Doing Something We Never Could with Spoken Language Technologies

Lin-shan Lee

National Taiwan University

What and Why

- **Some research effort tries to do something Better**
  - having aircrafts fly faster
  - having images look more beautiful
  - always very good

- **Some tries to do something we Never could**
  - developing the Internet to connect everyone over the world
  - selecting information from Internet with Google
  - usually challenging

What and Why

- **Those we could Never do before**
  - very far from realization
  - wait for the right industry to appear at the right time
  - new generations of technologies used very different from the earlier solutions found in research
- **Interesting and Exciting !**
- **Three examples**
  - (1) Teaching machines to listen to Mandarin Chinese
  - (2) Towards a Spoken Version of Google
  - (3) Unsupervised ASR (phoneme recognition)

(1) Teaching Machines to Listen to Mandarin Chinese

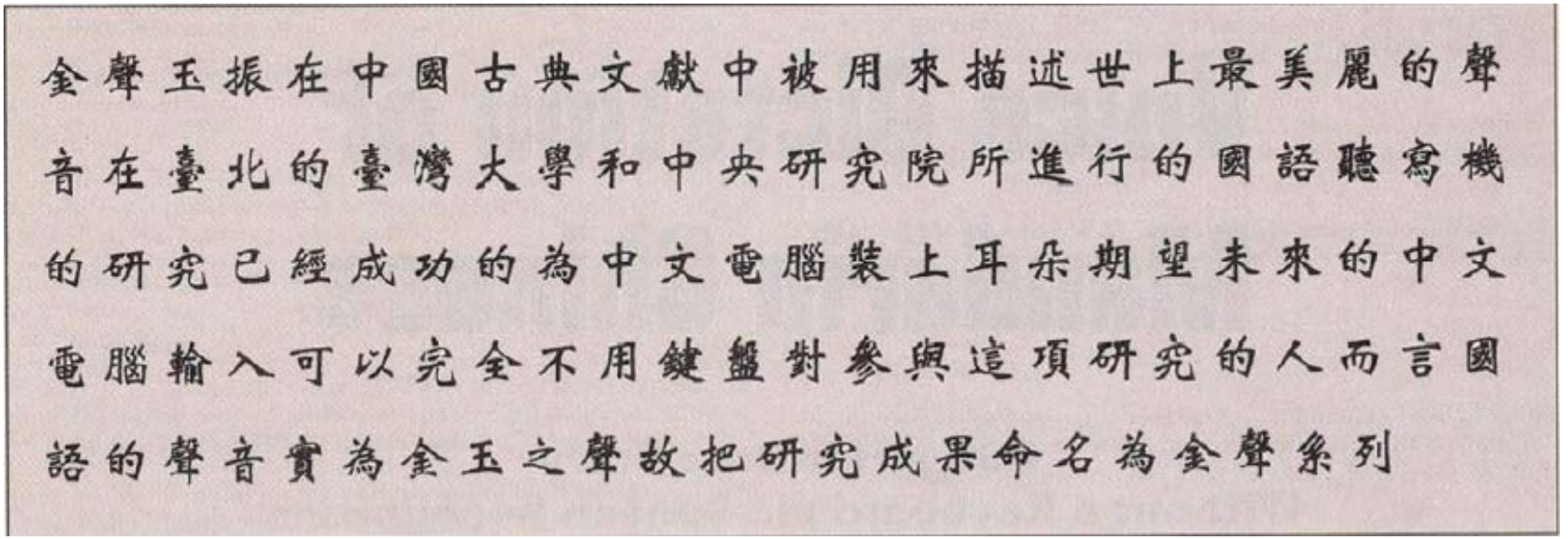
Talk to Machines in Writing – Typewriting

- **English typewriter**

- **Chinese typewriter**

Chinese Language is Not Alphabetic

- **Every Chinese character is a square graph**
  - many thousands of such characters



Talk to Machines in Writing - Typewriting

- **English typewriter**

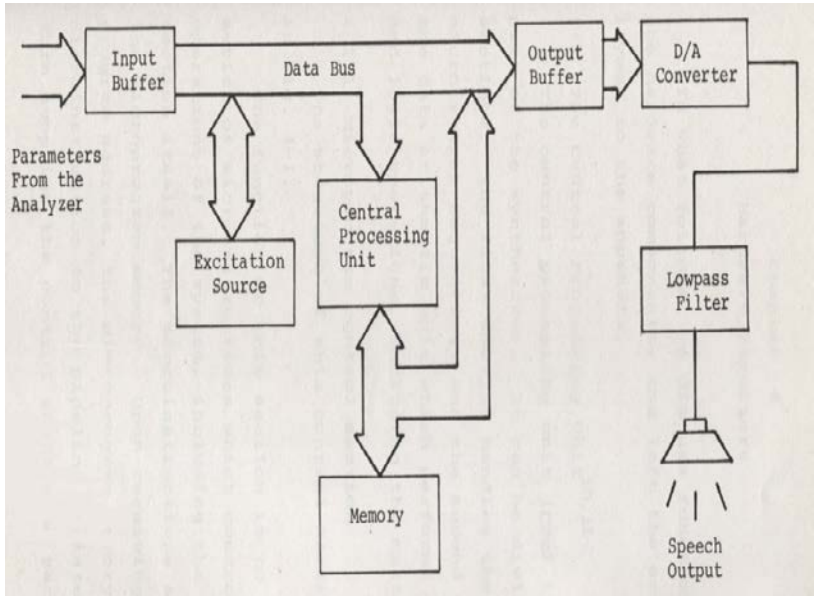
- **Chinese typewriter**

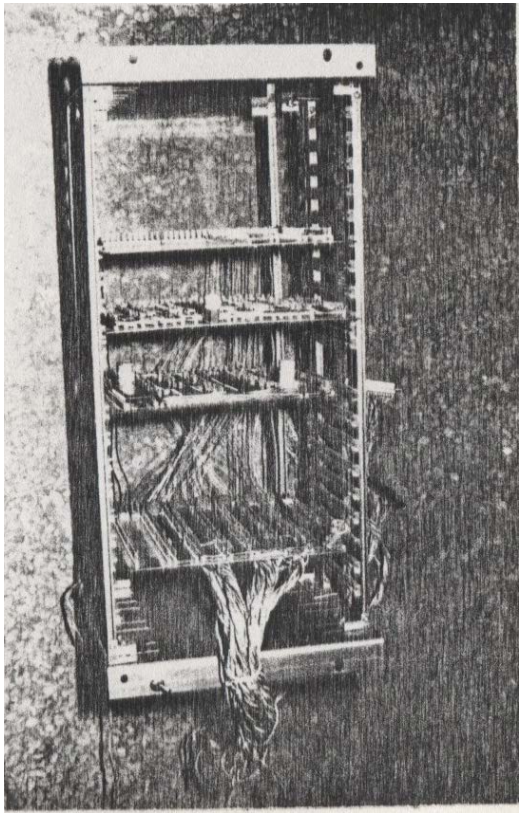
Age of Chinese Typewriters (1980)

- **Many people tried to represent Chinese characters by code sequences**
  - radicals (字根) : 口 火 木 山 人 手
  - phonetic symbols
  - corner codes (四角號碼)
- **A Voice-driven Typewriter for Chinese ?**
  - monosyllable per character
  - total number of distinct monosyllables limited
  - something we could Never do before
- **To Teach Machines to Listen to Mandarin ?**
  - too difficult (1980)
  - to teach machines to speak Mandarin first
    →Speech Synthesis (TTS)

Hardware-assisted Voice Synthesizer

- **Real-time requirement**
  - producing 1 sec of voice data within 1 sec
  - from Linear Predicative Coding (LPC) coefficients
- **Bit-slice microprocessor far too weak**





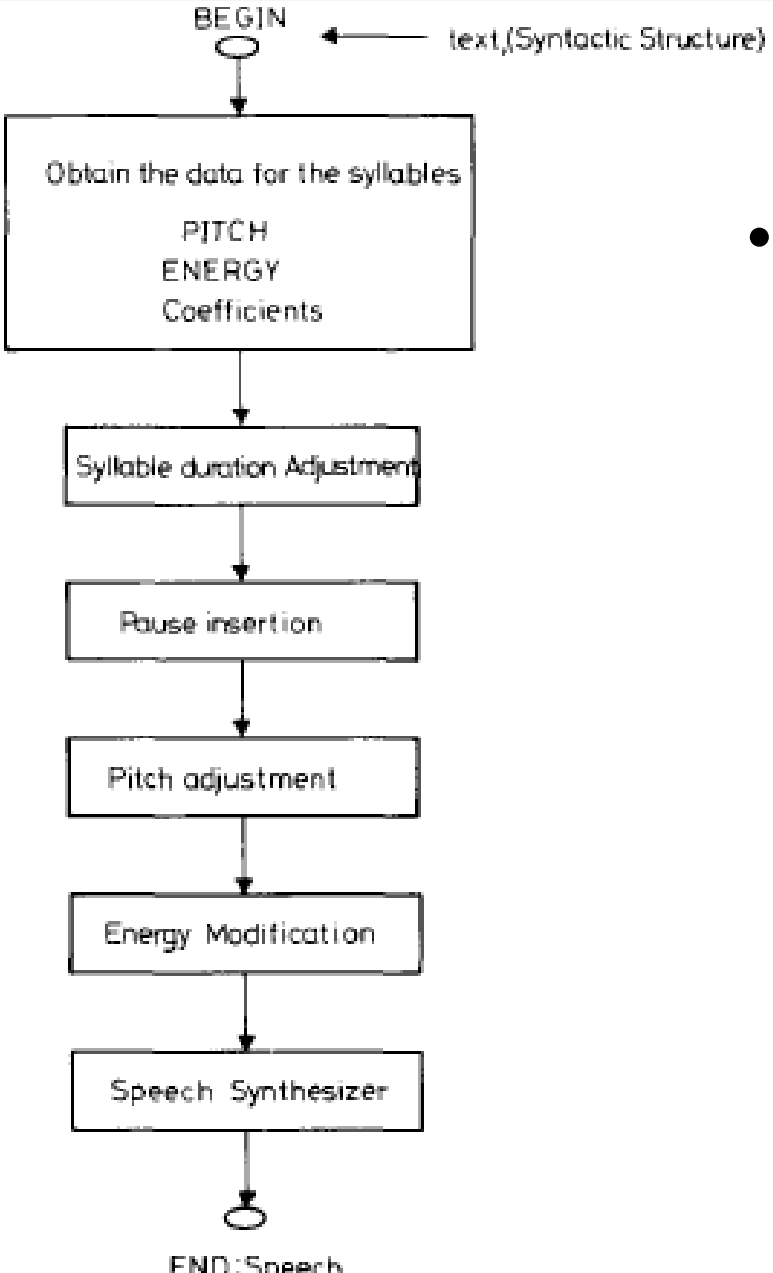
- **Completed 1981, used in next several years**

Initial Effort in Chinese Text-to-speech Synthesis

- **Calculated and stored the LPC coefficients and tone patterns for all Mandarin monosyllables**
  - concatenating isolated monosyllables into utterances directly
  - Never worked, stuck for long
- **Learned from a linguist**
  - the prosody (pitch, energy, duration, etc.) of the same monosyllable is different in different context in continuous speech (context dependency)
  - made sentences, recorded voice, analyzed the prosody
  - prosody rules for each monosyllable in continuous speech based on context
  - synthesized voice very poor, but intelligible
  - data science in early years
- **Lesson learned**
  - linguistics important for engineers

Prof. Chiu-yu Tseng

Mandarin Prosody Rules



- **Complete prosody rules (1984)**

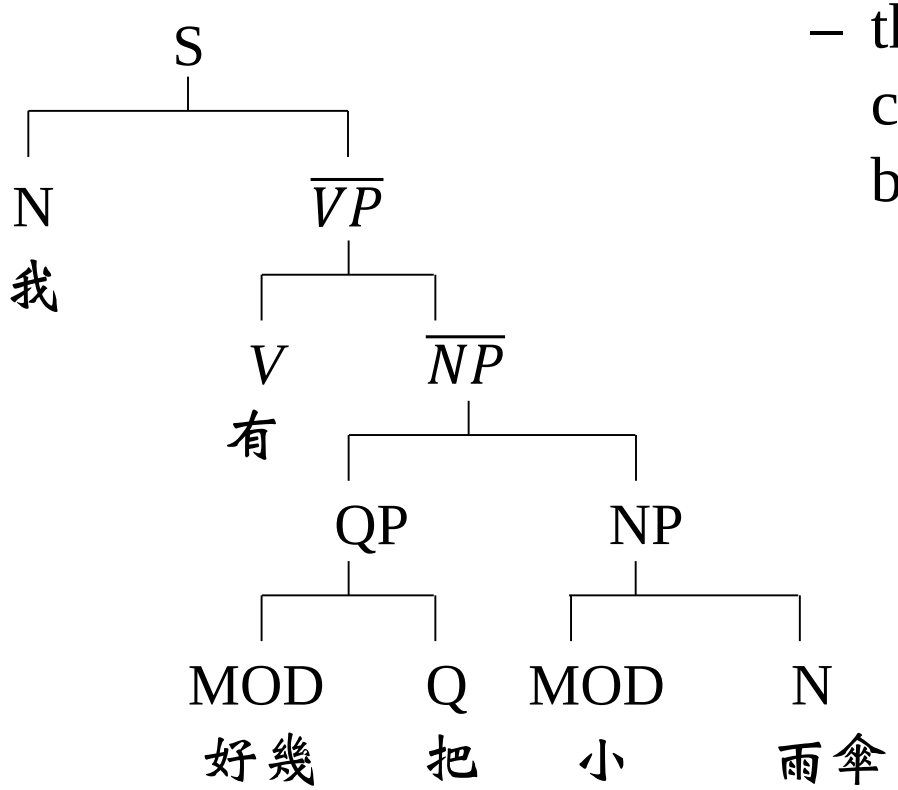
[ICASSP 1986][IEEE Trans ASSP 1989]

