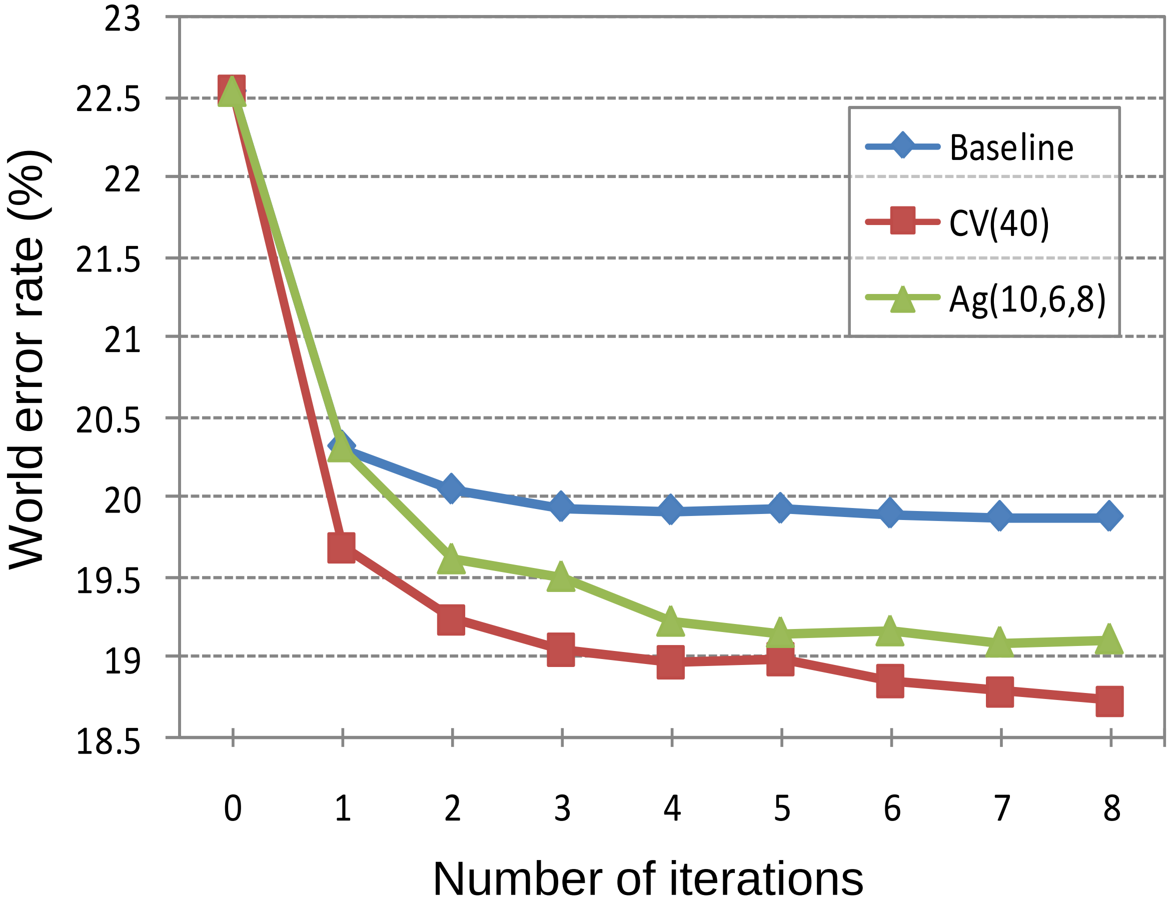
How to improve the adaptation performance by reducing the influence of the recognition errors

# Semi-supervised cross-validation (CV) adaptation



- Reducing the influence of recognition errors by separating the data used for the decoding step and the model update step

# Number of iterations and WER



Relative WER reductions

Batch-adapt : 12%

CV-adapt : 17%

Ag-adapt : 16%

CV:

K=40

Ag:

K=10, K’=6, N=8

# Summary

- Speech recognition technology has made very significant progress in the past 30 years with the help of computer technology.
- The majority of technological changes have been directed toward the purpose of increasing robustness of recognition.
- Major successful applications are spoken dialog and transcription systems.
- Much greater understanding of the human and physical speech processes is required before automatic speech recognition systems can approach human performance.
- Significant advances will come from data-intensive approach for knowledge extraction ("Big data" ASR).
- Active learning, and unsupervised, semi-supervised or lightly-supervised training/adaptation technologies are crucial (Machine learning).