Some Representation Learning Approaches to Automatic Speech Recognition and its Applications

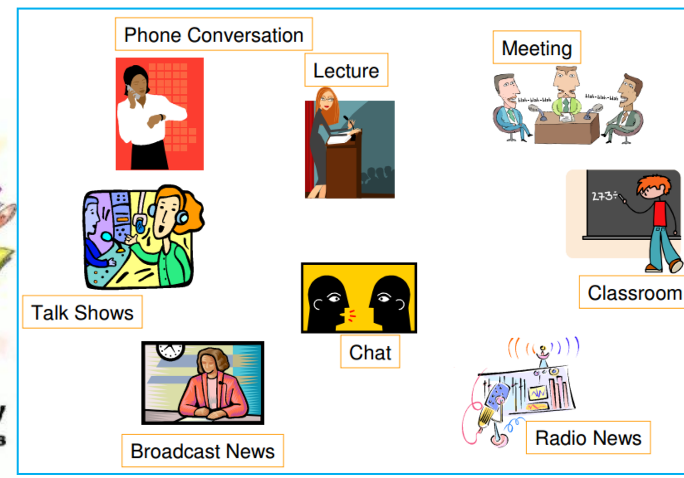
Berlin Chen (陳柏琳)
Professor, Department of Computer Science & Information Engineering
National Taiwan Normal University

Big Data Era – Information Overload

- *Too much information kills information!*

Written text

Speech, Audio, Image, Video, etc.



Some of the above figures are adapted from the presentation slides of Prof. Nenkova et al. at *ACL 2011*

Outline

- Introduction
- Machine Learning
- Automatic Speech Recognition (ASR)
- (Shallow & Deep) Representation Learning for ASR and its Applications
- Conclusions

Introduction (1/3)

- Communication and search are by far the most popular activities in our daily lives
  - Speech is the most nature and convenient means of communication between humans (and between humans and machines in the future)
    - A spoken language interface could be more convenient than a visual interface on a small device
    - Provide "*anytime*" and "*anywhere*" access to information
  - Already over half of the internet traffic consists of video data
    - Though visual cues are important for search, the associated spoken documents often provide a rich set of semantic cues (e.g., transcripts, speakers, emotions, and scenes) for the data

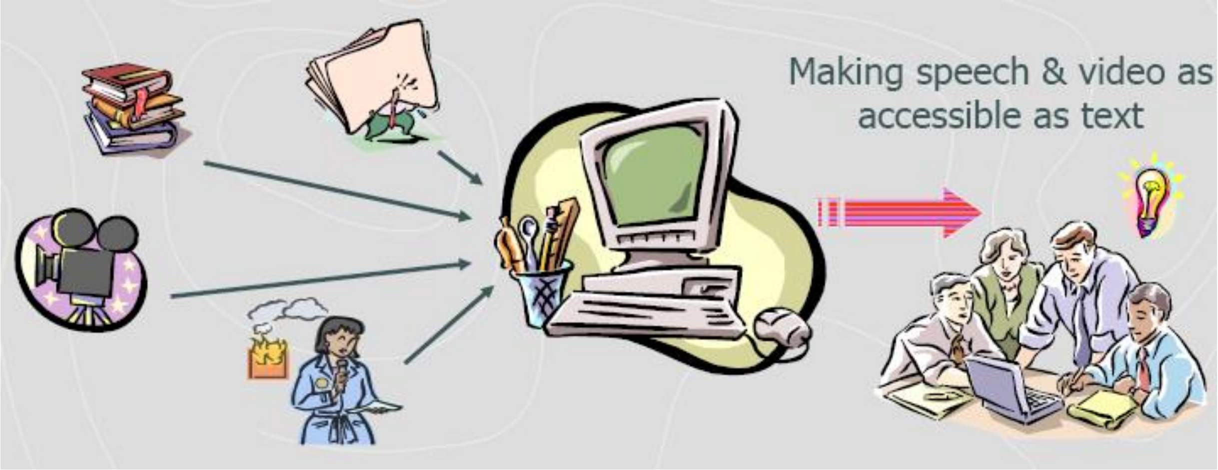
Tur and Mori, *Spoken language understanding – systems for extracting Semantic Information from speech*, Wiley 2011.

Introduction (2/3)

- Text Processing vs. Speech Processing
  - Recognition, Analysis and Understanding
    - **Text**: analyze and understand text
    - **Speech**: recognize speech (i.e., ASR), and subsequently analyze and understand the recognized text (propagations of ASR errors)
  - Variability
    - **Text**: different synonyms to refer to a specific semantic object or meaning, such as 台灣師範大學, 師大, 教育界龍頭, etc.
    - **Speech**: an infinite number of utterances with respect to the same word (e.g., 台灣師範大學)
      - Manifested by a wide variety of oral phenomena such as disfluences (hesitations), repetitions, restarts, and corrections
      - Gender, age, emotional and environmental variations further complicate ASR
      - No punctuation marks (delimiters) or/and structural information cues exist in speech

Introduction (3/3)

- Automatic Speech Recognition (**ASR**) or **Speech to Text**
  - Transcribe the linguistic contents of speech utterances
  - Play a vital role in multimedia information retrieval, summarization, organization, among others
    - Such as the transcription of spoken documents and recognition of spoken queries



The figure is adapted from the presentation slides of Prof. Ostendorf at *Interspeech 2009.*

Spectrum of **Machine Learning** Research

Training Data

- Supervised Learning (Labeled data)
- Semi-supervised Learning (Labeled and unlabeled data)
- Unsupervised
- Active Learning (Selectively labeled data)

Data (Input) Representation

- Dense Features
- Sparse Features
- Deep Learning for Multiple layers of Non-linearity

Evaluation Metrics

- Extrinsic
- Intrinsic

Training Criteria

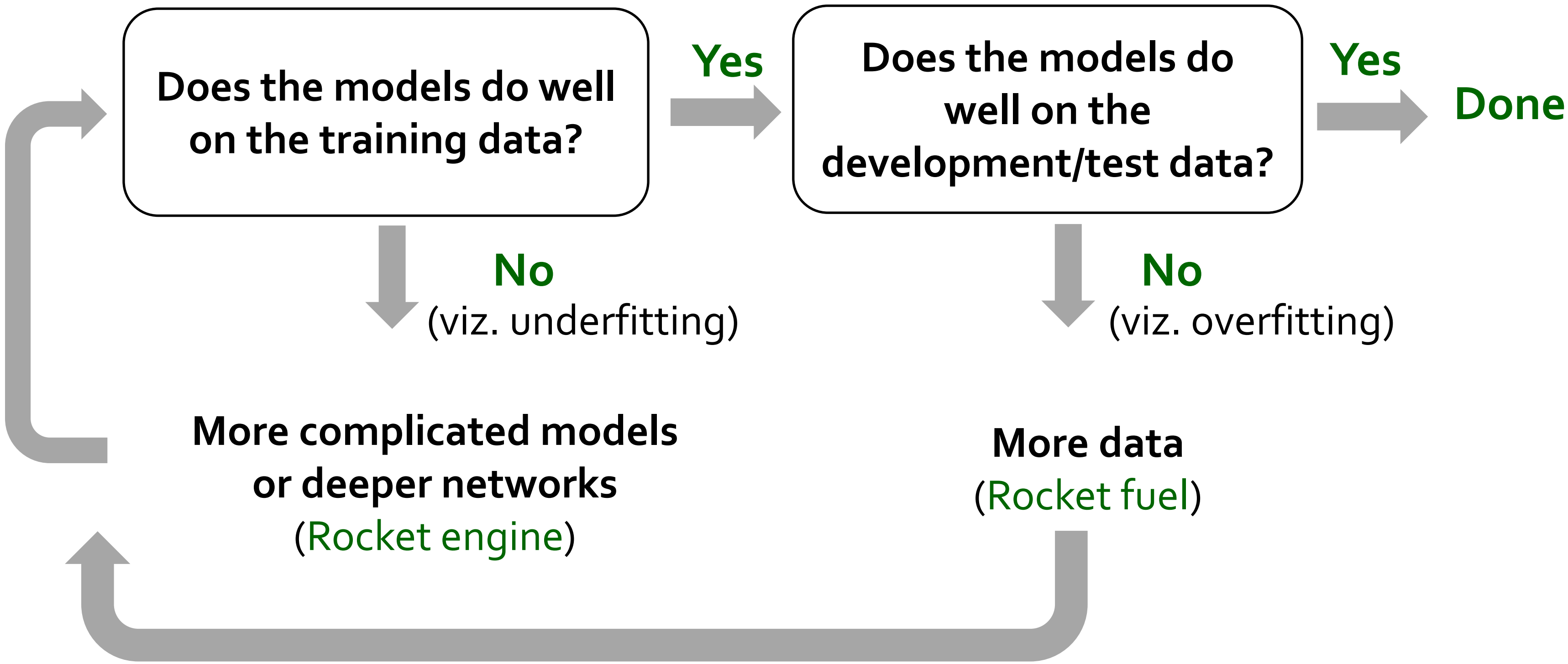
- Maximum Likelihood (Generative Learning)
- Maximum Discrimination (Discriminative Learning)
- Maximum Task Performance

Source and Target Distributions

- Single-Task Learning
- Model Adaptation
- Multi-Task Learning

The figure is adapted from the presentation slides of Dr. Li Deng at *Interspeech 2015*.

Typical Recipe for **Machine Learning** Research



There is no data like more data!

The figure is adapted from the presentation slides of Dr. Li Deng at *Interspeech 2015*.

Automatic Speech Recognition (ASR)

- Bayes Decision Rule (Risk Minimization)

$$W_{opt} = \arg\min_{W \in \mathbf{W}} Risk(W|O)$$

$$= \arg\min_{W \in \mathbf{W}} \sum_{W' \in \mathbf{W}} Loss(W, W') P(W'|O)$$

$$\approx \arg\max_{W \in \mathbf{W}} P(W|O)$$

Assumption: Using the "0-1" Loss Function
(Become a Typical **Maximum-a-Posteriori** Classification Problem)