Tone Sandhi Rule Example

- **Concatenation rules for Tone 3**

3 3 → 2 3

雨傘



– the rule applies across certain syntactic boundaries, but not for others

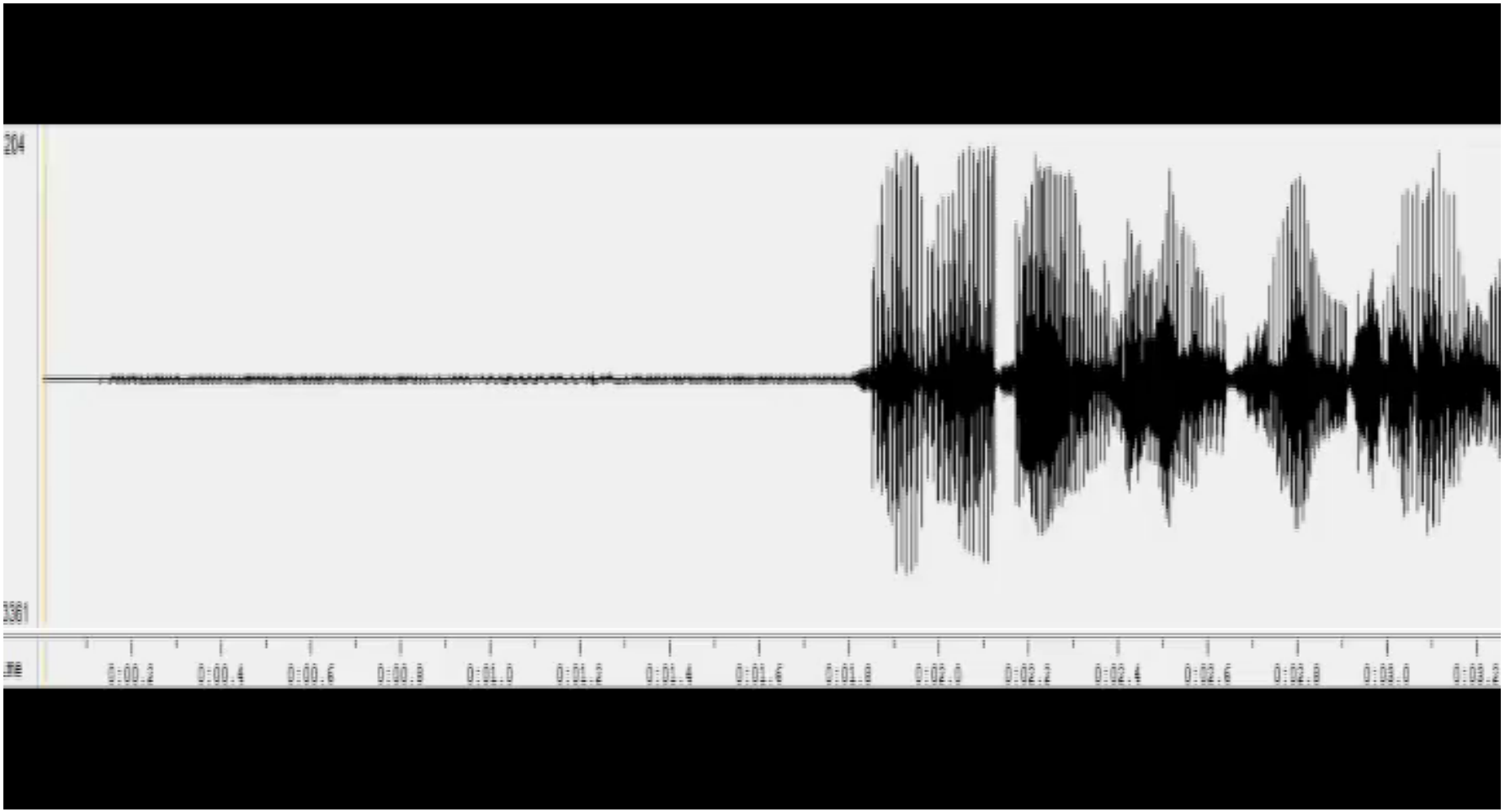
我 有 好 幾 把 小 雨 傘

2 3 2 2 3 3 2 3

[ICASSP 1986][IEEE Trans ASSP 1989]

Chinese Text-to-speech Synthesis

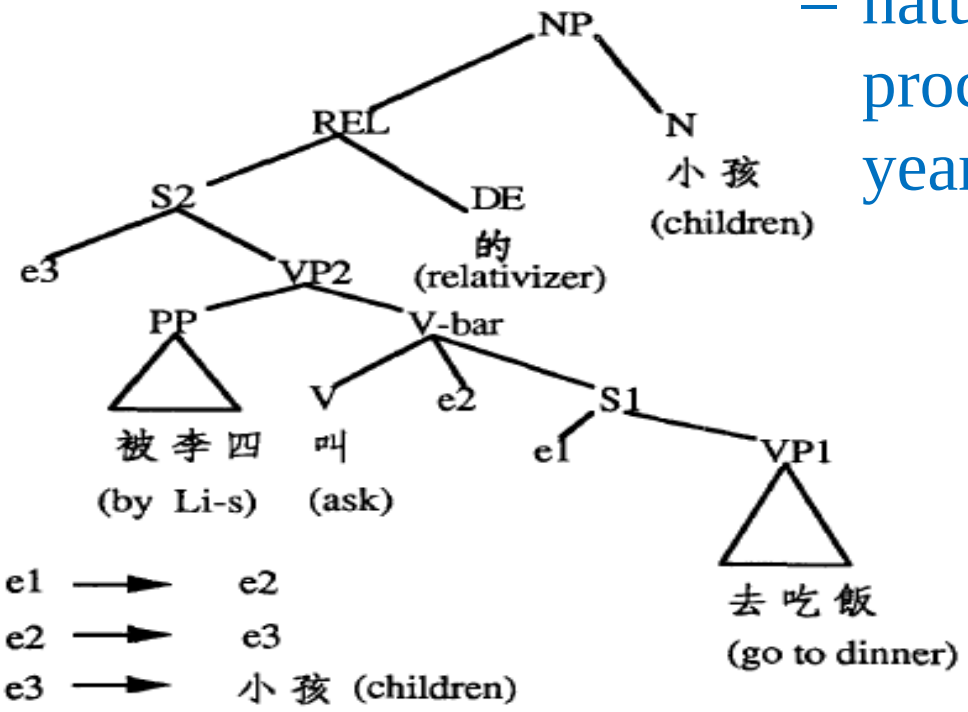
- **Concatenating stored monosyllables with adjusted prosody (1984)**



[ICASSP 1986] [IEEE Trans ASSP 1989]

Speech Prosody Depends on Sentence Structure

- **Chinese Sentence Grammar and Parser (1986)**
- **Empty category**
  - natural language processing in early years



被李四叫去吃飯的小孩

(The children who are asked to go to dinner by Li-s)

[AAAI 1986] [Computational Linguistics 1991]

Mandarin Speech Recognition

- **Very Large Vocabulary and Unlimited Text**
- **Isolated Monosyllable Input**
  - each character pronounced as a monosyllable
  - limited number of distinct monosyllables
  - a voice-driven typewriter
- **Decomposing the problem**
  - recognition of consonants/vowels/tones
  - identifying the character out of many homonym characters
- **System Integration very hard**
  - software very weak, each part assisted by different hardware

[Concept IJCAI 1987] [Initial Results ICASSP 1990]

An Example Circuit Diagram

- **1989**

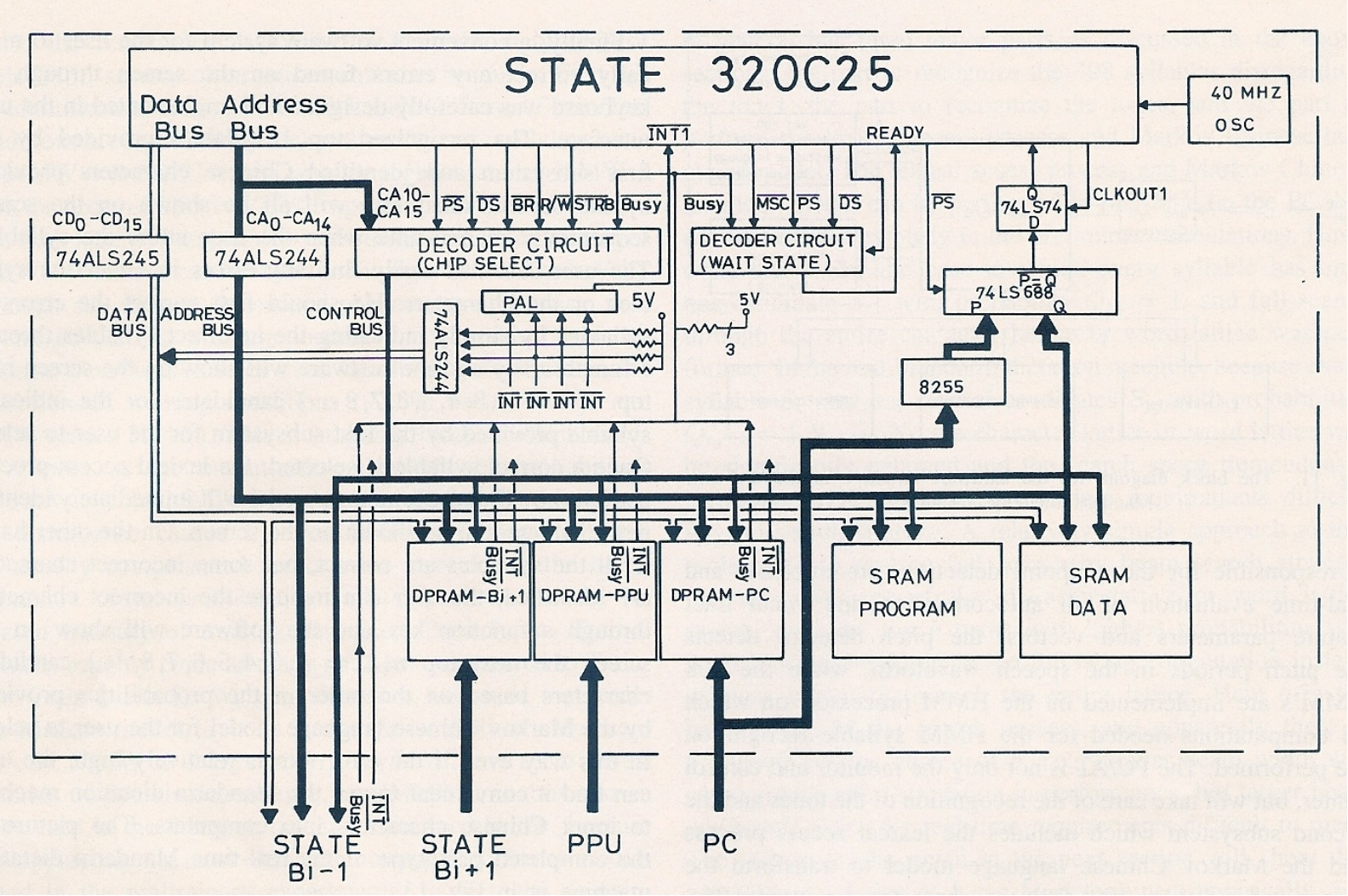
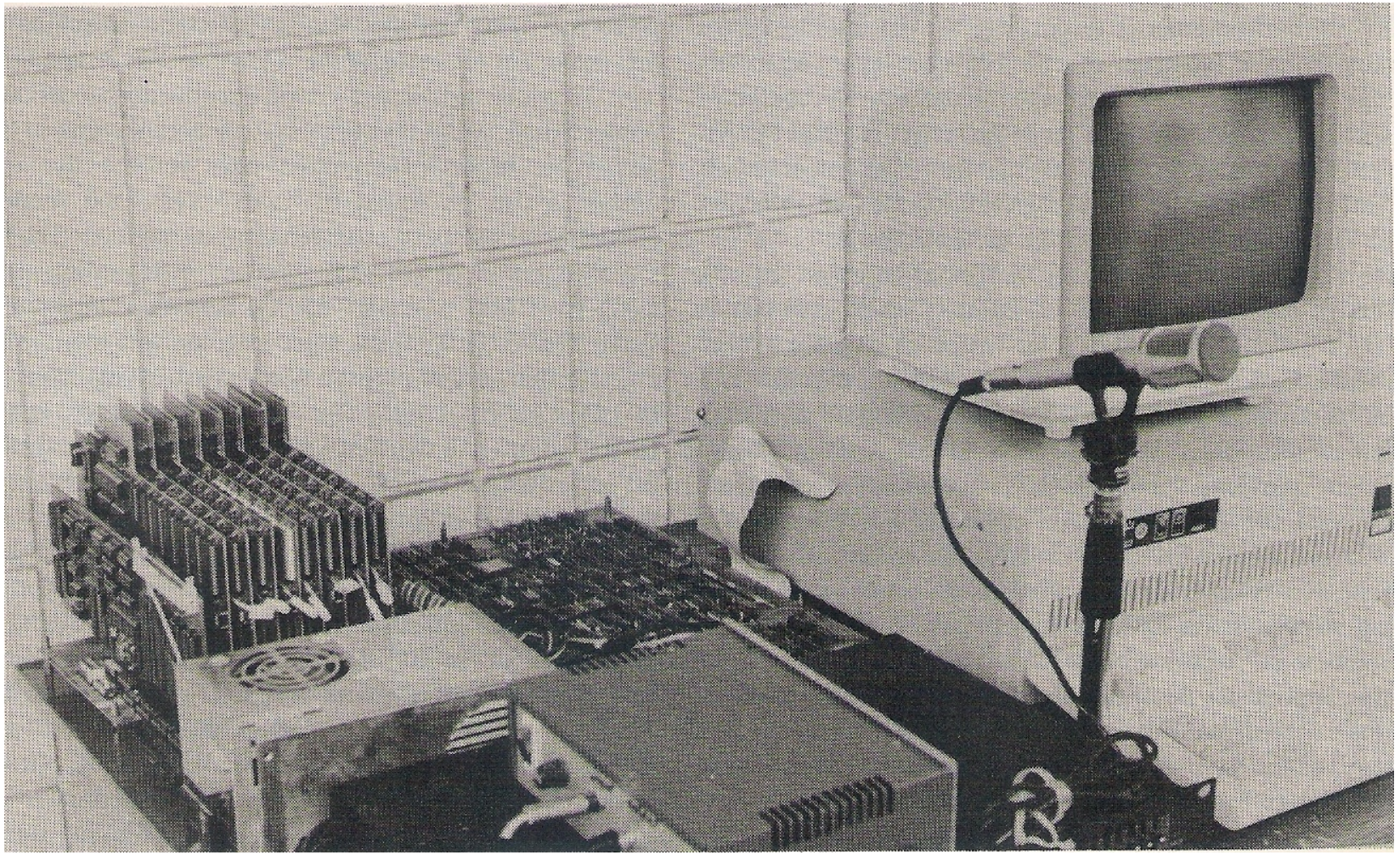


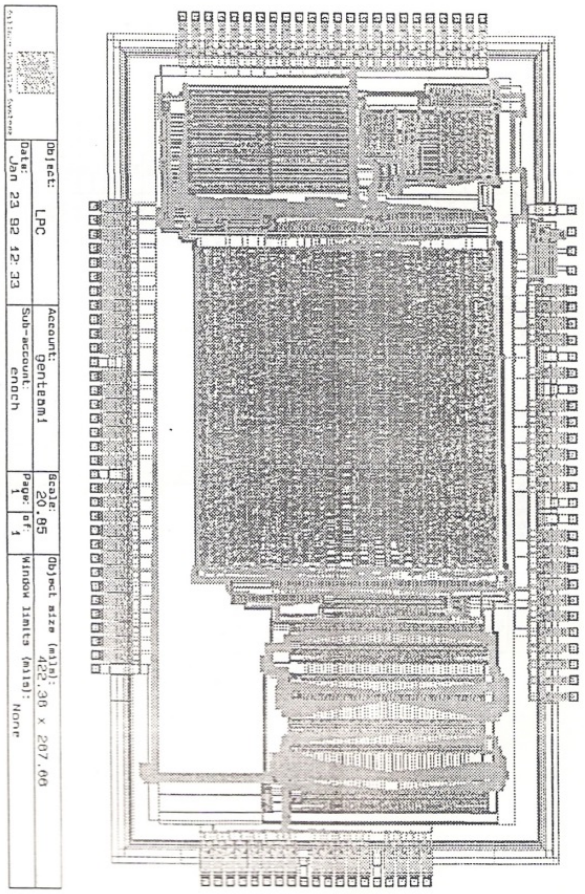
Photo of Completed Hardware

- **Completed in 1989, but Never worked**

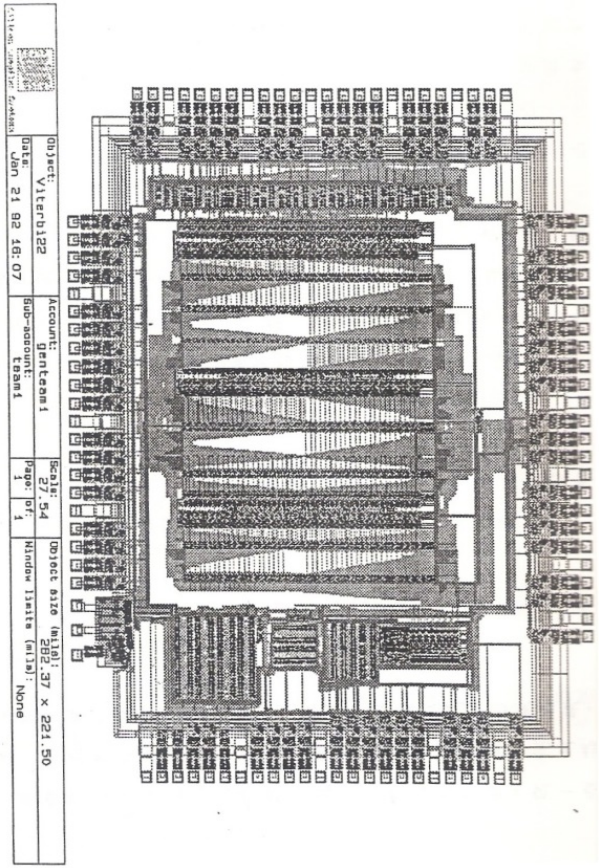


Chip Design

- **1992**
  - far from a complete system



LPC Analysis Chip



Viterbi Chip

Golden Mandarin I (金聲一號)