$$= \arg\max_{W \in \mathbf{W}} \frac{p(O|W)P(W)}{p(O)}$$

$$= \arg\max_{W \in \mathbf{W}} p(O|W)P(W)$$

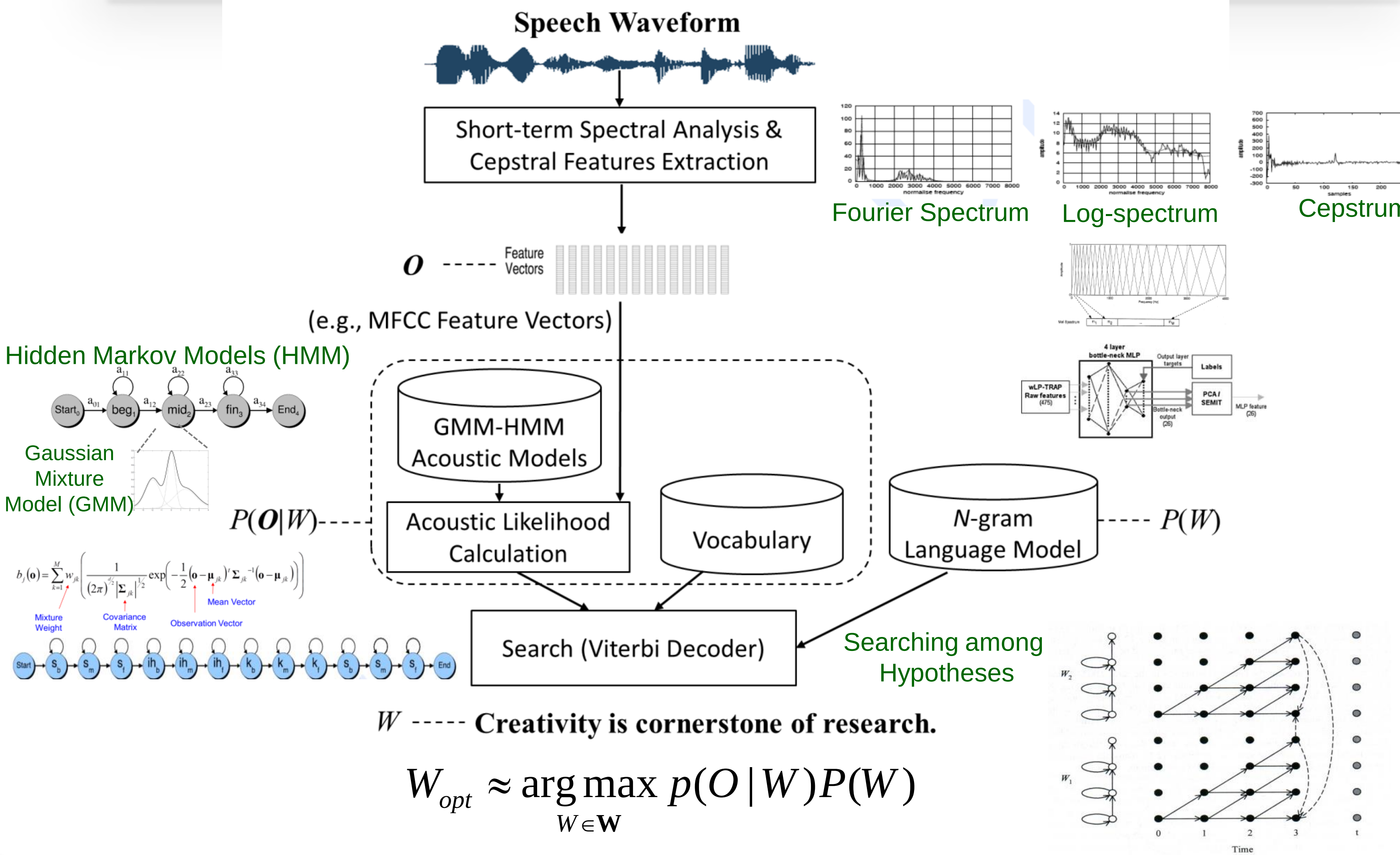
Linguistic Decoding

Feature Extraction & Acoustic Modeling — Language Modeling

Possible variations: speaker, pronunciation, environment, context, etc. and domain, topic, style, etc.

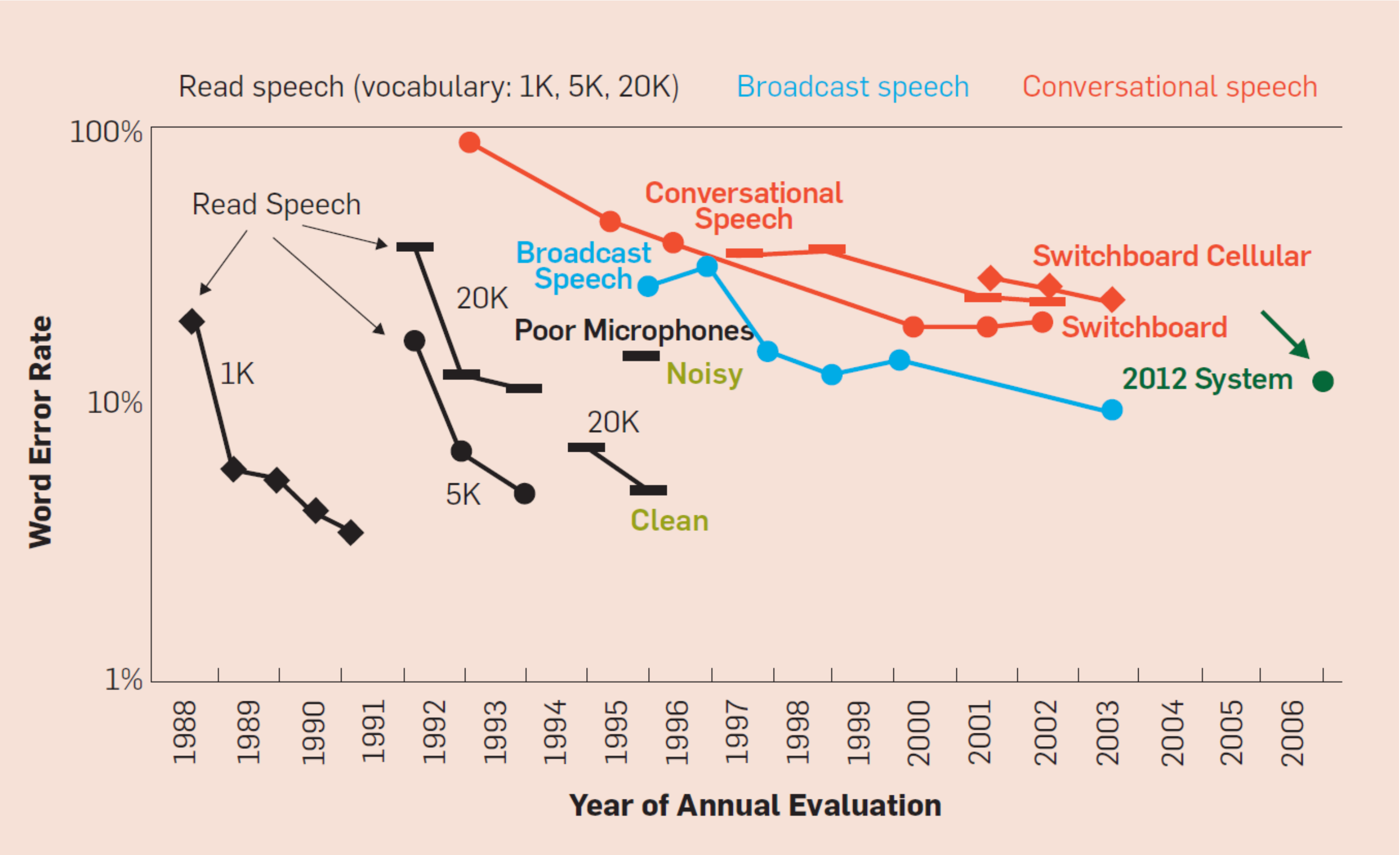
1. F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, 1999
2. X Huang, J. Backer, R. Reddy, "*A historical perspective of speech recognition*," *ACM Communications*, 2004

Schematic Diagram of ASR



$$W_{opt} \approx \arg\max_{W \in \mathbf{W}} p(O|W)P(W)$$

Historical Progress of ASR



X Huang, J. Backer, R. Reddy, "*A historical perspective of speech recognition,*" *ACM Communications*, 2014

What is Deep Learning?

WIKIPEDIA
The Free Encyclopedia

Deep learning

From Wikipedia, the free encyclopedia

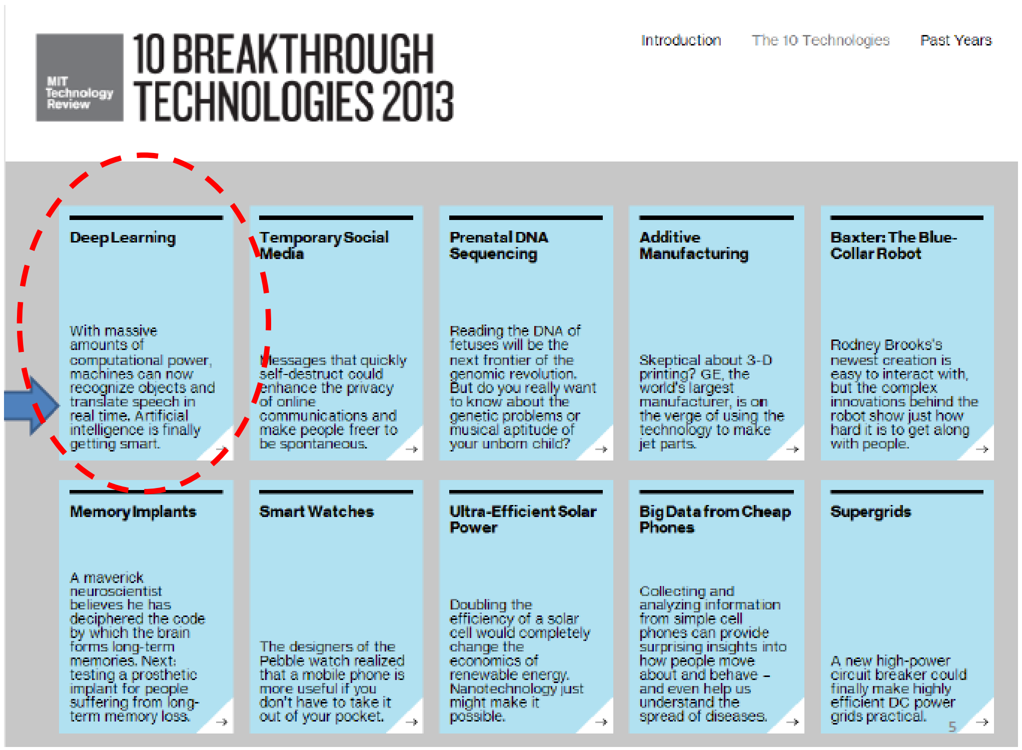
Deep learning (*deep machine learning*, or *deep structured learning*, or *hierarchical learning*, or sometimes *DL*) is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers with complex structures or otherwise, composed of multiple non-linear transformations.[1](p198)[2][3][4][5]

Representation Learning →
- **Shallow Learning** GMM, SVM, CRF, NMF/PLSA/LDA, Perception, Boosting, etc.
- **Deep Learning** DNN, CNN, RNN, LSTM, etc.
  (help to discover intricate structure in large data sets)

Deeper is better? vs. Simple is elegant?

A Surge of Research on Deep Learning (1/2)

- Our computers can learn and grow on their own
- Our computers are able to understand complex, massive amount of data (deep learning serves as a good foundation for effectively leveraging big data)



1. http://www.technologyreview.com/lists/breakthrough-technologies/2013/
2. Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, 521, pp. 436-444, 2015

A Surge of Research on Deep Learning (2/2)

MIT Technology Review

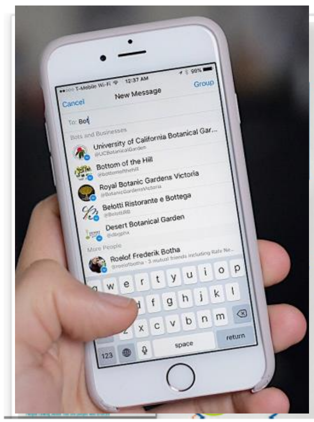
September 20, 2013

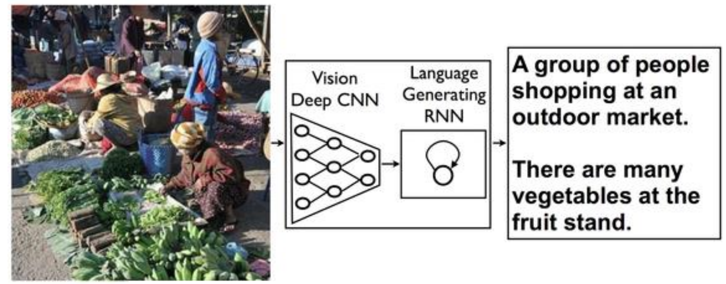
Facebook Launches Advanced AI Effort to Find Meaning in Your Posts

A technique called deep learning could help Facebook understand its users and their data better.

By Tom Simonite on September 20, 2013

……Facebook's foray into deep learning sees it following its competitors Google and Microsoft, which have used the approach to impressive effect in the past year. Google has hired and acquired leading talent in the field (see "10 Breakthrough Technologies 2013: Deep Learning"), and last year created software that taught itself to recognize cats and other objects by reviewing stills from YouTube videos. The underlying deep learning technology was later used to slash the error rate of Google's voice recognition services (see "Google's Virtual Brain Goes to Work")….Researchers at Microsoft have used deep learning to build a system that translates speech from English to Mandarin Chinese in real time (see "Microsoft Brings Star Trek's Voice Translator to Life"). Chinese Web giant Baidu also recently established a Silicon Valley research lab to work on deep learning.





A woman is throwing a **frisbee** in a park.

A little **girl** sitting on a bed with a teddy bear.

X. He,et al., "Deep learning for natural language processing and related applications," Tutorial given at ICASSP 2014.

Deep Learning for Acoustic Modeling in ASR (1/4)

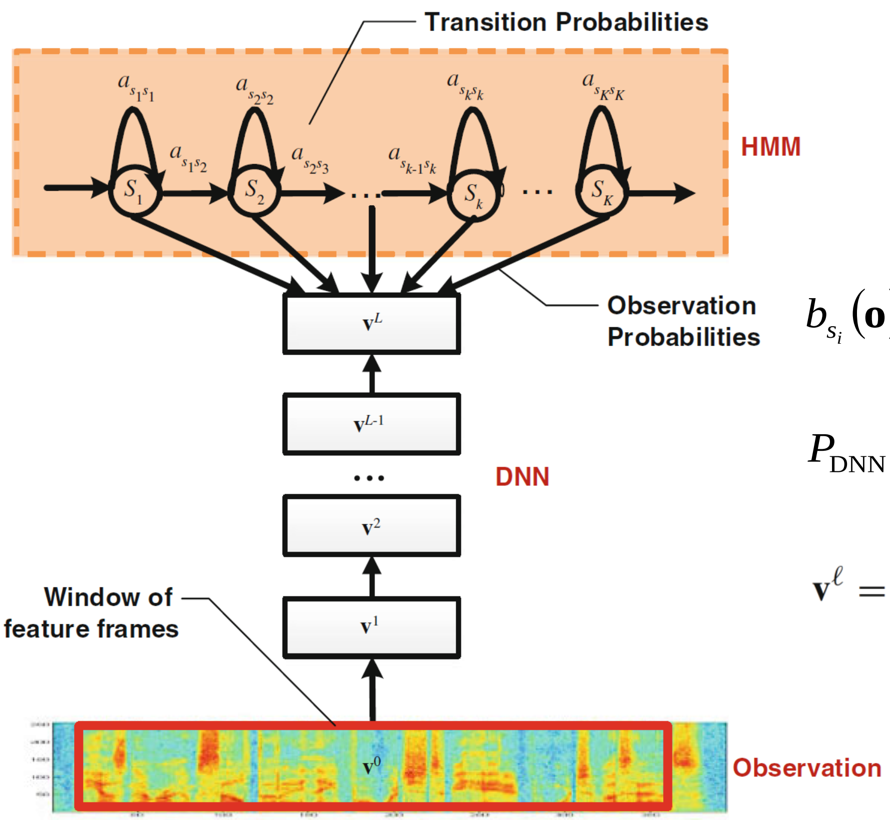
- **Deep Learning** is the cutting edge for acoustic modeling

- Dr. Li Deng pointed out that there are three major factors for the recent success of deep learning in ASR
  1. Remove modeling of dynamics by using a long time window to approximate the true effects of dynamics
  2. Reverse the direction of information flow in the deep models: **from top-down** as in the deep generative models **to bottom-up** as in the DNN
  3. Bypass the difficulty to train a DNN with many hidden layers: using Restricted Boltzmann Machines (RBM) or Deep Belief Networks (DBN) to initialize or pre-train the DNN

L. Deng , "Deep learning: from speech recognition to language and multimodal processing," APSIPA Transactions on Signal and Information Processing, January 2016

Deep Learning for Acoustic Modeling in ASR (2/4)

- **Deep Learning** is the cutting edge!
  - E.g., Leveraging **Deep Neural Networks (DNN)** for Feature Extraction and Acoustic Modeling (Context-Dependent **DNN-HMM**)



deeper layers,
longer features &
wider temporal contexts

$$b_{s_i}(\mathbf{o}) = p(\mathbf{o} \mid s_i) = \frac{P_{\mathrm{DNN}}(s_i \mid \mathbf{o})\, p(\mathbf{o})}{P_{\mathrm{ML}}(s_i)} \propto \frac{P_{\mathrm{DNN}}(s_i \mid \mathbf{o})}{P_{\mathrm{ML}}(s_i)}$$

$$P_{\mathrm{DNN}}(s_i \mid \mathbf{o}) = v_i^L = \mathrm{softmax}_i\left(\mathbf{z}^L\right) = \frac{e^{z_i^L}}{\sum_j e^{z_j^L}}$$

$$\mathbf{v}^\ell = f\left(\mathbf{z}^\ell\right) = f\left(\mathbf{W}^\ell \mathbf{v}^{\ell-1} + \mathbf{b}^\ell\right), \ \text{for } 0 < \ell < L$$

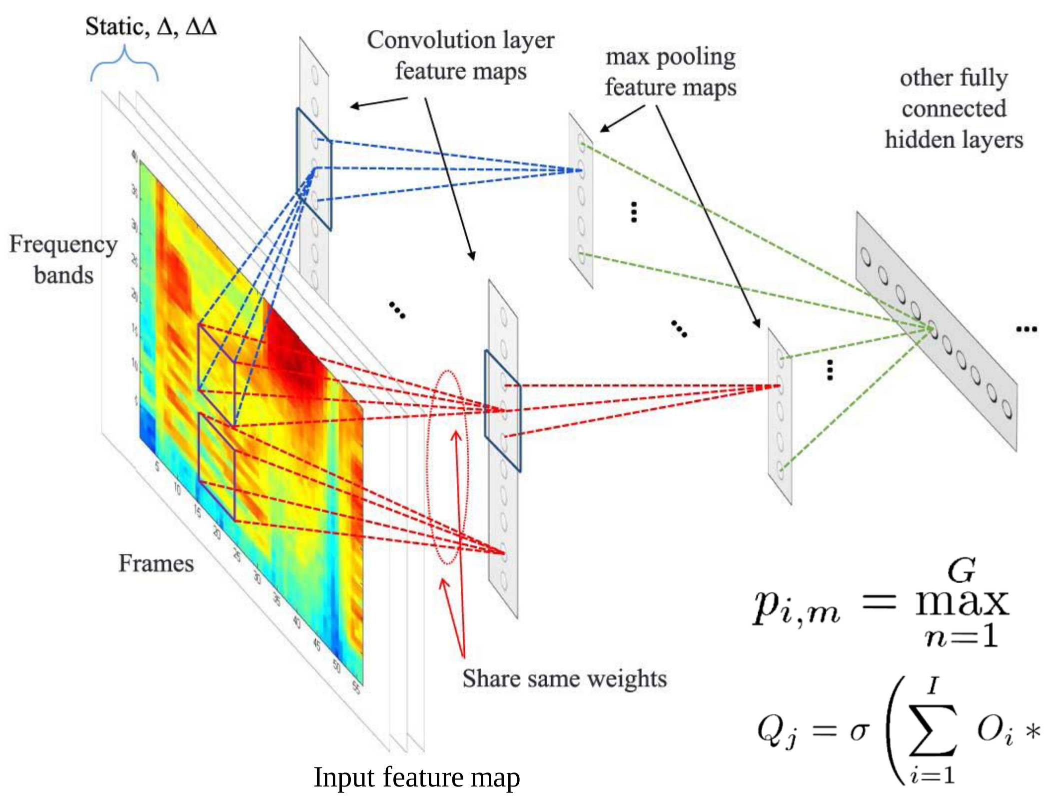
$f(\cdot)$: sigmoid, hyperbolic, or rectified linear unit (ReLU) functions

Model parameters of DNN can be estimated with **the error back-propagation** algorithm and **stochastic gradient decent** (SGD).

Dong Yu and Li Deng, "Automatic Speech Recognition: A Deep Learning Approach," Springer, 2015

Deep Learning for Acoustic Modeling in ASR (3/4)

- **CNN-HMM**
  - CNN: Convolutional Neural Networks



$$p_{i,m} = \max_{n=1}^{G} \quad q_{i,(m-1)\times s+n}$$

$$Q_j = \sigma\left(\sum_{i=1}^{I} O_i * \mathbf{w}_{i,j}\right) \qquad (j = 1, \ldots, J)$$

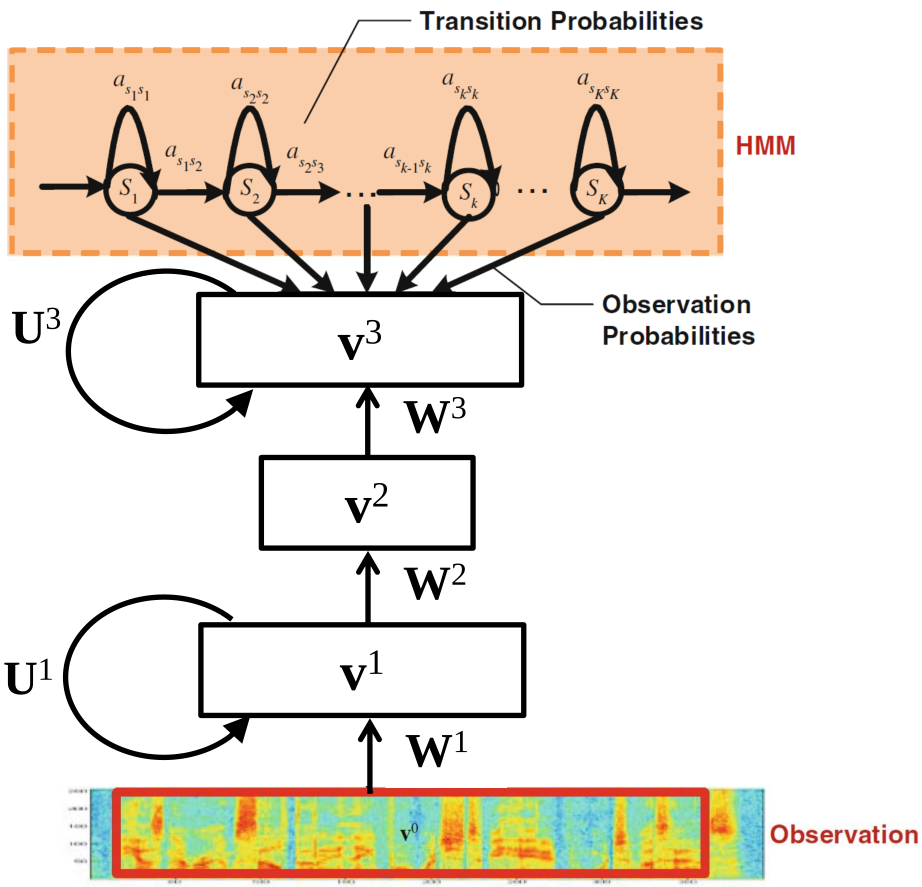
Convolution Layer:
- Locality: deal with noise
- Weight Sharing: facilitate model training

Pooling Layer:
- Maximum Pooling: less vulnerable to spectral and temporal varieties

Abdel-Hamid et al., Convolutional Neural Networks for Speech Recognition, IEEE/ACM Transactions on Adio, Speech, and Language Processing, Vol. 22, No. 10, 2014

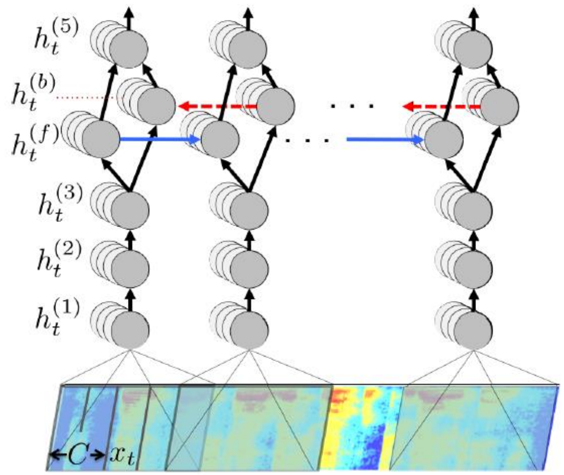
Deep Learning for Acoustic Modeling in ASR (4/4)

- Recurrent Neural Networks (**RNN-HMM**)



Baidu Research Approach

Output: Text (Letters, Words, ….)