- **Implemented with Transputer (10 CPUs in parallel, everything in software) purchased in 1990**
- **March 1992**
- **金聲玉振，金玉之聲**
- **Isolated monosyllable input**
- **Several seconds per monosyllable**
- **Lesson learned: software more powerful than hardware**

[ICASSP 1993] [IEEE Trans SAP 1993]

Golden Mandarin III (金聲三號)

- **March 1995**
- **Continuous read speech input**
- **Limitation by computation resources**
  - Version(a) on PC 486 for short utterance input
  - Version(b) on Workstation for long utterance input

[ICASSP 1995] [IEEE Trans SAP 1997] [IEEE SP Magazine 1997]

Mini-Remarks

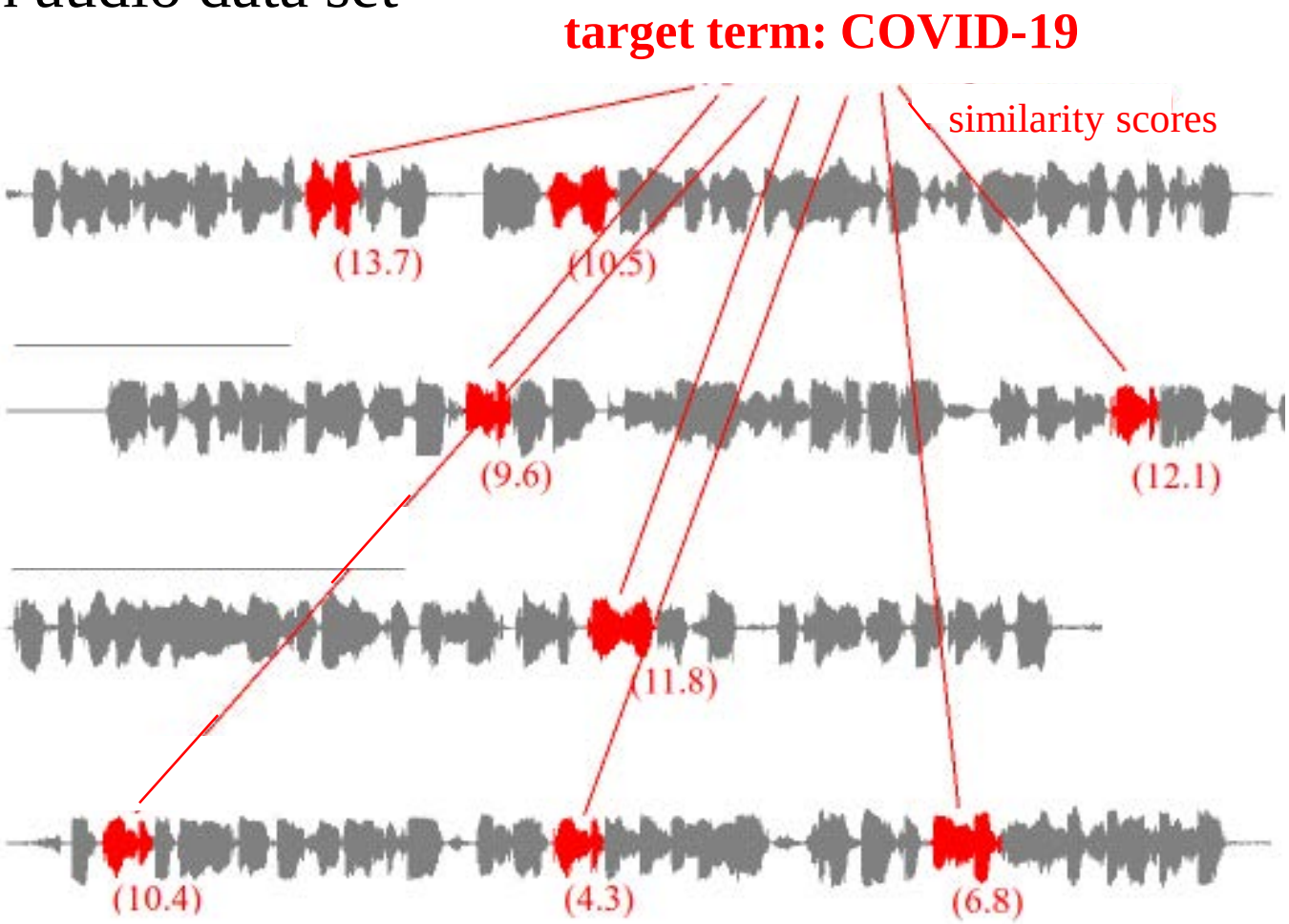
- **Today**

- **A dream many years ago was realized by the right industry at the right time**
  - with new generations of technologies
- **No worry for realization during research**
  - someone will solve the problem in the future

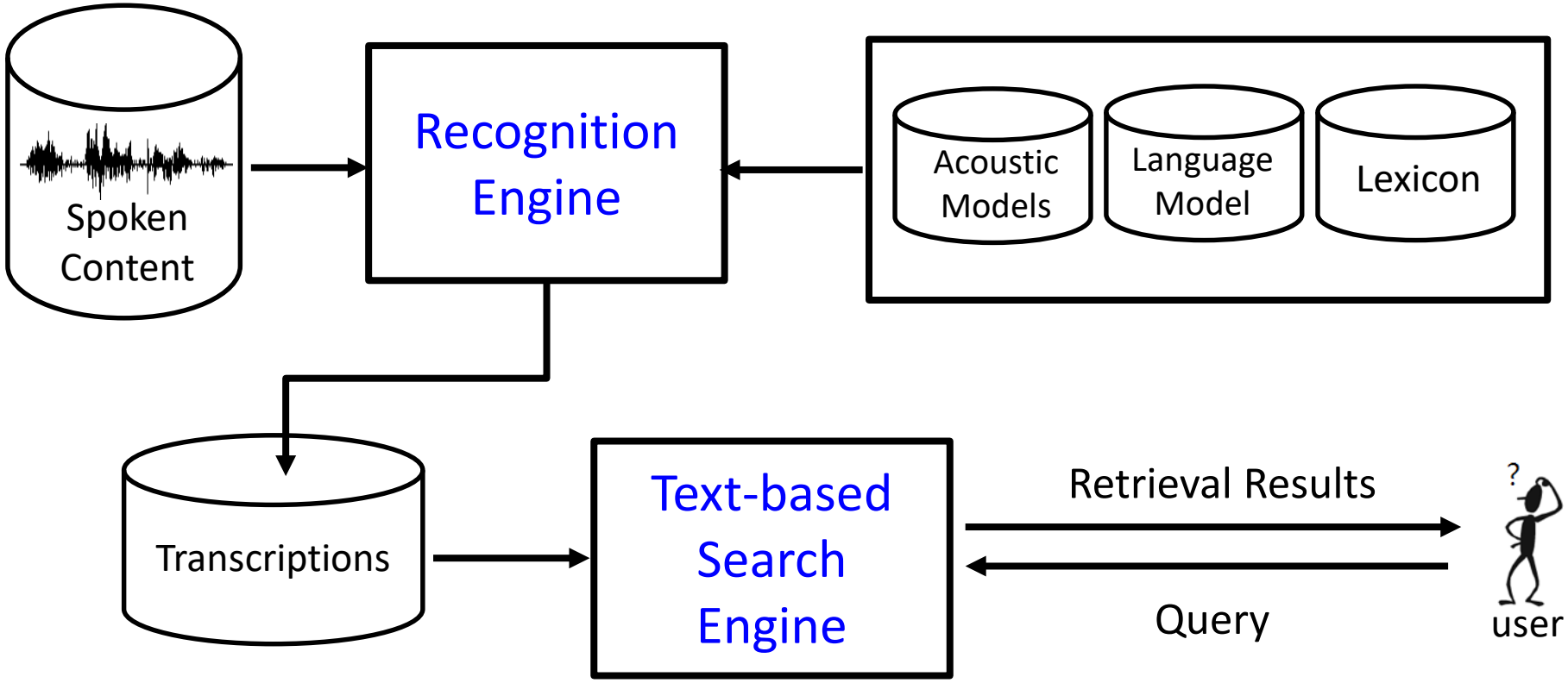
(2) Towards a Spoken Version of Google

Spoken Content Retrieval

- **Spoken term detection**
  - to detect if a target term was spoken in any of the utterances in an audio data set



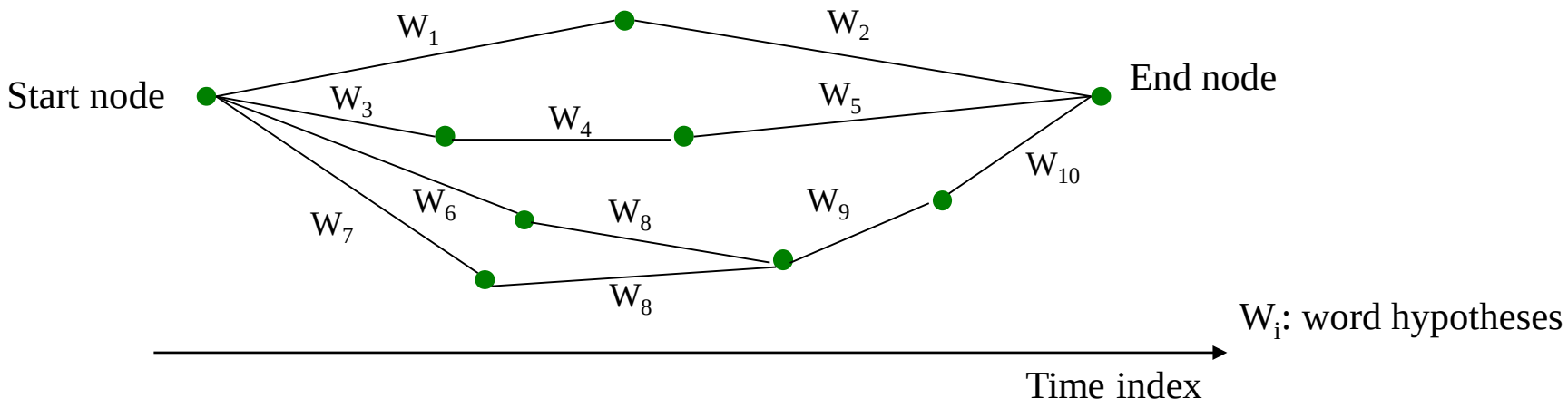
Spoken Content Retrieval – Basic Approach



- **Transcribe the spoken content**
- **Search over the transcriptions as they are text**
- **Recognition errors cause serious performance degradation**

Lattices for Spoken Content Retrieval

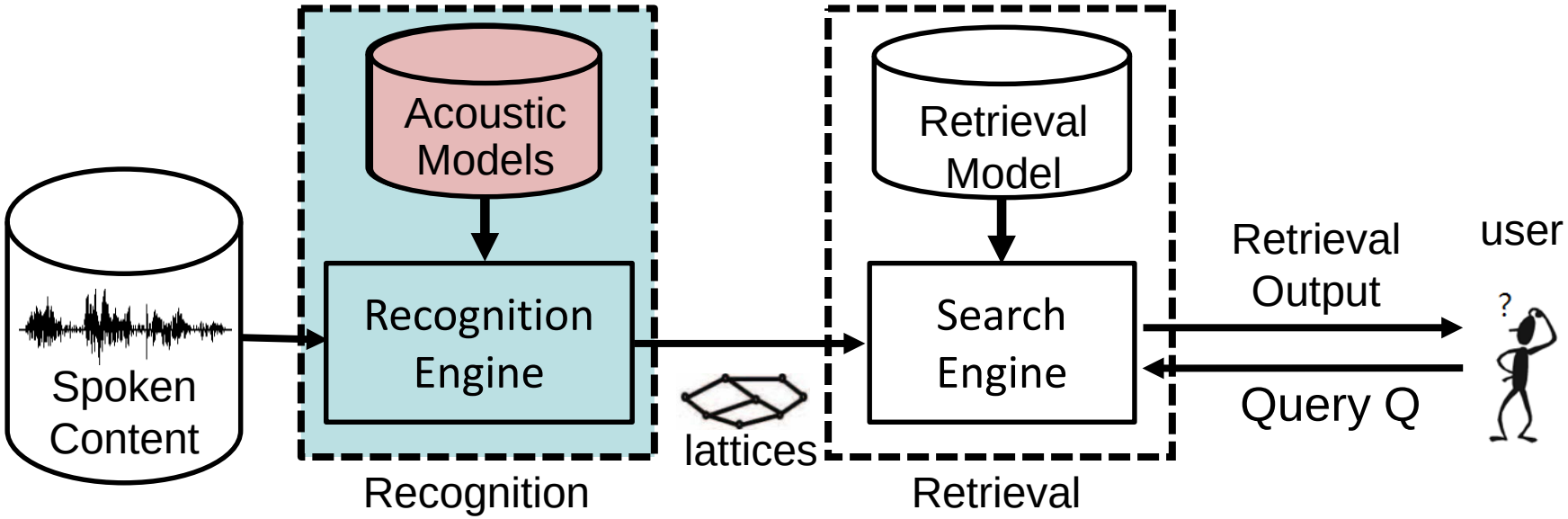
- **Low recognition accuracies for speech signals under various acoustic conditions**
  - considering lattices rather than 1-best output
  - lattices of subword units



[Ng 2000, Scarclar 2004, Chelba 2005, etc.]

ASR Error Problem Comes from Cascading Two Stages

- **Recognition stage cascaded with retrieval stage**
  - retrieval performance limited by recognition accuracy
- **Considering recognition and retrieval processes as a whole (2010)**
  - acoustic models re-estimated by optimizing overall retrieval performance



- **End-to-End Spoken Content Retrieval in early years**

[ICASSP 2010]

What can Spoken Content Retrieval do for us ?

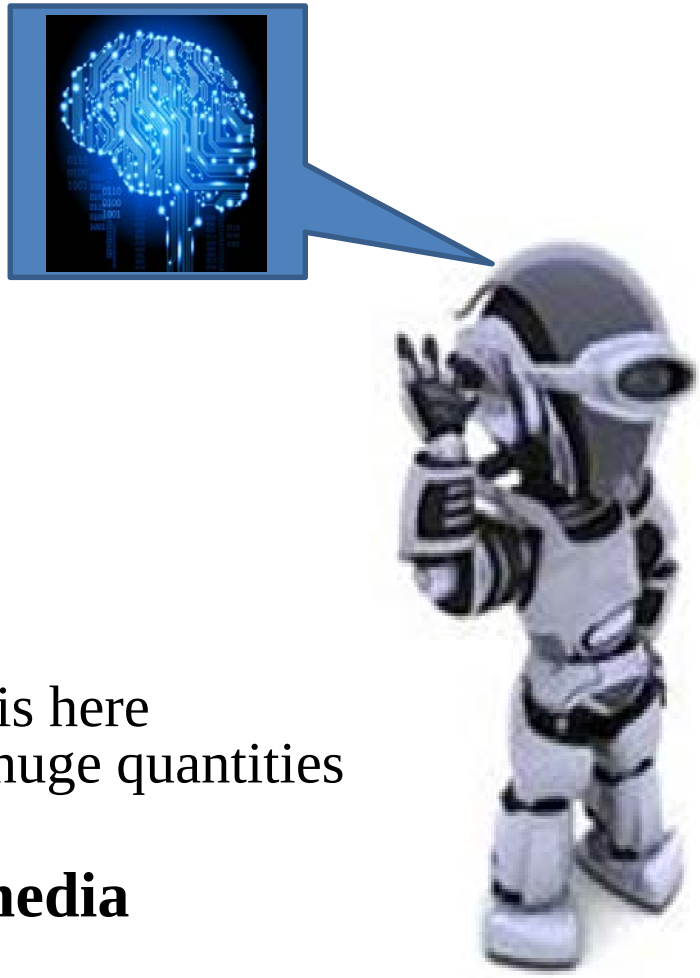
- **Google reads all text over the Internet**
  - can find any text over the Internet for the user
- **All Roles of Text can be realized by Voice**
- **Machines can listen to all voices over the Internet**
  - can find any utterances over the Internet for the user
- **A Spoken Version of Google**

What can we do with a Spoken Version of Google ?

- **Multimedia Content exponentially increasing over the Internet**

300hrs of videos uploaded per min (2015.01)

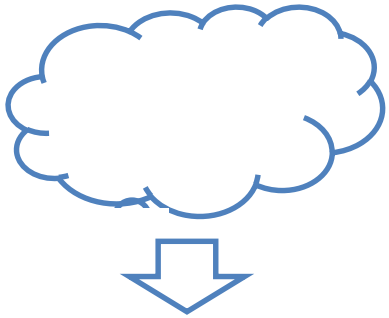
Roughly 2000 online courses on Coursera (2016.04)



  - best archive of global human knowledge is here
  - desired information deeply buried under huge quantities of unrelated information
- **Nobody can go through so much multimedia information, but Machines can**
- **Machines may be able to listen to and understand the entire multimedia knowledge archive over the Internet**
  - extracting desired information for each individual user

A Target Application Example : Personalized Education Environment

- **For each individual user**



Information from Internet

user

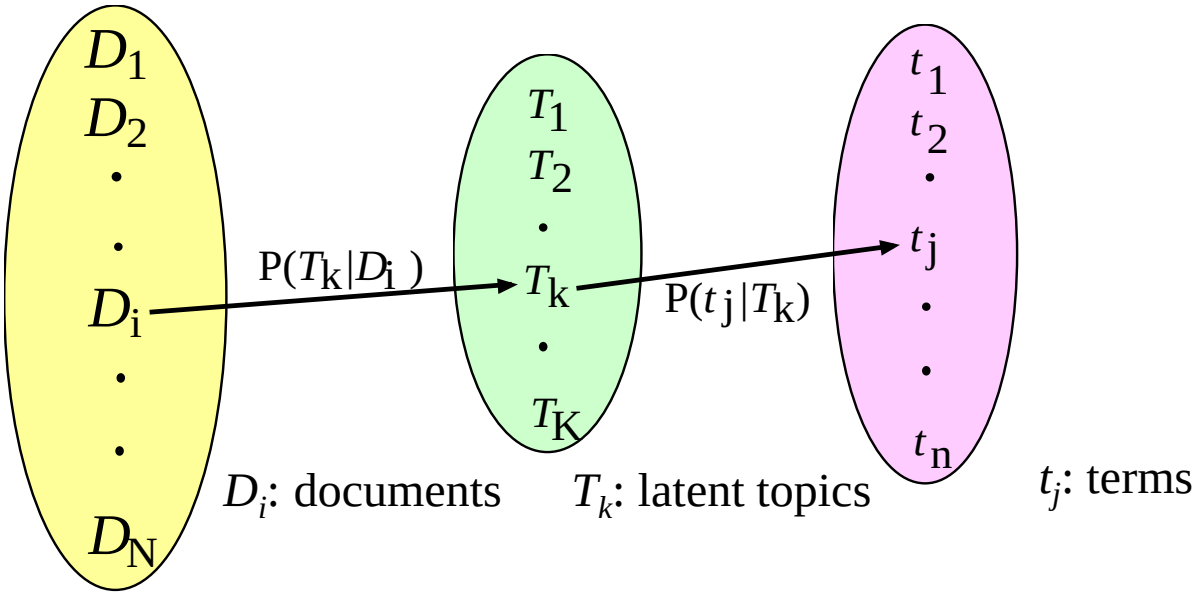
- I wish to learn about Wolfgang Amadeus Mozart and his music
- I can spend 3 hrs to learn

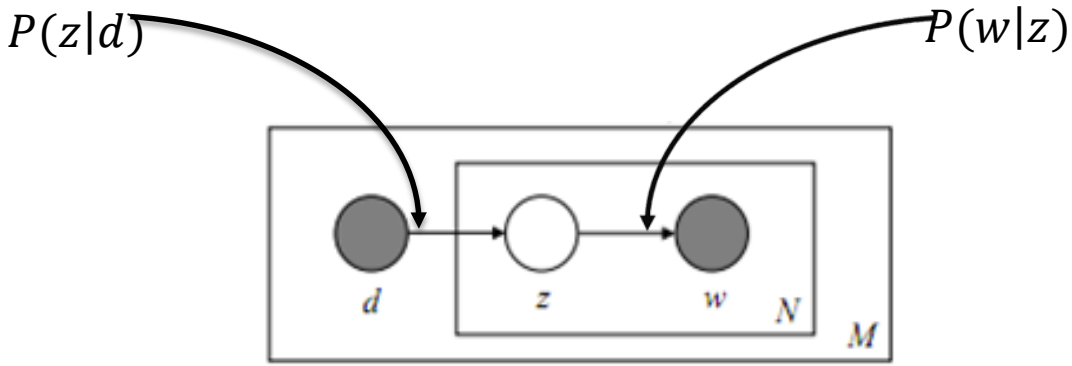
This is the 3-hr personalized course for you. I'll be your personalized teaching assistant. Ask me when you have questions.

- **Understanding, Summarization and Question Answering for Spoken Content**
  - something we could Never do (even today)
  - semantic analysis for spoken content

Probabilistic Latent Semantic Analysis (PLSA)

- **Unsupervised Learning of Topics from text corpus**

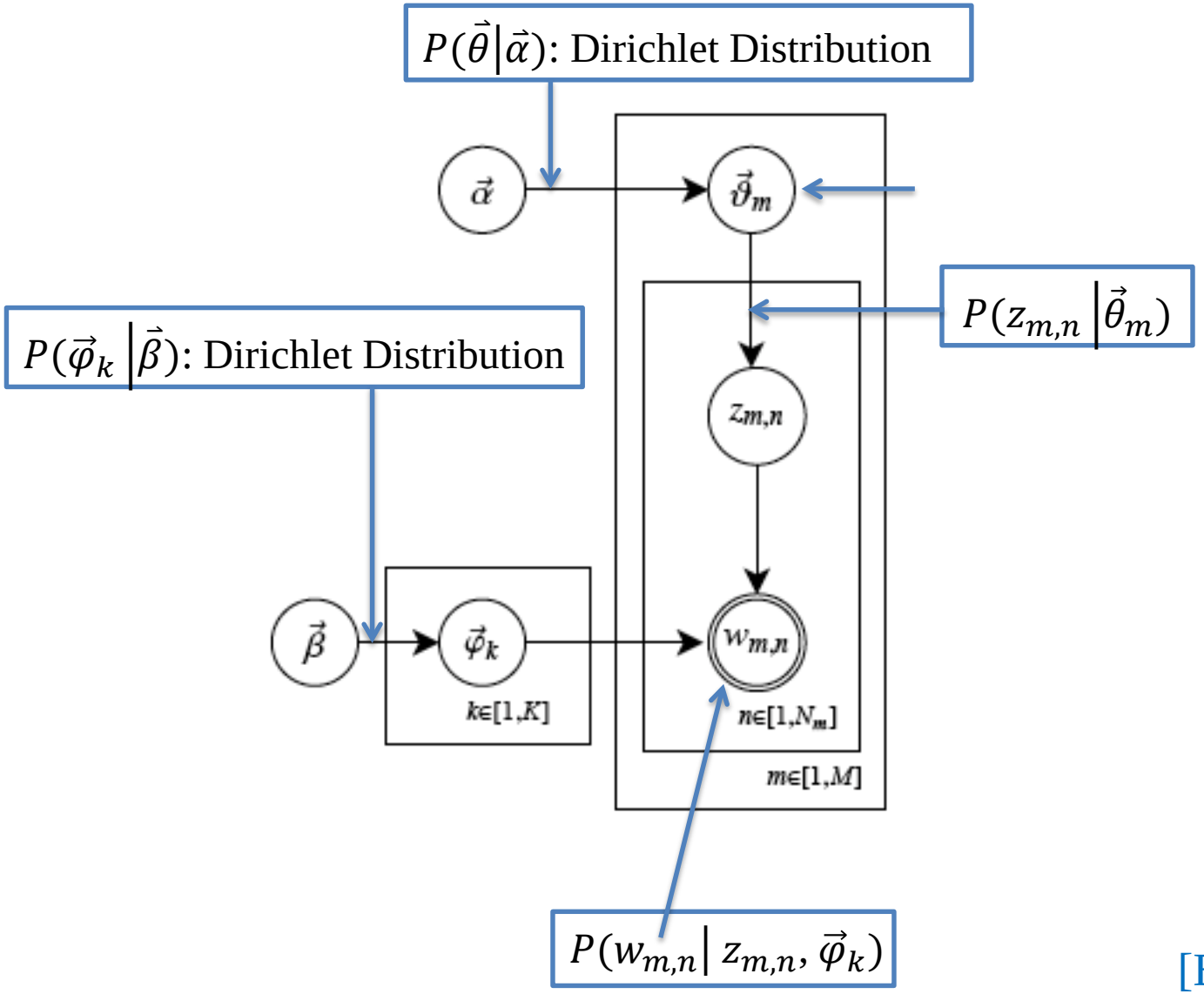




[Hofmann 1999]

Latent Dirichlet Allocation (LDA)

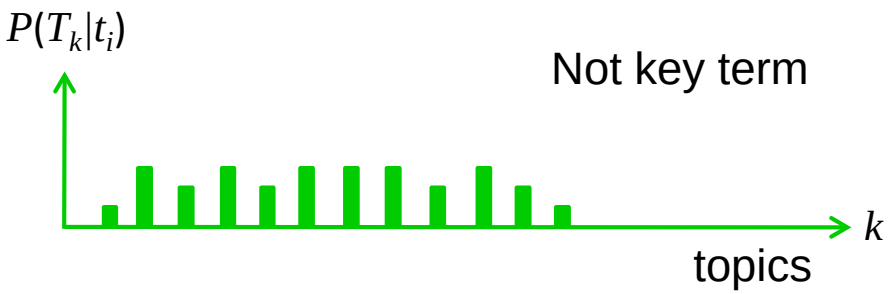
- **Unsupervised Learning of Topics from text corpus**



[Blei 2003]

Key Term Extraction and Summarization for Spoken Content

- **Key term extraction based on semantic features from PLSA or LDA**
  - key terms usually focused on smaller number of topics

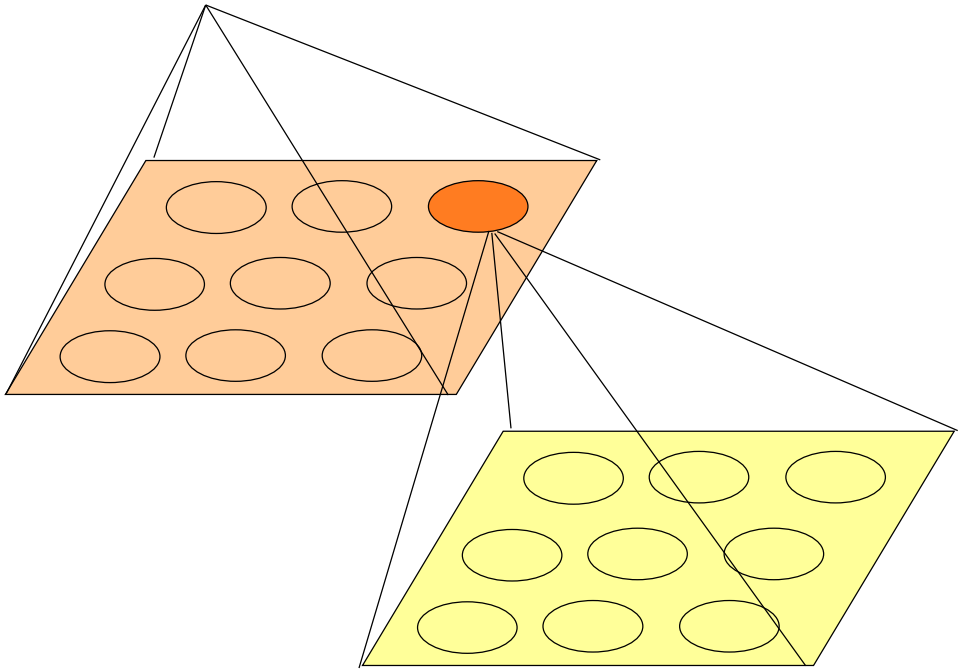


- **Summarization**
  - selecting most representative utterances but avoiding redundancy

[ICASSP 2006] [SLT 2010]

Semantic Structuring of Spoken Content (1/2)

- **Constructing the Semantic Structures of the Spoken Content**
- **Example Approach 1: Spoken Content categorized by Topics and organized in a Two-dimensional Tree Structure (2005)**
  - each category labeled by a set of key terms (topic) located on a map
  - categories nearby on the map are more related semantically
  - each category expanded into another map in the next layer



[Eurospeech 2005]