Structure of our RNN model and notation.

A. Hannun et al. (Lead by Andrew Ng), "Deep Speech: Scaling up end-to-end speech recognition," arXiv:1412.5567v2, December 2014.

Automatic Meeting Transcription (1/2)

Manual Transcripts

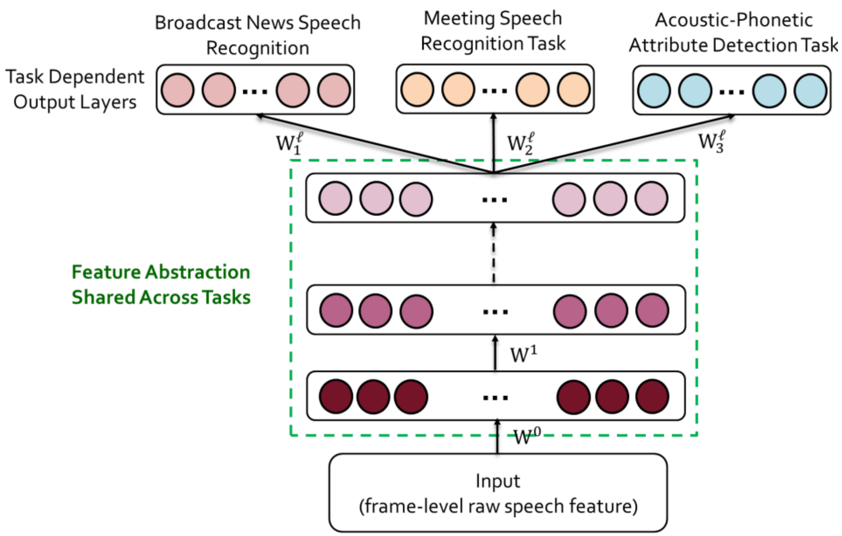
A: 那 會 在 二 a 那個 那 叫 什麼 二 b 啊 二 a
A: vip vip room
B: 欸
A: 就是 大家 開 all hands meeting 那裡
C: 錄音 的話 就 只 能 用 八爪魚 喔
A: 錄音 就 對 啊 那 場 就 反正 錄下 來 就 好了 對
A: 好 一 開始
D: 請問 一下
D: 上 次 二 a. 的 時候 那個 圓 方 不是 有 來 教 我們 怎麼 用 八爪魚 錄音 所以 那個 測試 設 定都 沒有 動
D: 就 直接 麥克風 可以 把 聲音 收進 來
A: 圓 圓形 會議 對 啊 圓形 會議 是 這樣
D: 好 好
A: 可是 我們 這 一 次 不是 在 圓形 我們 這 次 是 在 呃 vip
A: 就是 董事長 開會 的 地方

Automatic Transcripts

A: 那 會 在 二 a. h 那個 資料 怎麼 二 的 啊 把 二 a.
A: 七 vip 喔 vip vip room
B: 嘿
A: 可是 打開 過 hand meeting 那裡
C: 錄音 的話 是 怎麼 用 滑動 語料
A: 錄音 就 對 啊 那 一 場 就 反正 錄下 就 好了
A: 好 一 開始 了
D: 請問 一下
D:上 是 二月 的 時候 那個 員工 不是 來 教 我們 怎麼 跟 八爪魚 錄音 最 那個 測試 設定 檔 秒 鐘
D: 就 支 麥克風 可以 把 聲音 投 進來
A: 每 圓形 會議 對 啊 圓形 會議室 這樣
D: 好
A: 可是 我們 這 次 不是 在 圓形 我們 這 次 是 在 edge vip
A: 是 董事會 開會 的 地方

Automatic Meeting Transcription (2/2)

- Acoustic Modeling with Multitask Learning (MTL)
  - (A) Mono-Senones
  - (B) Multilingual Information
  - (C) Context State Label
  - (D) Context Phone Label
  - (E) Dark Knowledge



	Worr Error Rate, WER (%)	Character Error Rate, CER (%)	# Layers	# Neurons per Layer
GMM-HMM	58.71	51.88	-	-
DNN-HMM	43.20	36.45	6	2,048
LSTM-HMM	44.82	38.10	LSTM*3	1024
CNN-DNN-HMM	**42.20**	**35.60**	CNN*2+DNN*4	2,048
DNN-HMM+MTL(A)	45.87	39.42	6	2,048
DNN-HMM+MTL(B)	42.97	35.93	6	2,048
DNN-HMM+MTL(C)	45.89	38.83	6	2,048
DNN-HMM+MTL(D)	45.51	38.33	6	2,048
DNN-HMM+MTL(E)	42.72	35.91	6	2,048

1. G. E. Hinton, et al., “Distilling the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, 2015
2. J.W. Hung et al., "Robust speech recognition via enhancing the complex-valued acoustic spectrum in modulation domain," IEEE/ACM Transactions on Audio, Speech, and Language Processing, February 2016.

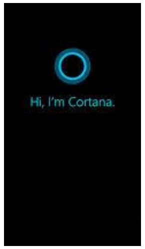
Some Applications of ASR

- Multimedia (spoken document) retrieval and organization
  - Speech-driven Interface and multimedia content processing
  - Work in concert with natural language processing (NLP) and information retrieval (IR) techniques
  - A wild variety of potential applications (to be introduced later)
- Computer-Aided Language Learning (CALL)
  - Speech-driven Interface and multimedia content processing
  - Work in in association with natural language processing techniques
  - Applications
    - Synchronization of audio/video learning materials
    - Automatic pronunciation assessment/scoring
    - Read student essays and grade them
    - Automated reading tutor
- Others

IEEE Signal Processing Magazine 25(3), 2008 (*Spoken Language Technology*)
IEEE Signal Processing Magazine 22(5), 2005 (*Speech Technology and Systems in Human-Machine Communication*)

Speech-based Multimedia Retrieval , Organization, Question Answering, Machine Translation

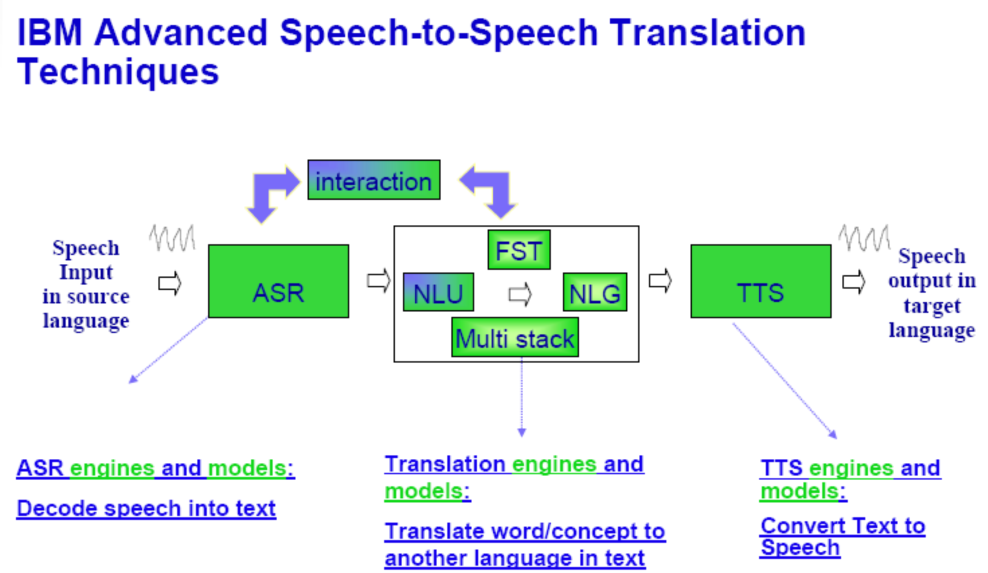
- Continuous and substantial efforts have been paid to speech-driven multimedia retrieval and organization in the recent past
  - *Informedia* System at Carnegie Mellon Univ.
  - MIT Lecture Browser
  - IBM Speeh-to-Speech Translation, ***Waston*** (QA)
  - Google Voice Search (*GOOG-411*, *Audio Indexing*, *Translation*), *Google Now*
  - Apple's ***Siri*** (QA)
  - Microsoft ***Cortana*** (QA), ***Skype Translator***
  - Amazon ***Echo*** (QA)
  - Facebook ***chatbot***

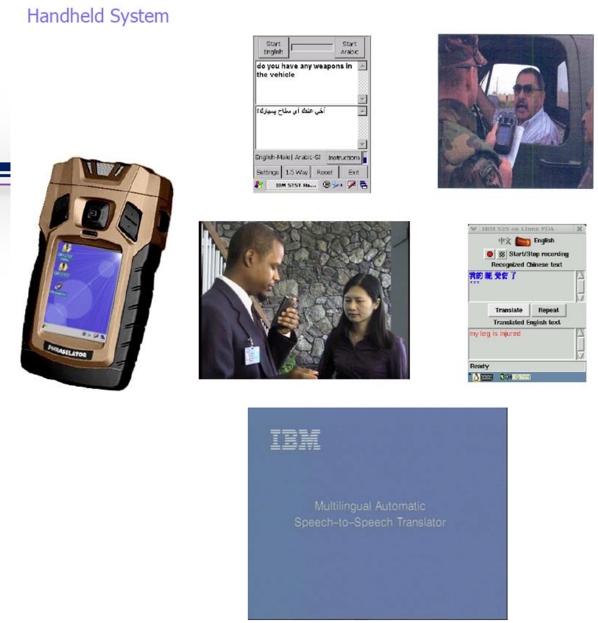


We are witnessing the golden age of ASR!

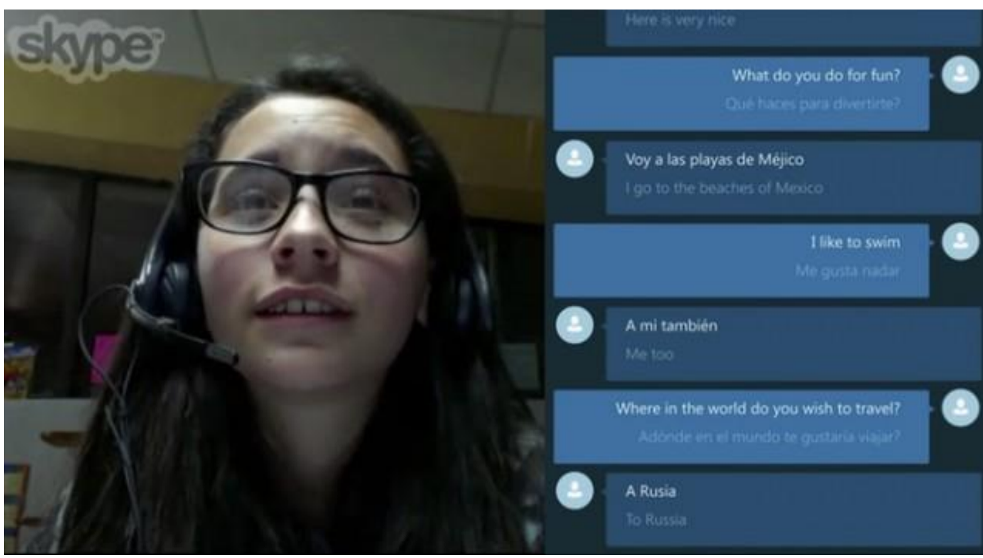
IEEE SLTC eNewsletter - Spring 2010 : *Following Global Events with IBM Translingual Automatic Language Exploration System (TALES)*

Speech-to-Speech Translation





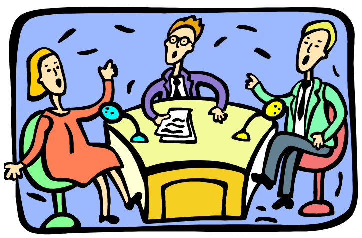
Adapted from the presentation slides of Dr. Yuqing Gao's at ISCSLP2008