An Example of Two-dimensional Trees

- **Broadcast News Browser (2006)**

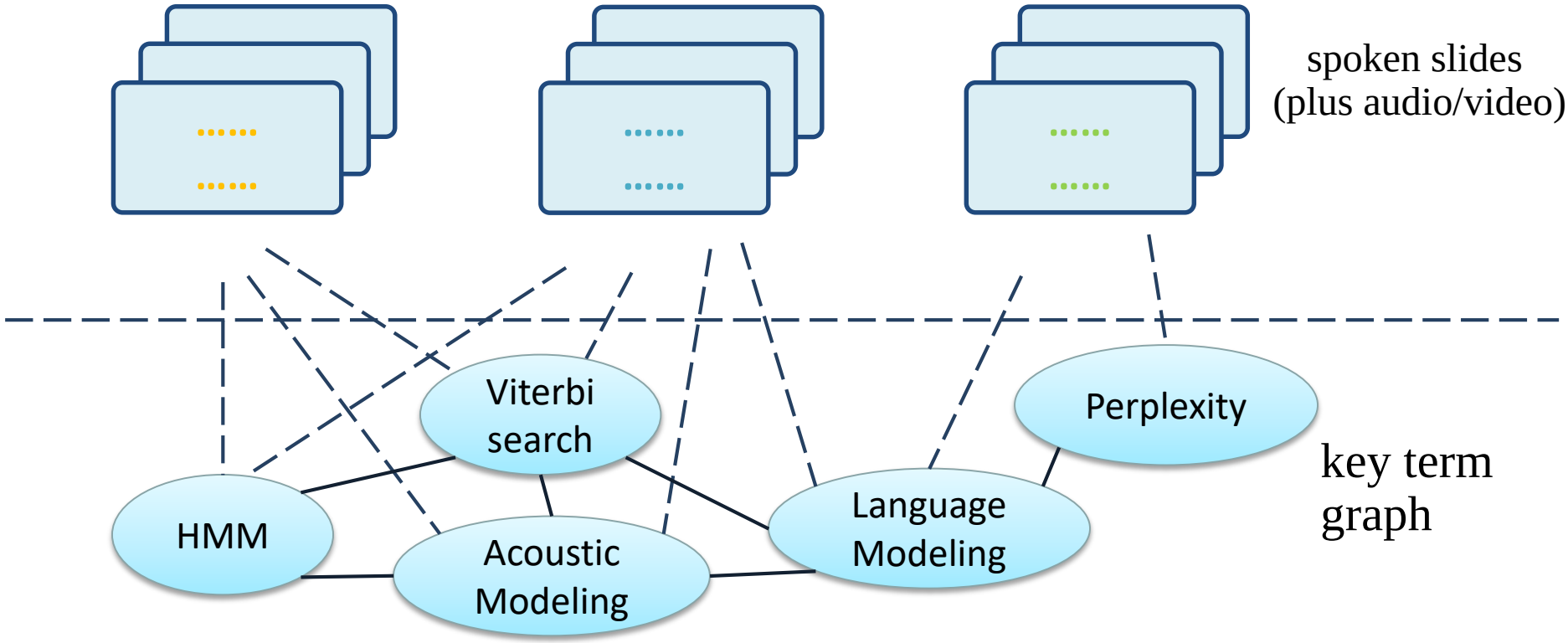
[Interspeech 2006]

Online Courses

- **Sequential knowledge transfer lecture by lecture**
- **When a lecture in an online course is retrieved for a user**
  - difficult for the user to understand this lecture without listening to previous related lectures
  - not easy to find out background or related knowledge

Semantic Structuring of Spoken Content (2/2)

- **Example Approach 2: Key Term Graph (2009)**
  - each spoken slide labeled by a set of key terms (topics)
  - relationships between key terms represented by a graph

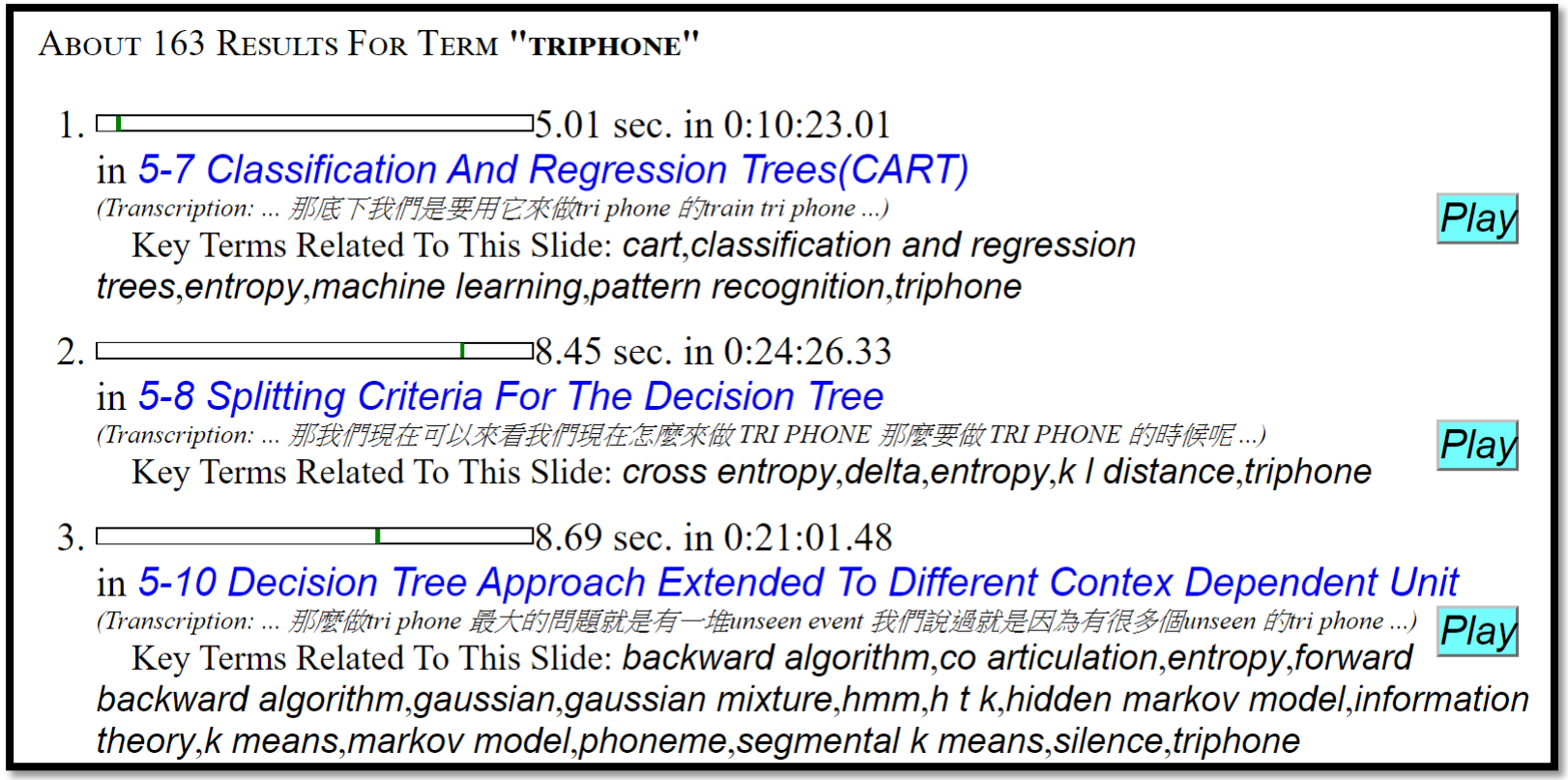


- **Very Similar to Knowledge Graph**

[ICASSP 2009][IEEE Trans ASL 2014]

An Example of Retrieving with an Online Course Browser (1/2)

- **Course : Digital Speech Processing (2009)**
  - Query : “triphone”
  - retrieved utterances shown with the spoken slides they belong to specified by the titles and key terms



[ICASSP 2009][IEEE Trans ASL 2014]

An Example of Retrieving with an Online Course Browser (2/2)

- **User clicks to view the spoken slide (2009)**
  - including a summary, key terms and related key terms from the graph
  - recommended learning path for a specific key term

5-7 Classification And Regression Trees(CART)

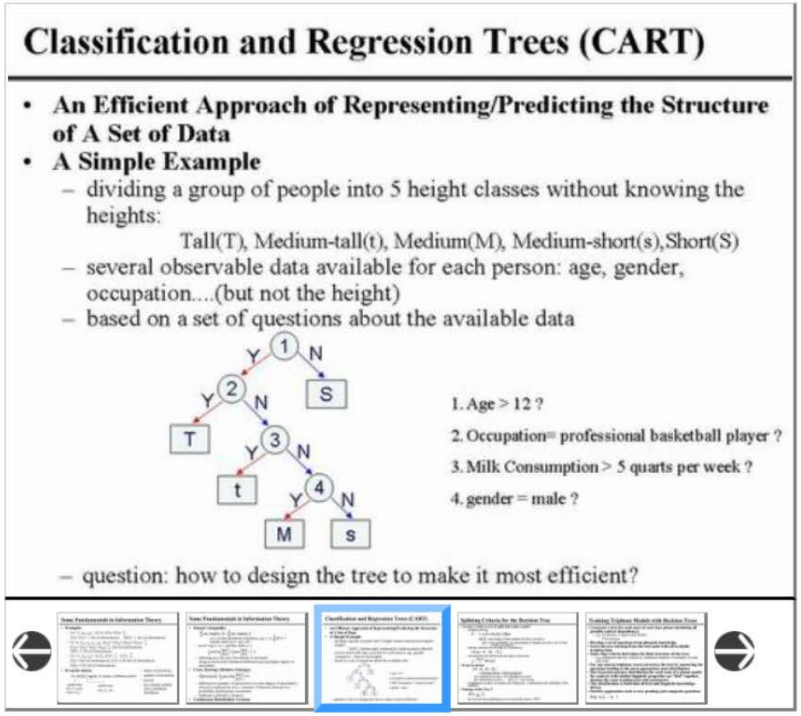
LENGTH:
0:10:23.0
TIME SPAN OF THIS CHAPTER:
CH 1 CH 18
TIME SPAN OF THIS SLIDE:
5-1 5-13

Play Summary (0:01:2.3) | Play Whole

KEY TERMS:

triphone	entropy	pattern recognition	m
phone	n gram	machine learning	p
hmm	triphone	eigen value	e
gaussian	language model	l d a	e
language model	hmm	gaussian	g
syllable	information theory	m c e	l

Related Key Terms

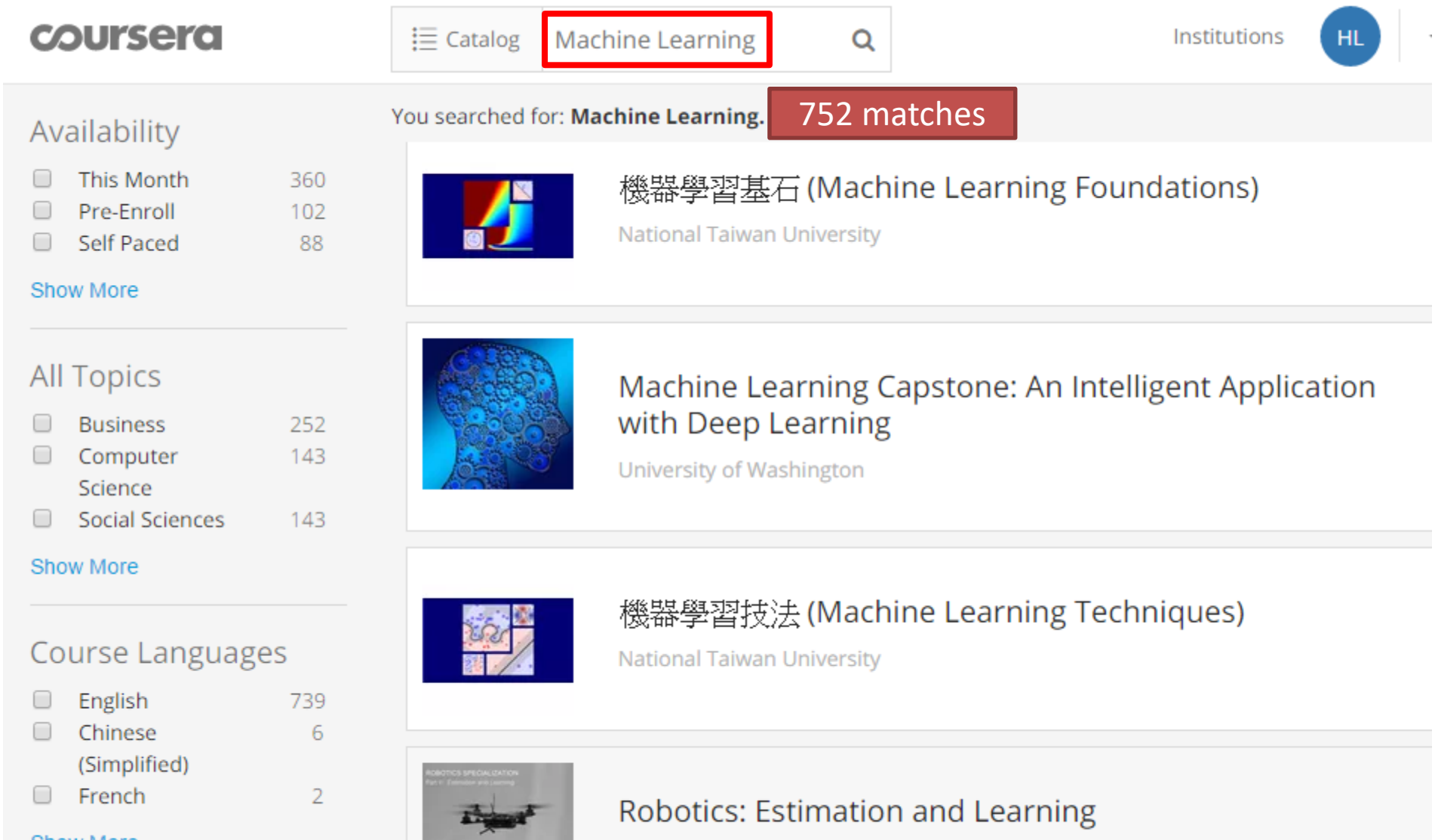


This key term(entropy) first appears in 5-4
Also appears in slide(s): 5-5 5-6 5-7 5-8 5-9 5-10 6-1 6-2 6-5 6-10 9-5 12-1 12-8 13-6

[ICASSP 2009][IEEE Trans ASL 2014]

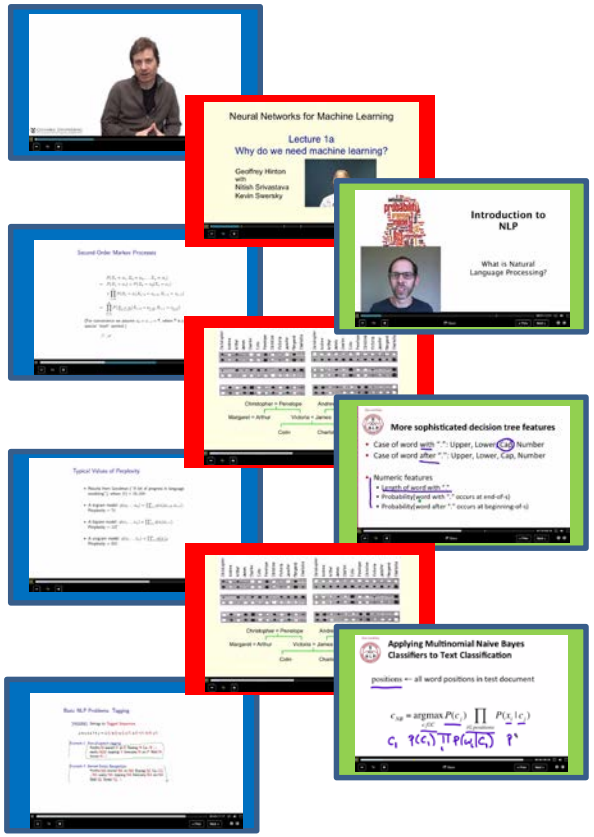
A Huge Number of Online Courses

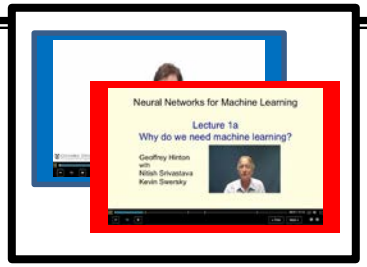
- **A user enters a keyword or a key phrase to coursera**

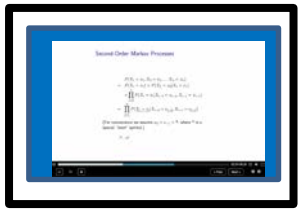


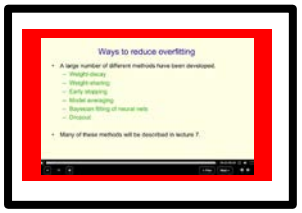
Having Machines Listen to all the Online Courses

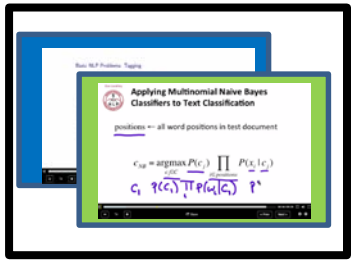
three courses on
some similar topic









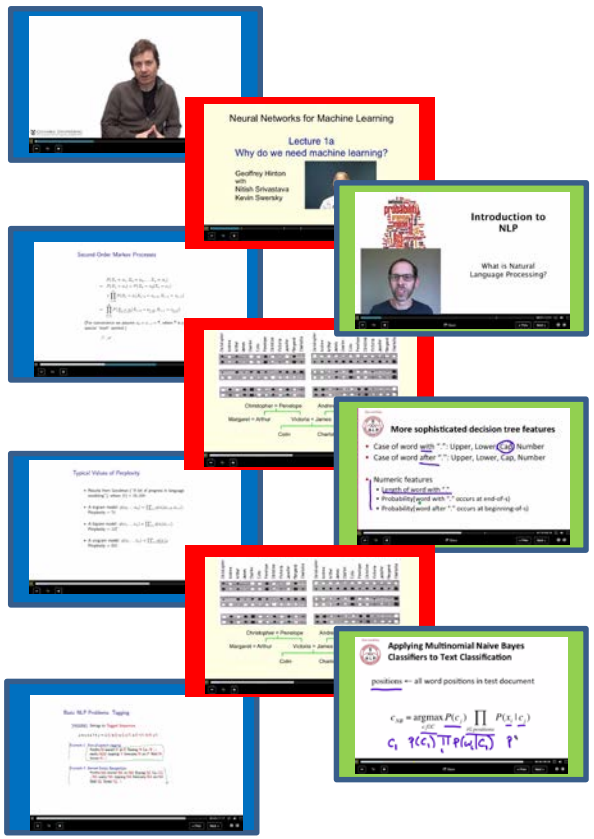


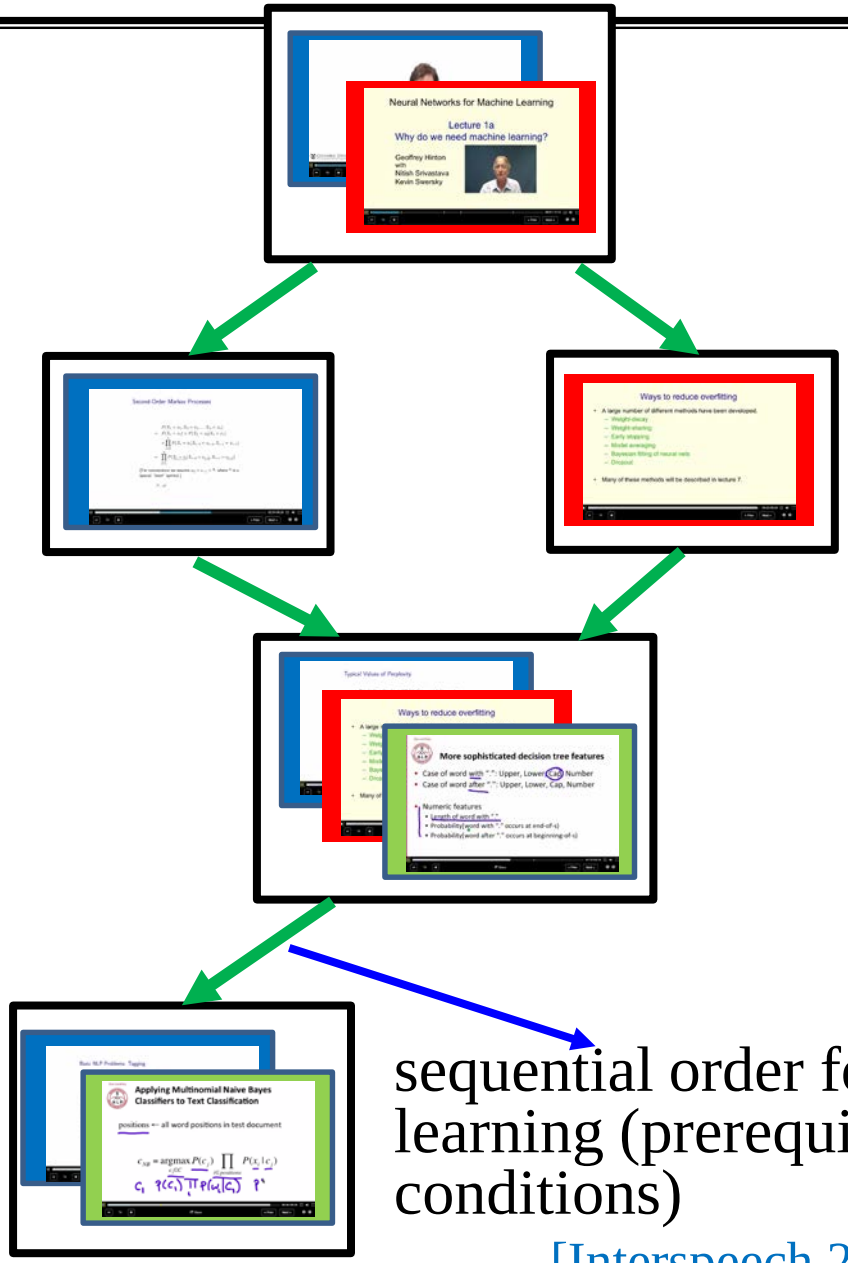
Lectures with very
similar content

[Interspeech 2015]

Having Machines Listen to all the Online Courses

three courses on
some similar topic





sequential order for learning (prerequisite conditions)

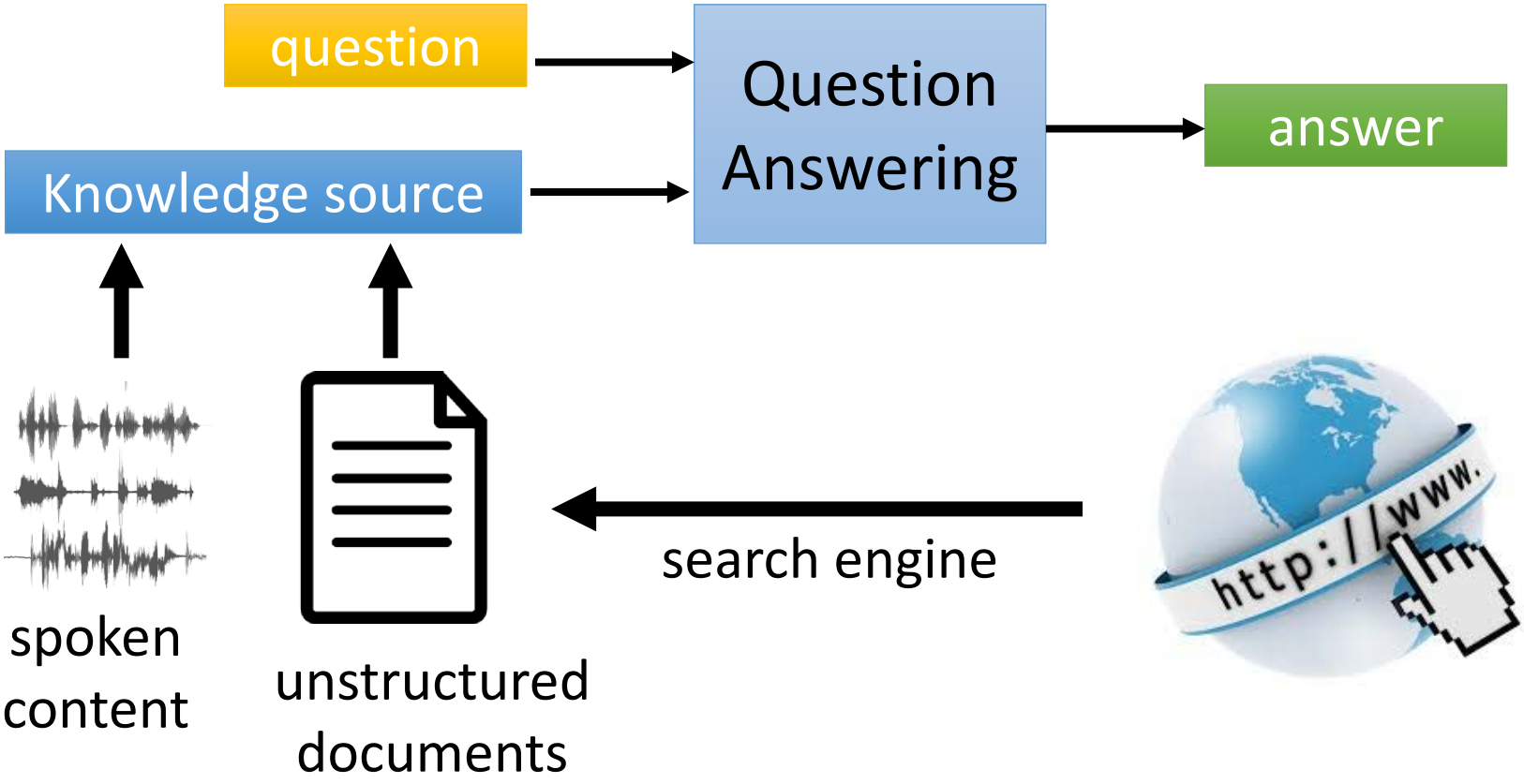
[Interspeech 2015]

Demonstration
Learning Map Produced by Machine (2014)

[Interspeech 2015]

Question Answering

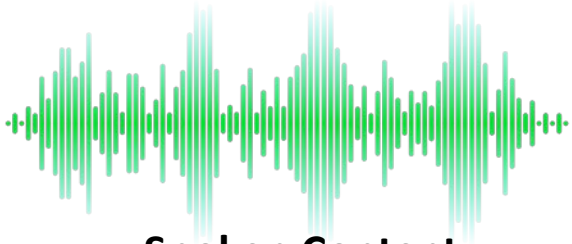
- **Machine answering questions from the user**



Text v.s. Spoken QA (Cascading v.s. End-to-end)

- **Text QA**

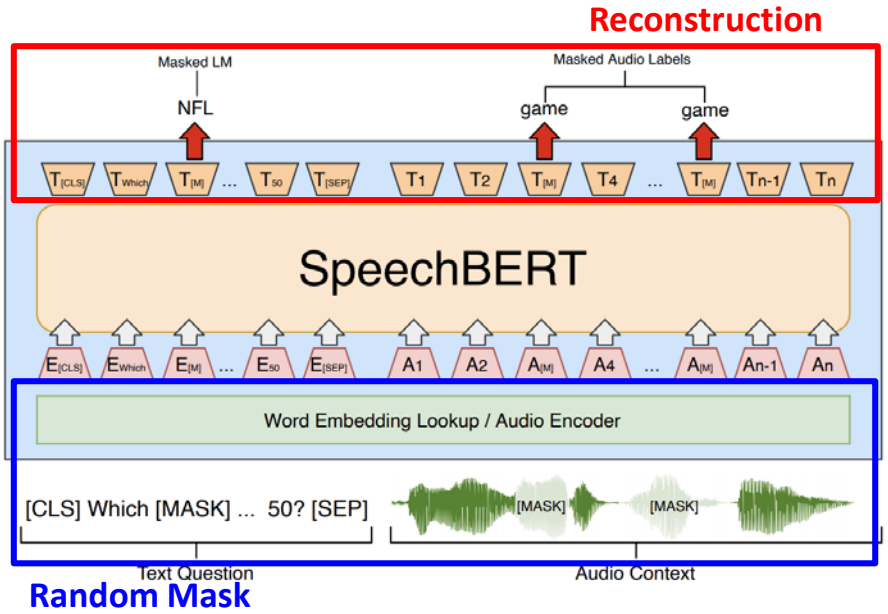
- **Spoken QA**



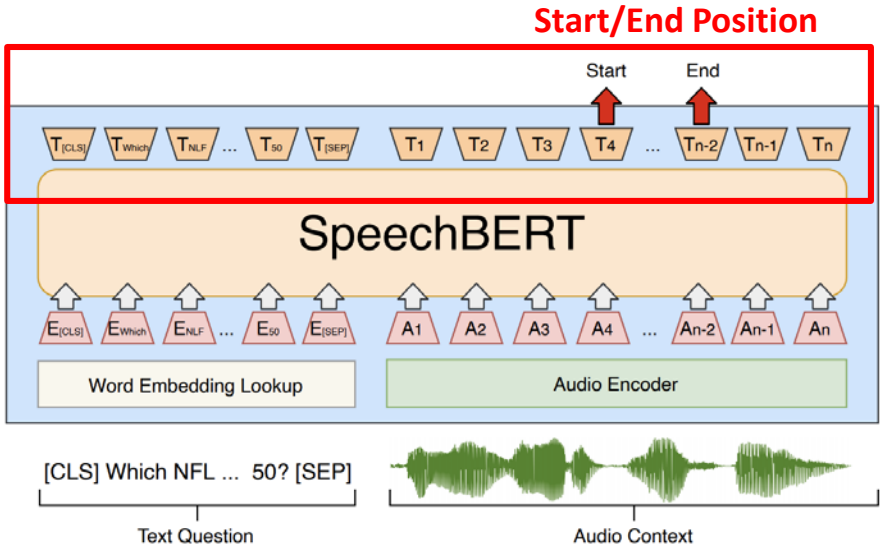
[Interspeech 2020]

Audio-and-Text Jointly Learned SpeechBERT

- **Pre-training**



- **Fine-tuning**



- **End-to-end Globally Optimized for Overall QA Performance**
  - not limited by ASR errors (no ASR here)
  - extracting semantics directly from speech, not from words via ASR

[Interspeech 2020]

Mini-Remarks

- **Internet is the only largest archive for global human knowledge**
  - spoken language technologies offer a bridge to that knowledge
- **Still a long way to go towards the personalized education environment considered**
  - many small steps may lead to the final destination some day
  - someone may realize it at the right time in the future
- **Machines capable of handling huge quantities of Data**
  - using machines in the way they are specially capable and efficient