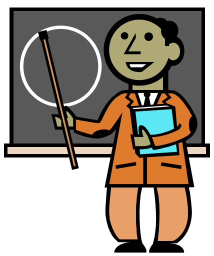
Speech Summarization

conversations

meetings



lectures



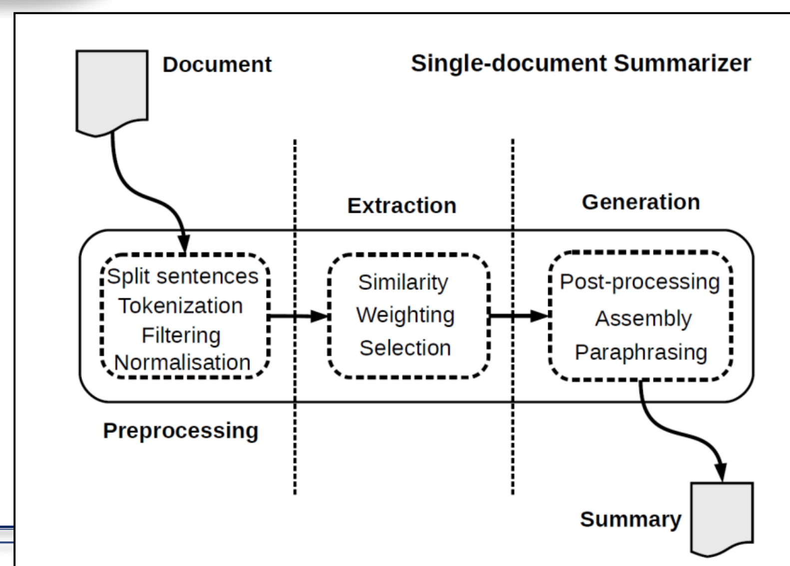
broadcast and TV news

distilling important information

abstractive vs. extractive

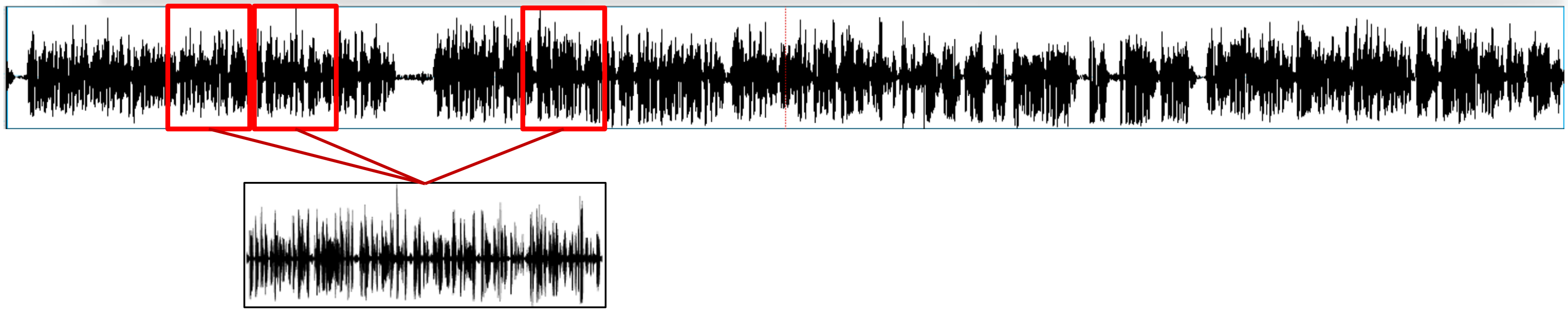
generic vs. query-oriented

single- vs. multi-documents



Torres-Moreno , "Automatic Text Summarization," Wiley-ISTE, 2014.

Speech Summarization: A Running Example



Manual transcript

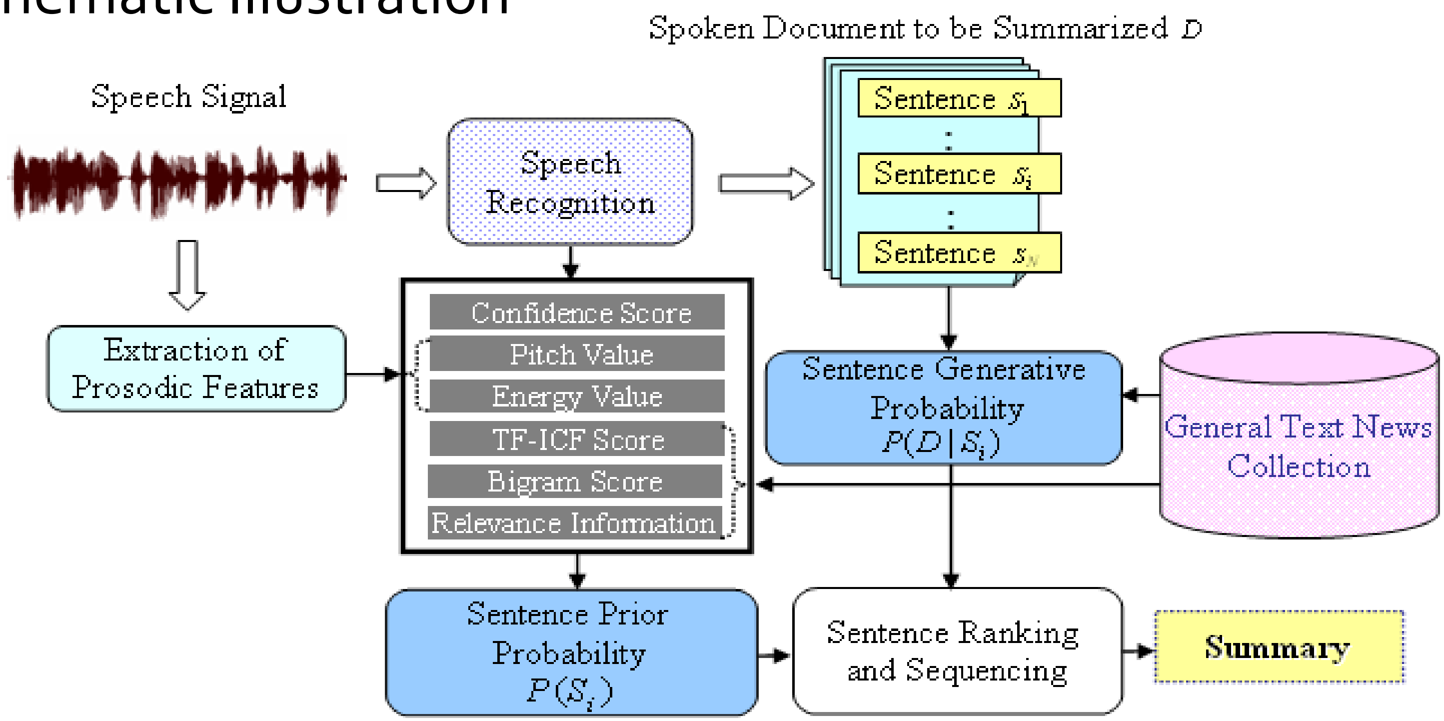
我們 繼續 來 關心 的 是 身體 的 健康
曾經 因為 心臟病 車禍 等 因素 而 接受 過 醫院急救 的 民眾 請 您 特別 留意 下面 這 則 醫療 訊息
因為 病患 在 急救 進行 插 管 治療 時常 常 容易 傷 到 氣管 出現 呼吸 困難 的 後遺症
醫師 提醒 曾經 急救 插 管 的 民眾 要 注意 呼吸 道 的 癒 後 狀況
今年 二十 一 歲 的 蘇 先生 兩 年 前 遭到 電擊 意外 差點 送命
雖然 經過 急救 撿 回 性命 卻 在 一 年 後 出現 了 呼吸 困難 的 後遺症
呼吸 的 時候 都 覺得 快 喘 不過 氣 連 講話 講 個 一 兩 個 字 或者 是 走路 走 走 一下 子 就 覺得 快 喘 不過 氣 來
蘇 先生 後來 才 知道 當初 在 醫院急救 時 醫師 處理 頸部 插 管 不 小心 導致 他 的 氣 管 受傷
氣管 周圍 長出 肉 芽 組織 整個 呼吸 道 因此 阻塞
插 管 的 問題 傷害到 這個 黏膜 以致 於 這 黏膜 長 了 一 圈 這個 肉 芽 組織
你 可以 看到 這邊 這個 洞 只剩下 大概 三 變成 只 靠 這 三 在 呼吸
這 肉 芽 組織 是 不應該 有所 以 本來 應該 有 這麼 大一 個 洞 可以 呼吸 現在 只剩下 這麼 小 一個 洞 可以 呼吸
所以 解決 蘇 先生 呼吸 困難 的 唯一 方法 就是 進行 氣管 環狀 軟骨 的 切除 手術
將 周圍 的 肉 牙 組織 去除 恢復正常 的 呼吸 道
這種 手術 對上 呼吸 道 阻塞 的 病患 有 很 大 的 幫助
不過 醫師 也 提醒 民眾 如果 肉 牙 組織 擴散 到 聲帶 部位 就 不能夠 做 這樣 的 手術 以免 影響 發音
公視 新聞 洪 蕙 竹 郭 俊 麟 採訪 報導

ASR output

風景 在 關心 的 是 身體 的 健康
曾經 因為 心臟病 車禍 的 因素 而 接受 過 醫院 七 九 的 民眾 去年 特別 留意 下 明哲 則 要 去 七
一 位 病患 在 急救 情形 插 管 治療 師 常常 中 英 上午 到 氣管 出現 呼吸 困難 等 後 遺症
醫師 提醒 才 引進 七 九 場 館 的 民眾 要 注意 布希 高 的 北投 狀況
今年 二十 一 歲 的 蘇 先生 兩 年 前 遭到 電擊 意外 差點 送命
雖然 經過 急救 前 回 性命 謝 在 一 年 後 出現 了 呼吸 困難 的 後遺症
賈西亞 所 作 這個 款 傳 不過 七 億元 講話 講 課 另 兩 個 字 或 失蹤 五 宗 座 一下子 就 覺得 會 從中 國 企
福建省 後來 才 知道 當初 在 醫院急救 時 一些 處理 經過 查辦 不 小心 導致 它 的 器 官 受傷
習慣 這 位 長出 中亞 組織 整合 呼吸 道 因此 足賽
曹 文 特 問題 妨礙 到 真面目 被 行政院 模範 的 權責 若要 出資
米 可 抗 痙攣 這個 棟 指出 有的 的 三 名 漁民 特 電子 扣 著 三 名 女 特色 主題
除了 住宿 等 人 罪 本來 應該 在 末代 的 東北 虎 旗 新竹縣 調 增 為 效率
的 動 可以 忽視
隨 解決 簇 先生 呼吸 困難 的 唯一 方法 就是 進行 氣管 換 裝 冷酷 的 切除 手術
將 朝 威 的 中亞 組織 取出 恢復正常 的 體細胞
這種 手術 對上 科技 島 足賽 的 病患 有 很 大 的 幫助
不過 醫師 也 提醒 民眾 若 中亞 組織 擴散 到 省 逮捕 為止 共 構 多 張 的 手術 以免 影響 他 因
公視 新聞 宏 輝 杜 家 駿 明 採訪 報導

A Novel Framework for Speech Summarization

- Schematic Illustration



$$S^* = \arg\min_{S_i \in D} \sum_{S_j \in D} Loss(S_i, S_j) \cdot P(S_j | D)$$

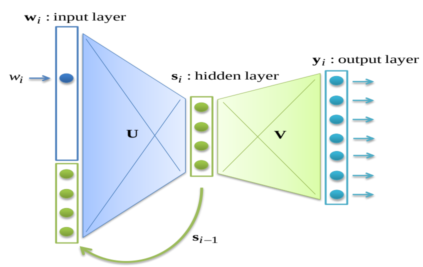
$$= \arg\min_{S_i \in D} \sum_{S_j \in D} Loss(S_i, S_j) \cdot \frac{P(D|S_j)P(S_j)}{\sum_{S_m \in D} P(D|S_m)P(S_m)}$$