(3) Unsupervised ASR (Phoneme Recognition)

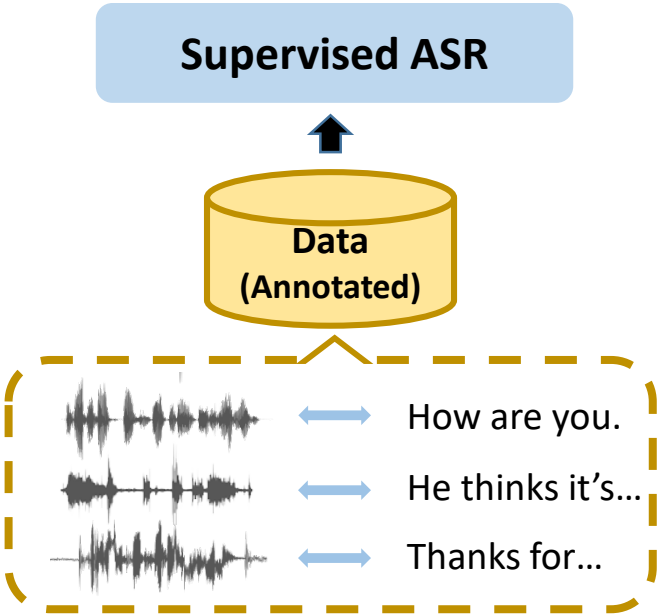
Hung-yi Lee (left) and Lin-shan Lee

Supervised/Unsupervised ASR

- **Supervised ASR**
  - Has been very successful
  - Problem : requiring a huge quantity of annotated data

    - Thousands of languages spoken over the world
      - most are low-resourced without enough annotated data

- **Unsupervised ASR**
  - Train without annotated data

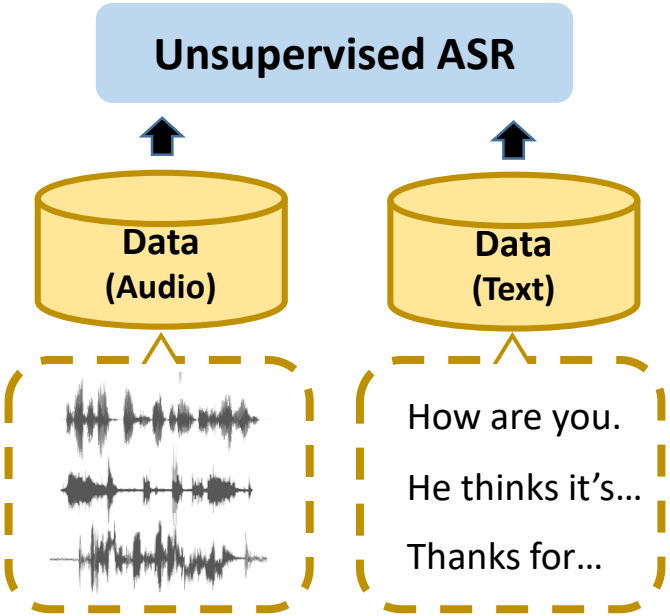


Supervised/Unsupervised ASR

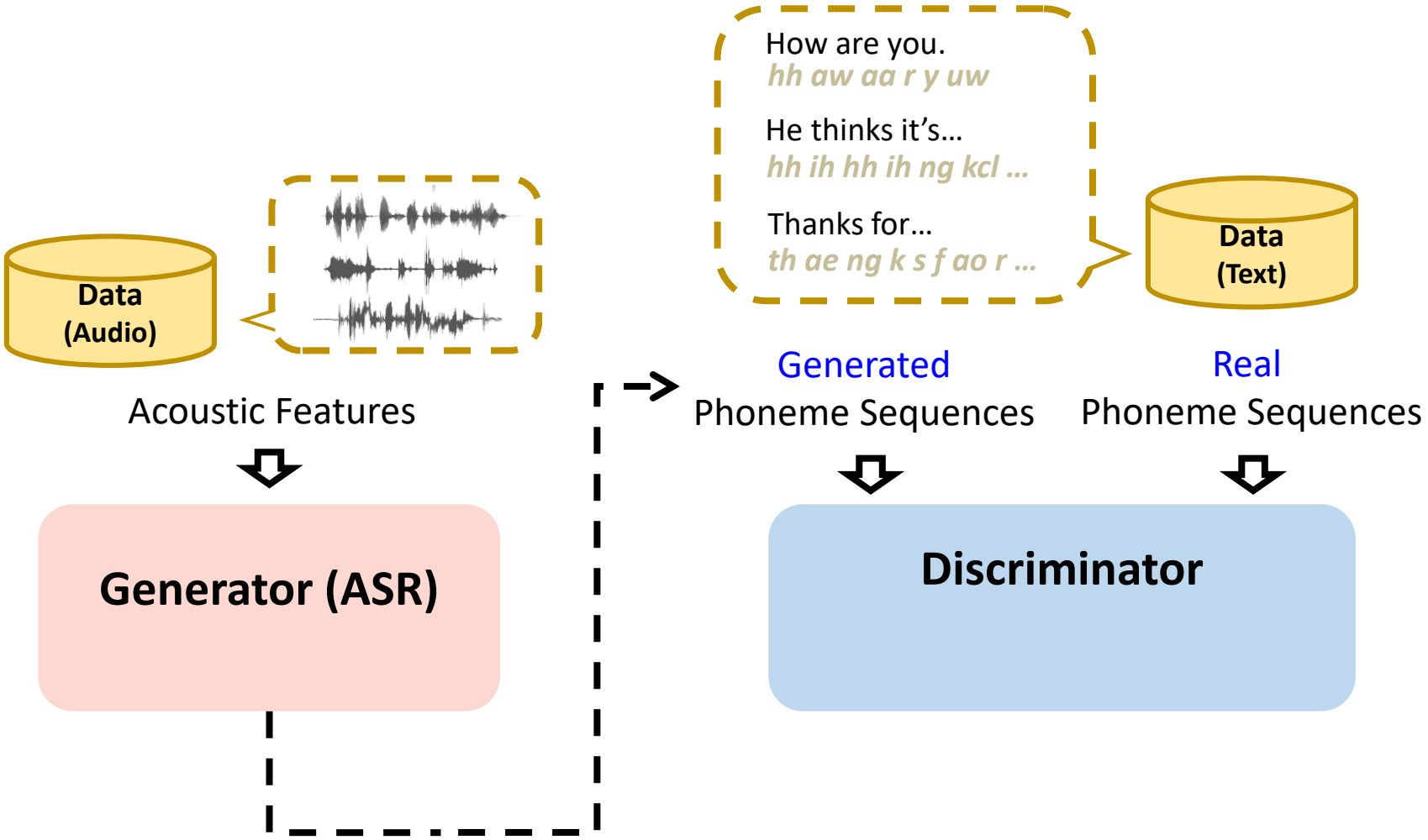
- **Supervised ASR**
  - Has been very successful
  - Problem : requiring a huge quantity of annotated data

    - Thousands of languages spoken over the world
      - most are low-resourced without enough annotated data

- **Unsupervised ASR**
  - Train without annotated data
  - Unlabeled, unpaired data are easier to collect
  - something we could Never do before (2018)

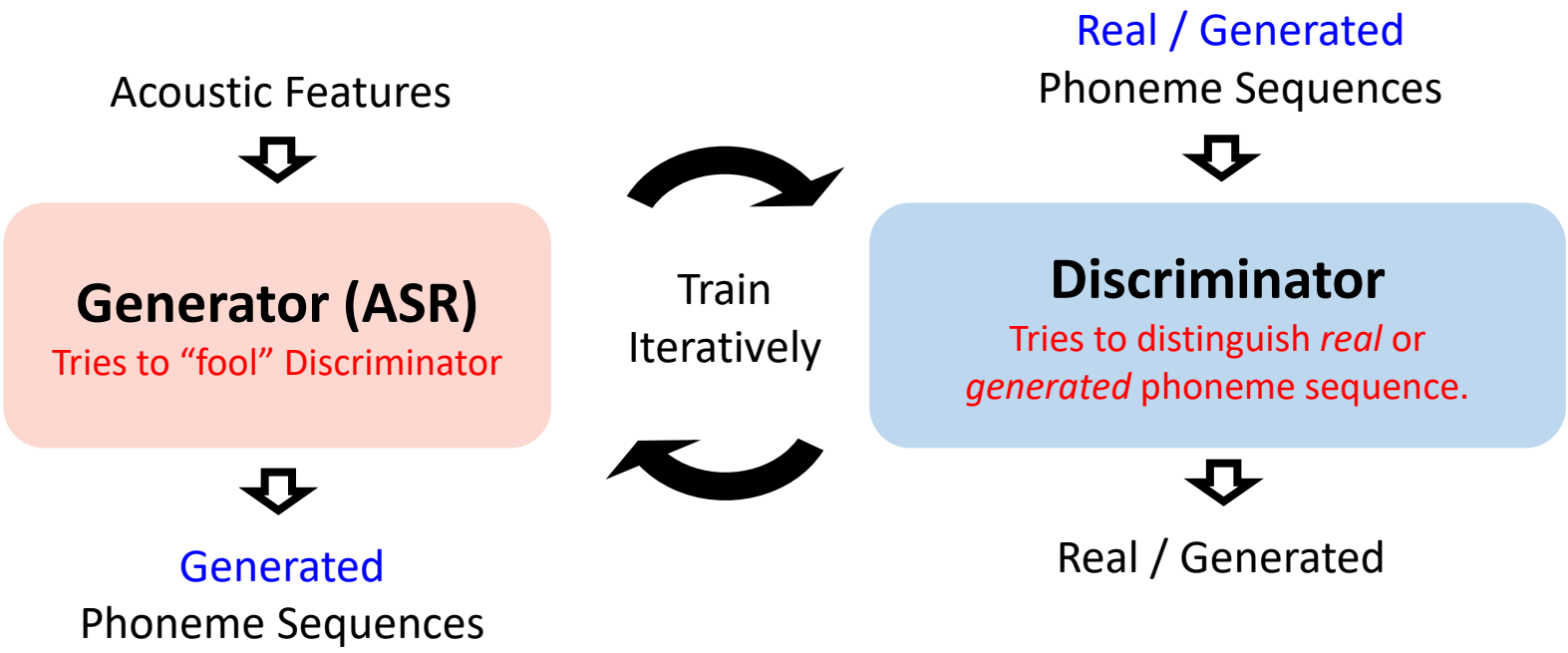


Use of Generative Adversarial Networks (GAN)



Use of Generative Adversarial Networks (GAN)

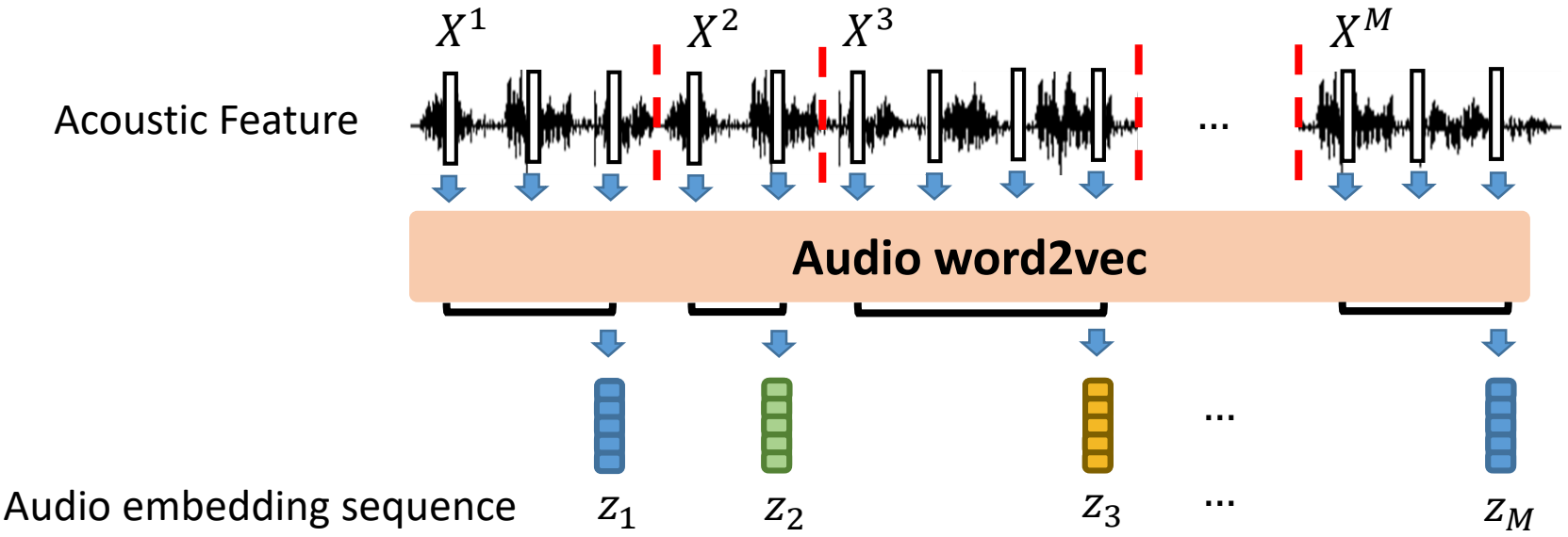
- **Generative Adversarial Network (GAN)**



- **Discriminator / Generator improve themselves individually and iteratively**

Model 1 (2018)

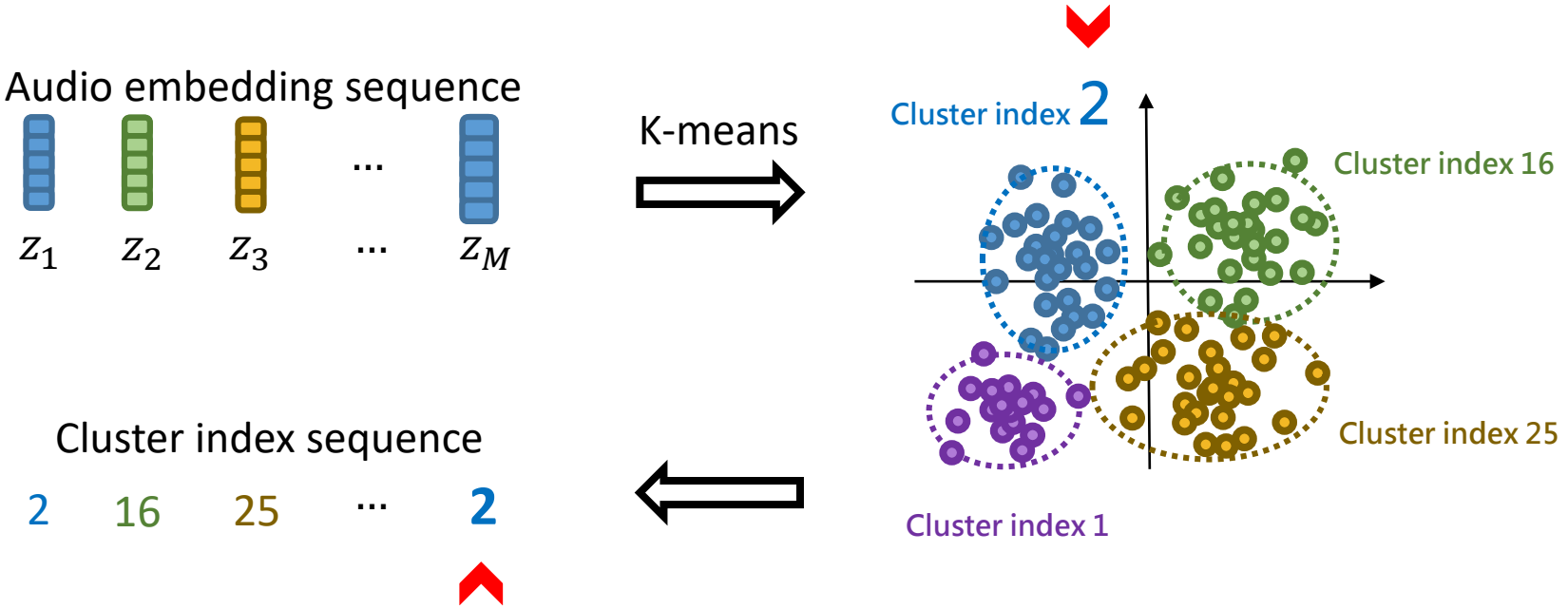
- **Waveform segmentation and embedding**
  - divide the features into acoustically similar segments of different lengths
  - transform each segment into a fixed-length vector (audio embedding)