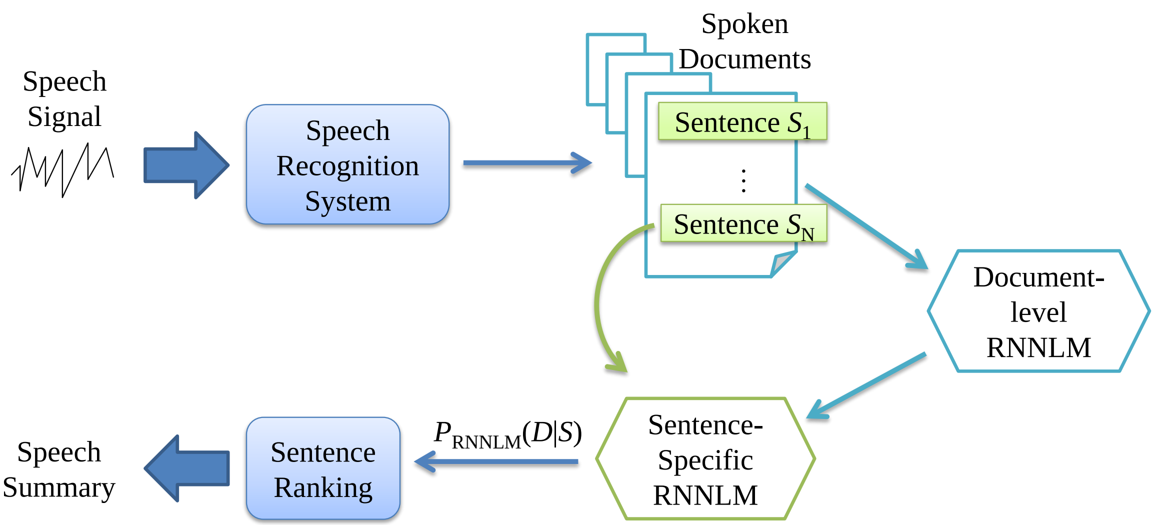
B. Chen and S.-H. Lin, "A risk-aware modeling framework for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 20(1), 2012

Speech Summarization with Recurrent Neural Networks (RNNs)

- Recurrent Neural Networks (RNN) for sentence modeling

$$P_{\text{RNNLM}}(D \mid S) = \prod_{i=1}^{L} P_{\text{RNNLM}}(w_i \mid w_1, \ldots, w_{i-1}, S)$$





Input:
H: Number of Hidden Layer Neurons
$\mathbf{D} = \{D_1, \cdots, D_m, \cdots, D_M\}$
$D_m = \{S_1^{D_m}, \cdots, S_j^{D_m}, \cdots, S_{|D_m|}^{D_m}\}$

Model Training & Important Sentence Ranking:

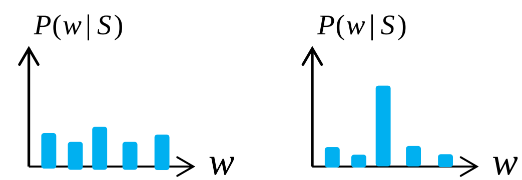
for D_1 to D_M do
  document-level RNNLM model training
  $\mathcal{L}(\mathbf{U}_m, \mathbf{V}_m) = \sum_{i=1}^{|D_m|} \log(y_i)$
  for $S_1^{D_m}$ to $S_{|D_m|}^{D_m}$ do
    sentence-level RNNLM model training
    $\mathcal{L}\left(\mathbf{U}_{S_j^{D_m}}, \mathbf{V}_{S_j^{D_m}} | \mathbf{U}_m, \mathbf{V}_m\right) = \sum_{i=1}^{|S_j^{D_m}|} \log(y_i)$
  end for
  for $S_1^{D_m}$ to $S_{|D_m|}^{D_m}$ do
    calculate document likelihood
    $P(D_m | S_j^{D_m}) = \prod_{i=1}^{|S_j^{D_m}|} P(w_i | w_1, \ldots, w_{i-1}, S_j^{D_m})$
    $= \prod_{i=1}^{|S_j^{D_m}|} P\left(w_i \middle| \mathbf{U}_{S_j^{D_m}}, \mathbf{V}_{S_j^{D_m}}, S_j^{D_m}\right)$
  end for
  Sentence selection according to $P(D_m | S_j^{D_m})$
end for

The design of learning curriculum for RNN is of paramount importance here

Chen et al., "Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques," IEEE/ACM Transactions on Audio, Speech, and Language Processing, August 2015

Speech Summarization with Clarity Measure

- A **clarity score** is defined for each sentence
  - The clarity score incorporates both **intrinsic** and **extrinsic** cues from the sentence



$$Clarity(S) \overset{def}{=} CE(B \| S) - H(S)$$

Extrinsic

$$-\sum_{w \in V} P(w|B) \log P(w|S)$$

Intrinsic

$$-\sum_{w \in V} P(w|S) \log P(w|S)$$

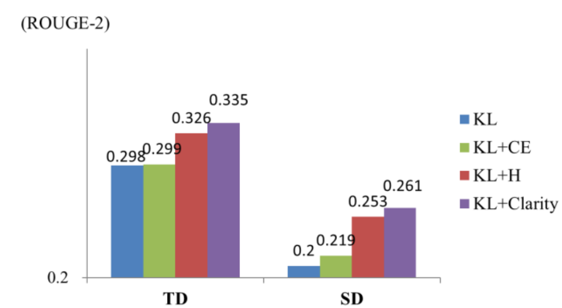
$P(w|B)$: background unigram model

	$CE(B \| S)$	$H(S)$
Low	Close to N_D	Specific
High	Away from N_D	Uniform

  - The clarity score can be combined with KL-Divergence Measure for selecting salient sentences:

$$-KL(D \| S) + Clarity(S)$$
$$= -KL(D \| S) + CE(N_D \| S) - H(S)$$

The higher the score, the more salient the sentence.



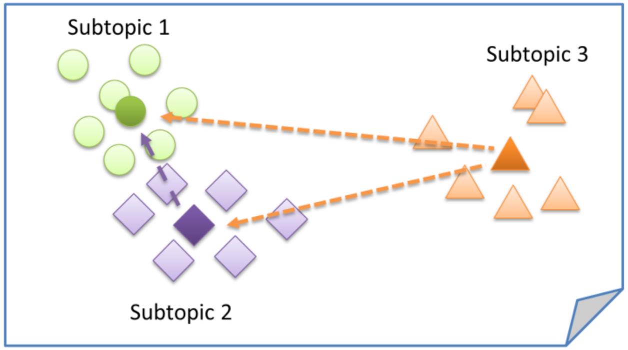
Liu et al., "Combining relevance language modeling and clarity measure for extractive speech summarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, June 2015

Speech Summarization with Density Peaks Clustering

- **Fundamental Premise**: Summary Sentences should Have
  1. A higher density score than other sentences
  2. A higher divergence score than other sentences that also have high density scores
- The density score for any sentence S_i in a document D to be summarized can be defined by

$$density(S_i) = \frac{1}{K-1} \sum_{j=1, j \neq i}^{K} \chi(sim(S_i, S_j) - \delta)$$

$$\chi(x) = \begin{cases} 1 & , if \;\; x > 0 \\ 0 & , otherwise \end{cases}$$

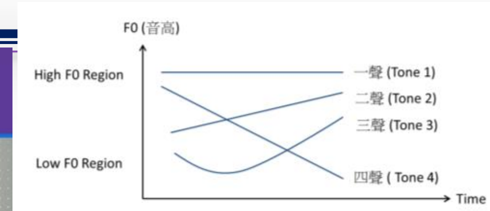


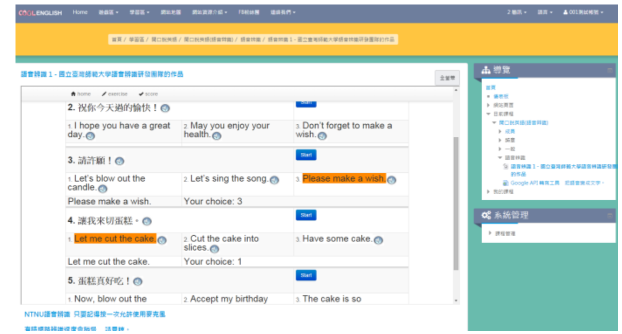
- After the density score for each sentence is obtained, the divergence scores of the sentences are calculated by

$$divergence(S_i) = 1 - \max_{\substack{\forall S_j \in D \\ density(S_j) > density(S_i)}} sim(S_i, S_j)$$

1. A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks, Science, 2014
2. Chen et al., " Incorporating paragraph embeddings and density peaks clustering for spoken document summarization," ASRU 2015

Computer-Assisted Language Training (CAPT)



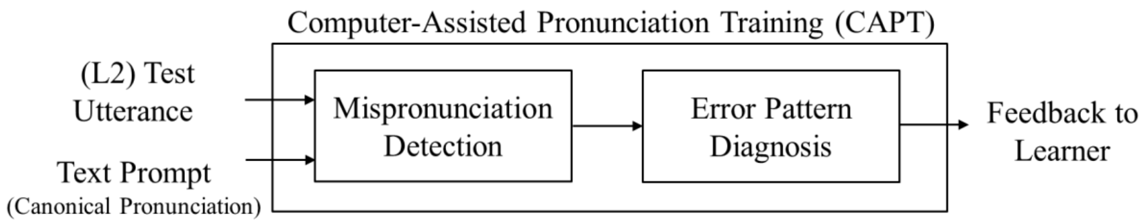


- **Pronunciation of Lexical Tones**: Detection and Assessment
- **Pronunciation of Sub-word (Syllable, INITIAL/FINAL) Units**: Detection and Assessment
- **Speaking Style (Duration, Fluency): Assessment**
- **Overall Scoring** (word-, phrase-, sentence-levels)

1. Mandarin Chinese CAPT: http://140.122.96.191/ALS/assessment.aspx
2. English CAPT: http://www.coolenglish.edu.tw/

CAPT: Motivation

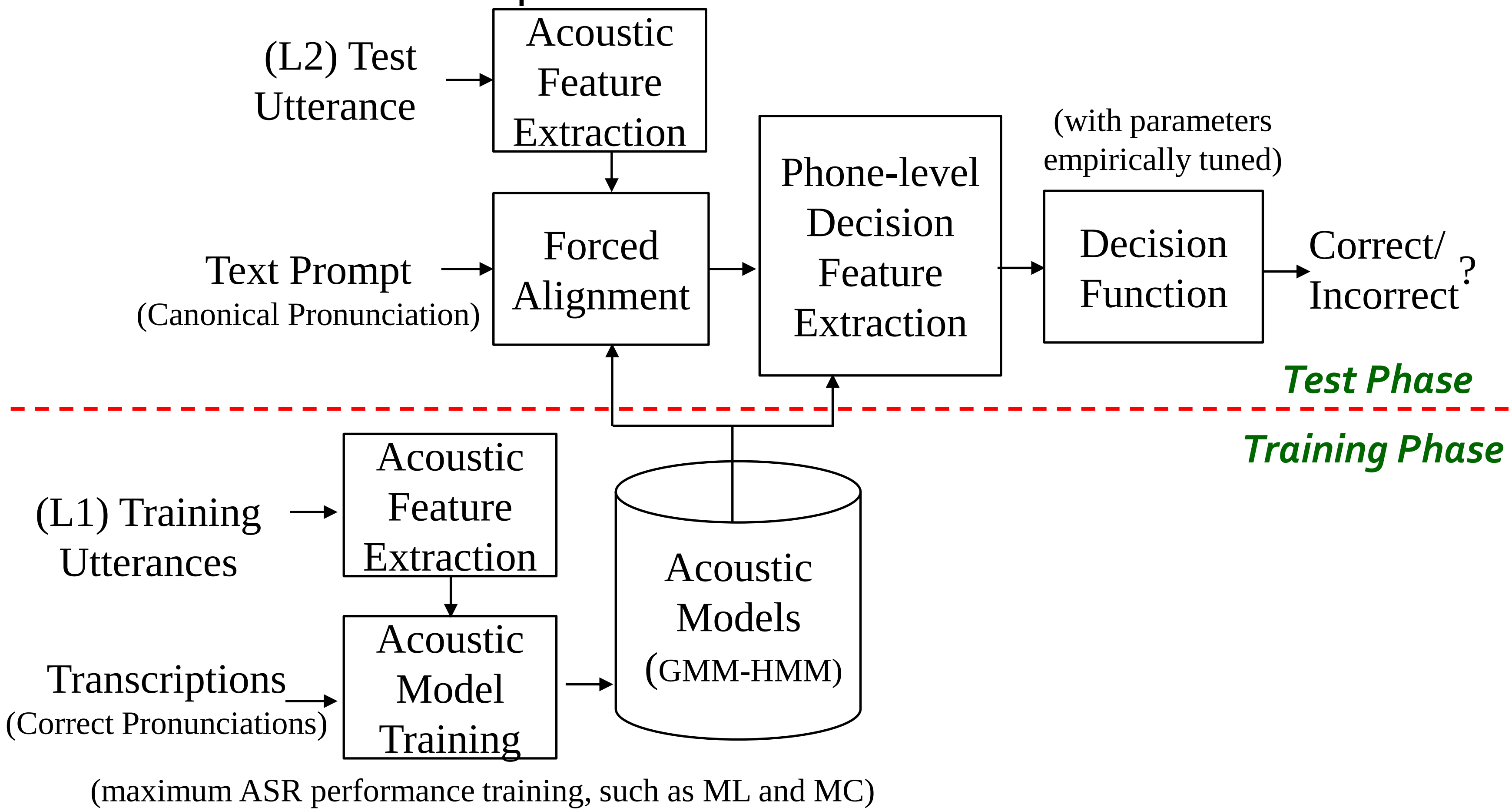
- Computer assisted pronunciation training (CAPT) has attracted increasing research interest recently, partly due to the rapid progress of automatic speech recognition (ASR) technology
  - Deep Learning + Increasing Computational Power + Big Data + …



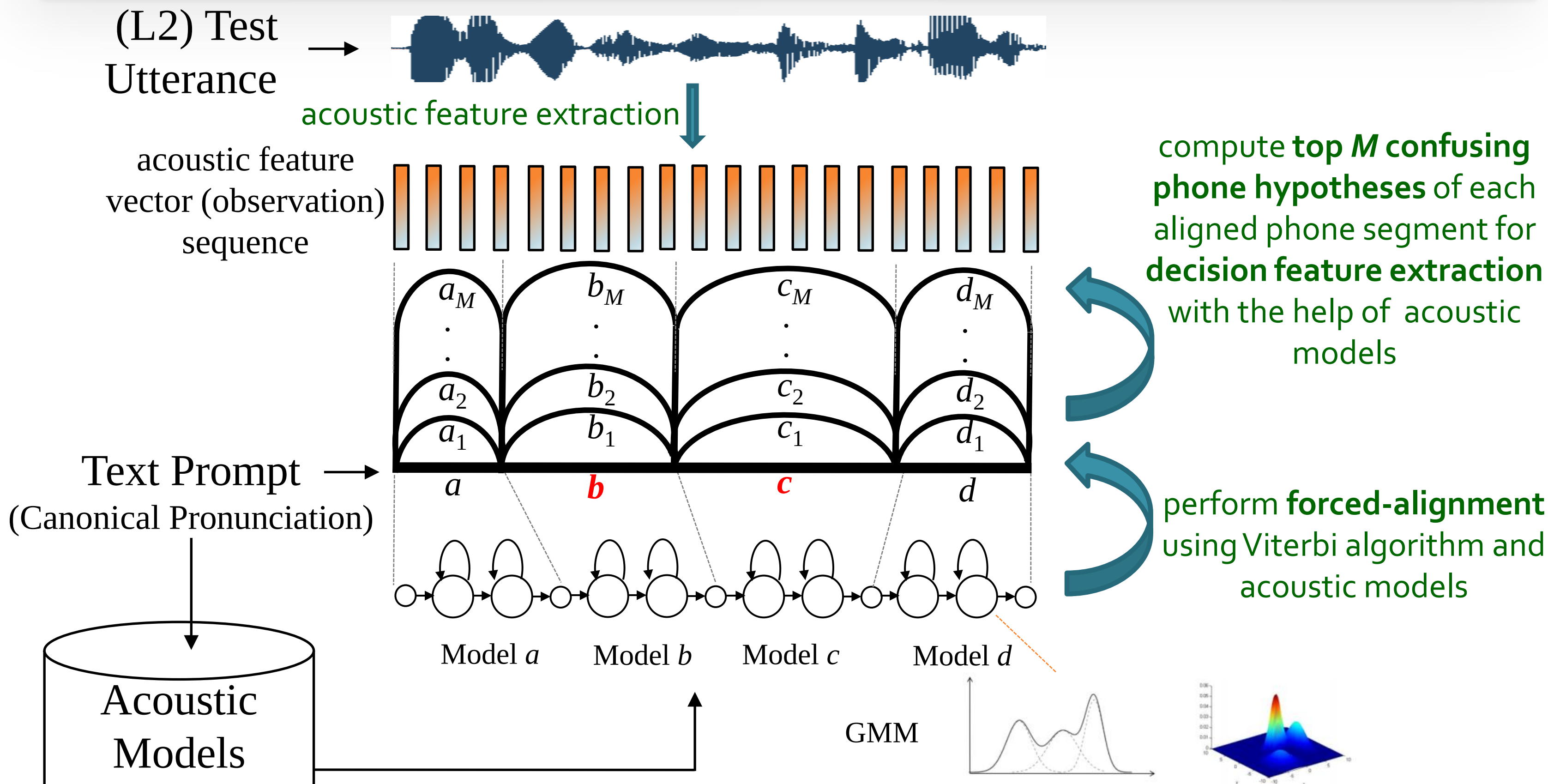
- Mispronunciation detection (MD) is an essential module in a CAPT system
  - Assist second-language (L2) learners to pinpoint incorrect pronunciations in a given utterance in order to improve their spoken proficiency
  - E.g., phone-level or word-level substitution errors, insertion errors, deletion errors, among others

Technical Framework for MD

- Schematic diagram of **a conventional (mainstream) framework** for mispronunciation detection



Forced Alignment & Generating Competing Phone Hypotheses (in the Test Phase)



$$b_j(\mathbf{o}) = \sum_{k=1}^{M} w_{jk} \left(\frac{1}{(2\pi)^{d/2} \left|\mathbf{\Sigma}_{jk}\right|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{o}-\boldsymbol{\mu}_{jk})' \mathbf{\Sigma}_{jk}^{-1} (\mathbf{o}-\boldsymbol{\mu}_{jk}) \right) \right)$$

Mixture Weight; Covariance Matrix; Mean Vector; Observation Vector

GMM-HMM: hidden Markov model (HMM) with Gaussian mixture models (GMM) for estimating state-level observation probability

Phone-level Decision Feature Extraction

- Adopt the commonly-used **goodness of pronunciation** (**GOP**) measure for decision feature extraction, based on the phone-level posterior probabilities computed with forced alignment and acoustic models

$$\mathrm{GOP}(u,n) = \frac{1}{T_{u,n}} \log P(q_{u,n} \mid \mathbf{O}_{u,n})$$

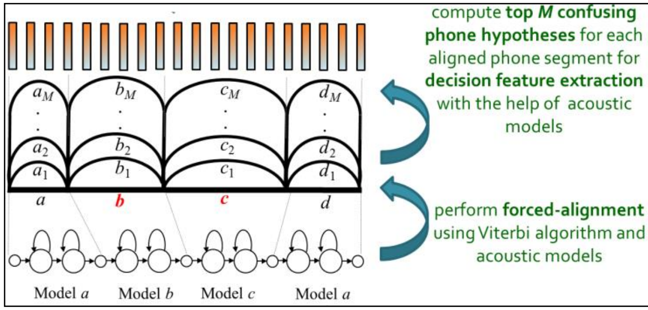
posterior probability

$$\approx \frac{1}{T_{u,n}} \log \frac{P(\mathbf{O}_{u,n} \mid q_{u,n})}{\sum_{\tilde{q} \in \{\mathrm{Top}\ M\}} P(\mathbf{O}_{u,n} \mid \tilde{q})}$$

log likelihood ratio

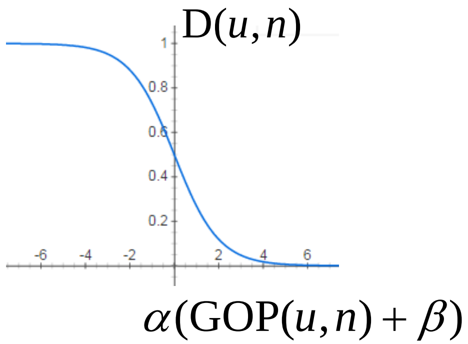
or

$$\mathrm{GOP}(u,n) \approx \frac{1}{T_{u,n}} \log \frac{P(\mathbf{O}_{u,n} \mid q_{u,n})}{\max_{\tilde{q} \in \{Top\ \mathrm{M}\}} P(\mathbf{O}_{u,n} \mid \tilde{q})}$$



Phone-level Decision Functions

- As to the decision function, we can adopt the **logistic sigmoid function** for our purpose



$$D(u,n) = \frac{1}{1+\exp\left[\alpha(\text{GOP}(u,n)+\beta)\right]}$$

- Take the GOP score as the input and output a decision score, ranging between 0 and 1
- $D(u,n) \geq \tau$ implies the occurrence of mispronunciation for phone $q_{u,n}$
  - The higher the decision score, $D(u,n)$, the more likely the phone $q_{u,n}$ is mispronounced
- The parameters α, β and the threshold τ are empirically tuned in practice (one size fits all: all phones share the same set of parameters/threshold)

Our Research Contributions for MD (1/2)

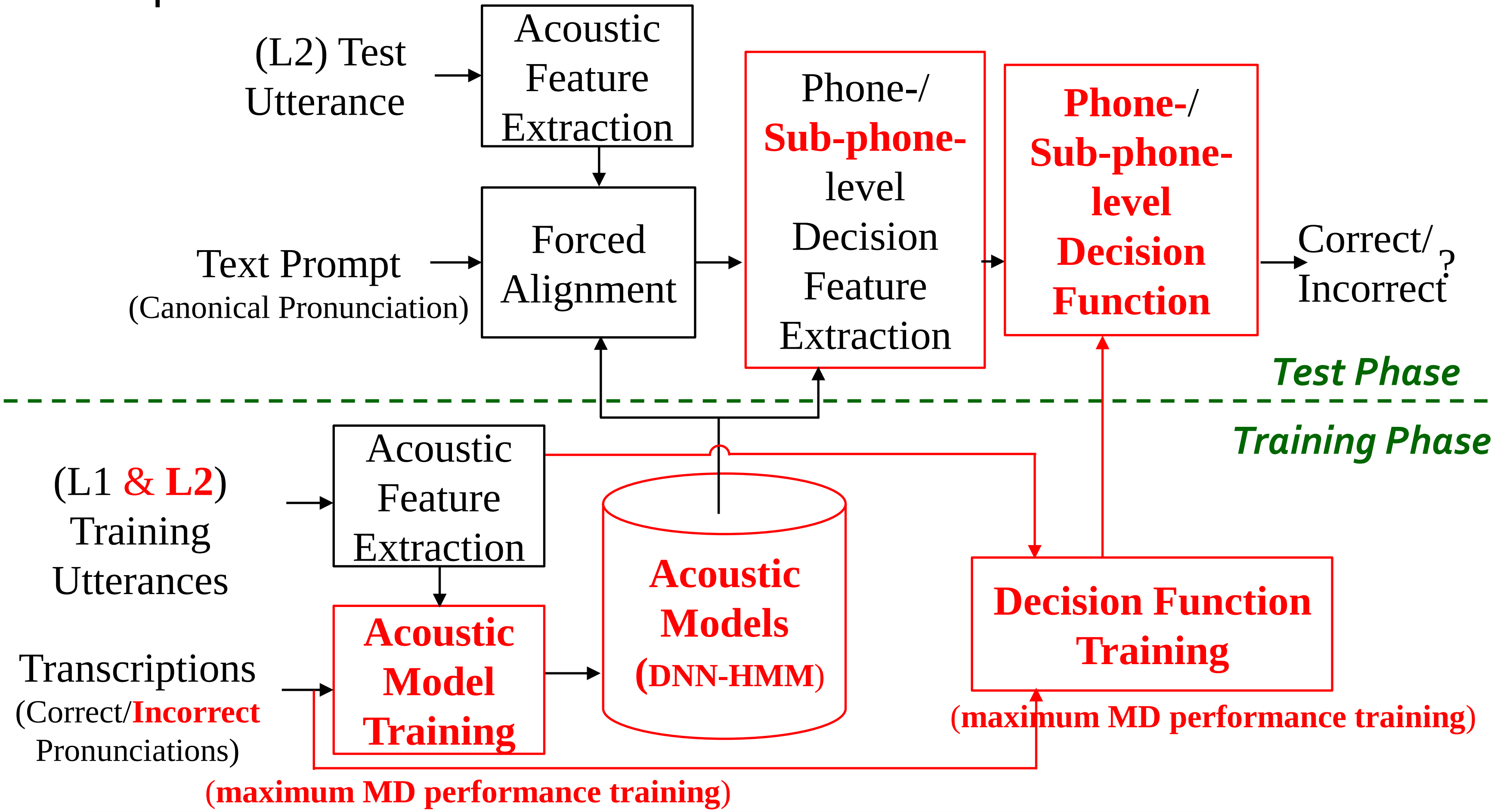
1. We explore recent advances in **deep learning** (especially **deep neural networks**, **DNN**) to achieve better speech feature extraction and acoustic modeling

2. An effective learning approach is proposed, which estimates the DNN-based acoustic models by optimizing an objective directly linked to the ultimate evaluation metric of mispronunciation detection

3. Decision functions of different levels of granularity, with either phone- or sub-phone(senone)-dependent parameterization, are also explored for mispronunciation detection

Our Research Contributions for MD (2/2)

- Schematic diagram of **our proposed approach** to mispronunciation detection



Maximum Performance Training for MD

- Instead of training the acoustic models with criteria that maximize the ASR performance, we attempt to train the acoustic models with an objective function that directly maximizes the performance of MD
  - For example, the **maximum F1-score criterion** (**MFC**)

$$\Xi(\boldsymbol{\theta}) = \frac{2C_{\mathrm{D}\cap\mathrm{H}}}{C_{\mathrm{D}} + C_{\mathrm{H}}} = \frac{2 \cdot \sum_{u=1}^{U} \sum_{n=1}^{N_u} \mathrm{I}(\mathrm{D}(u,n)) \cdot \mathrm{H}(u,n)}{[\sum_{u=1}^{U} \sum_{n=1}^{N_u} \mathrm{I}(\mathrm{D}(u,n))] + C_{\mathrm{H}}}$$

$$\approx \frac{2 \cdot \sum_{u=1}^{U} \sum_{n=1}^{N_u} \mathrm{D}(u,n) \cdot \mathrm{H}(u,n)}{[\sum_{u=1}^{U} \sum_{n=1}^{N_u} \mathrm{D}(u,n))] + C_{\mathrm{H}}}$$

    - Where $\boldsymbol{\theta}$ denotes the set of **parameters** of both **the DNN-HMM based acoustic models** and **the decision function**
    - $C_{\mathrm{D}\cap\mathrm{H}}$ is the total number of phone segments in the training set that are identified as being mispronounced simultaneously by both the current mispronunciation detection module and the majority vote of human assessors
    - Optimized by stochastic gradient ascent algorithm + chain rule for differentiation

Appendix: F1 Score for Performance Evaluation

- The default evaluation metric for **mispronunciation detection** employed in this work is the F1 score, which is a harmonic mean of precision and recall

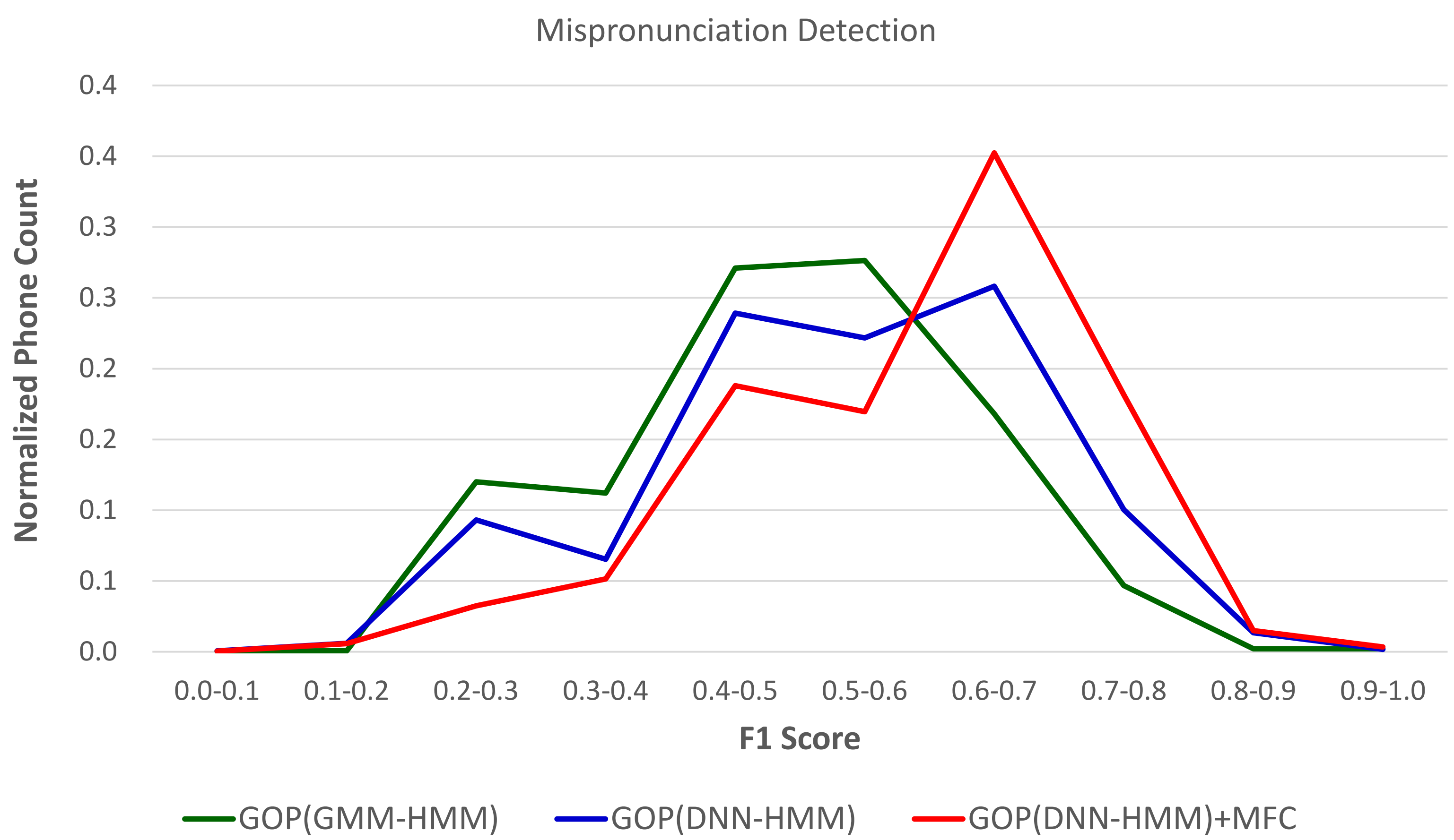
$$\text{F1Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{C_{D \cap H}}{C_D + C_H}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{C_{D \cap H}}{C_D}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{C_{D \cap H}}{C_H}$$

NTNU

Performance Evaluation of MD



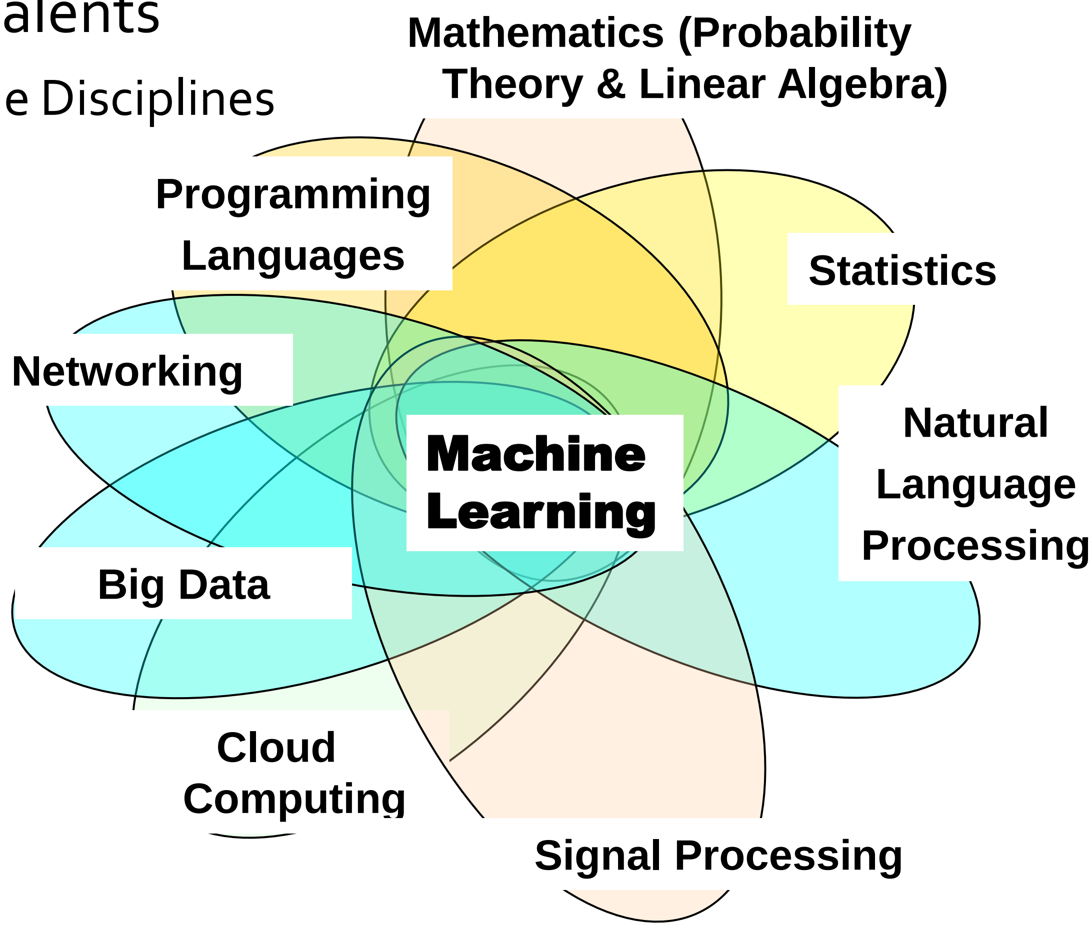
A Running Example of MD

Conclusions (1/2)

- Multimedia information access (over the Web) using speech will be very promising in the near future
- Speech processing technologies are expected to play an essential role in computer-aided (language) learning
- We have observed an increasing surge of interest in developing deep learning techniques for text and multimedia processing
  (as pointed out by Dr. Li Deng at *Interspeech 2015*)
  - Speech recognition: **all** low-hanging fruits are taken
  - Image recognition: **most** low-hanging fruits are taken
  - Natural language processing: **not many** low-hanging fruits are there
  - Big data analytics (recommendations, user behaviors, business strategies) would be a new frontier

Conclusions (2/2)

- Machine Learning (ML) emerges to be an attractive realm of research for young talents
  - Confluence of Multiple Disciplines



Exploring Known Unknowns
vs.
Exploring Unknown Unknowns