[Interspeech 2018]

Model 1 (2018)

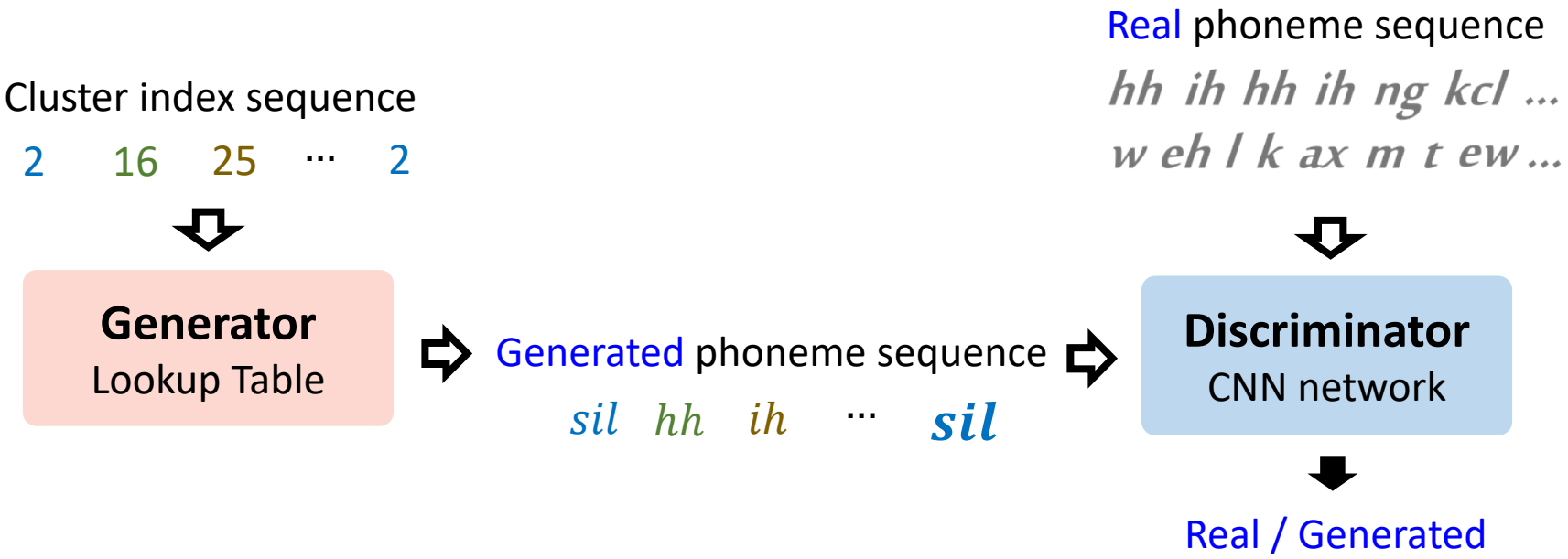
- **Cluster the embeddings into groups**



[Interspeech 2018]

Model 1 (2018)

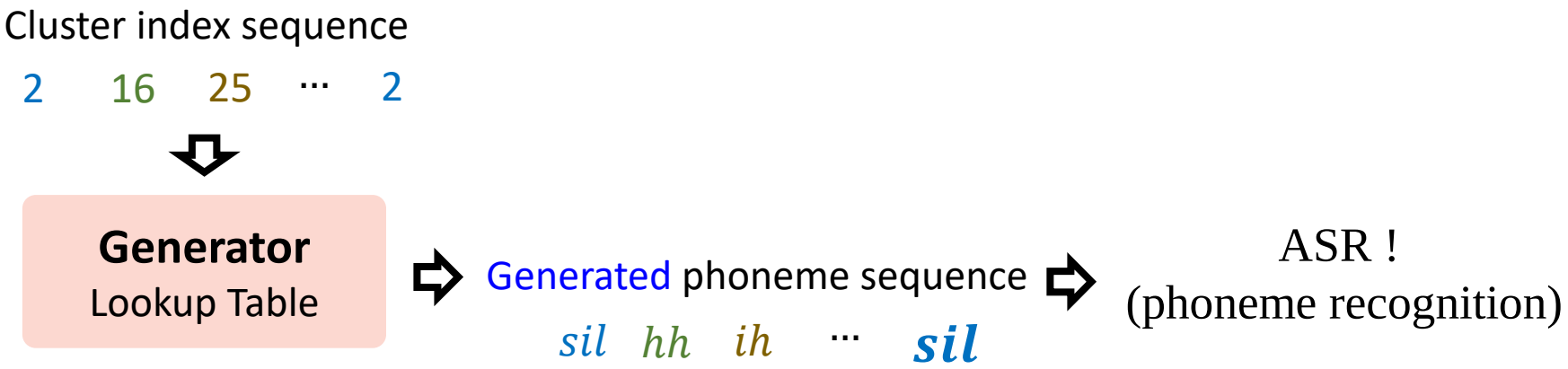
- **Learning the mapping between cluster indices and phonemes with a GAN**
  - embedding clustering followed by (cascaded with) a GAN



[Interspeech 2018]

Model 1 (2018)

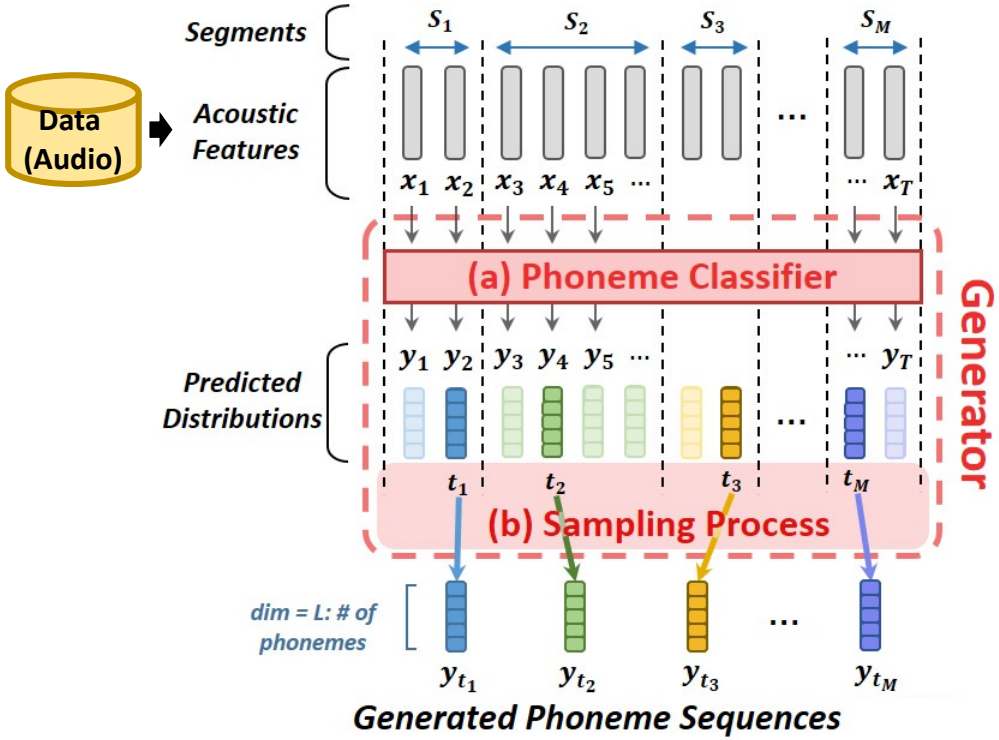
- **Learning the mapping between cluster indices and phonemes with a GAN**
  - embedding clustering followed by (cascaded with) a GAN



[Interspeech 2018]

Model 2 (2019)

- **Generator** consists of two parts
  - **(a) Phoneme Classifier (DNN)**
  - **(b) Sampling Process**

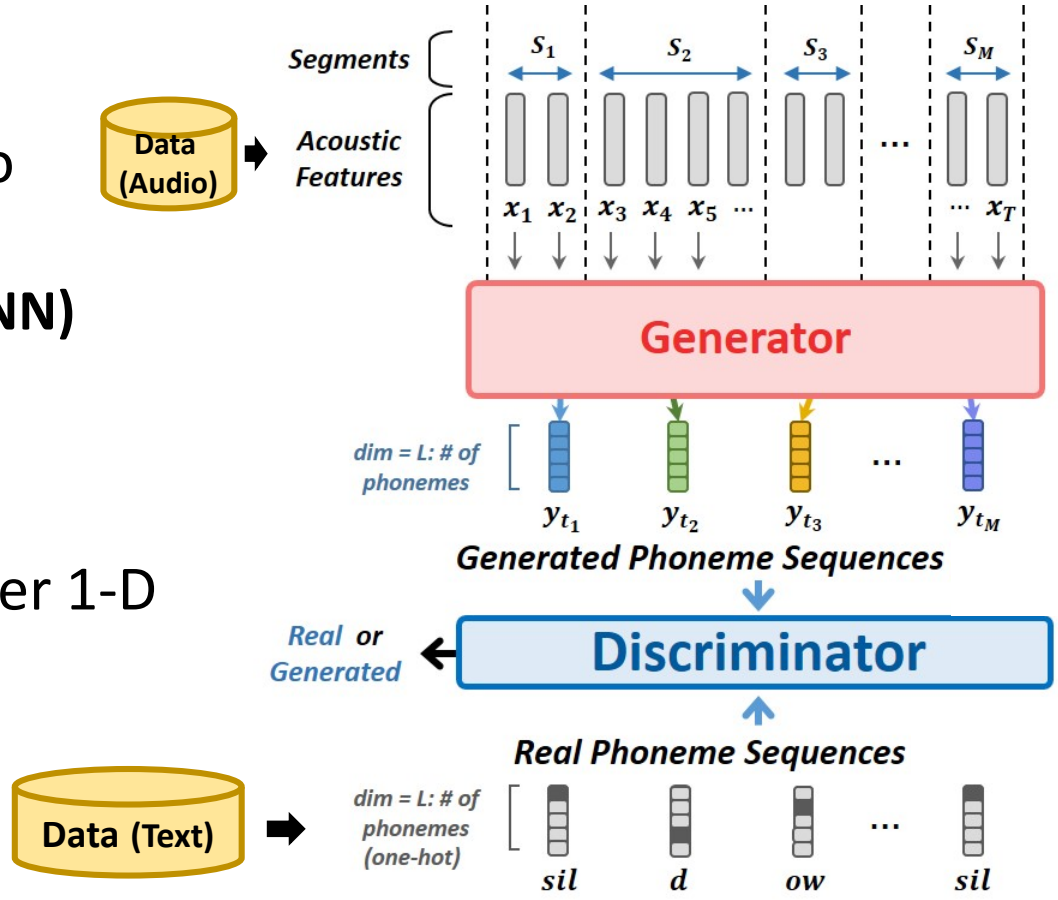


[Interspeech 2019]

Model 2 (2019)

- **Generator** consists of two parts
  - **(a) Phoneme Classifier (DNN)**
  - **(b) Sampling Process**

- **Discriminator** is a two layer 1-D CNN.



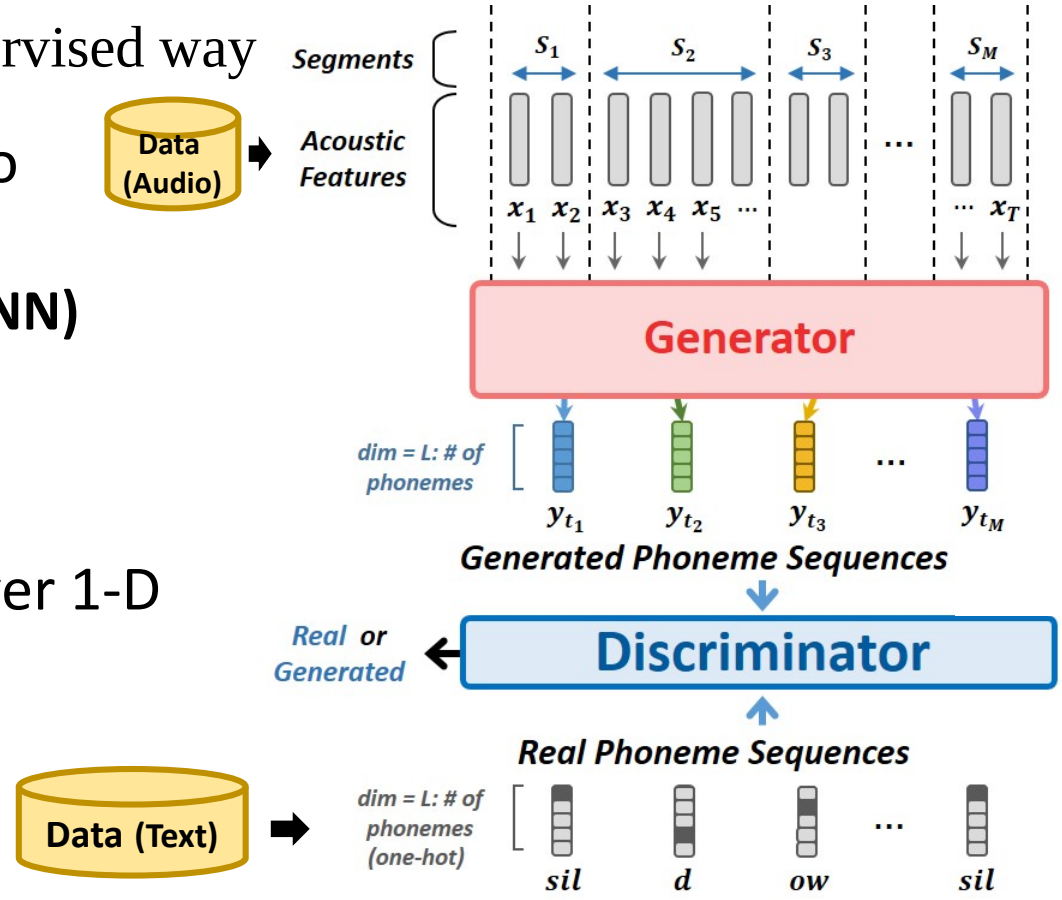
[Interspeech 2019]

Model 2 (2019)

- **A GAN (Generator and Discriminator) trained End-to-end**
  - DNN trained in an unsupervised way

- **Generator** consists of two parts
  (a) Phoneme Classifier (DNN)
  (b) Sampling Process

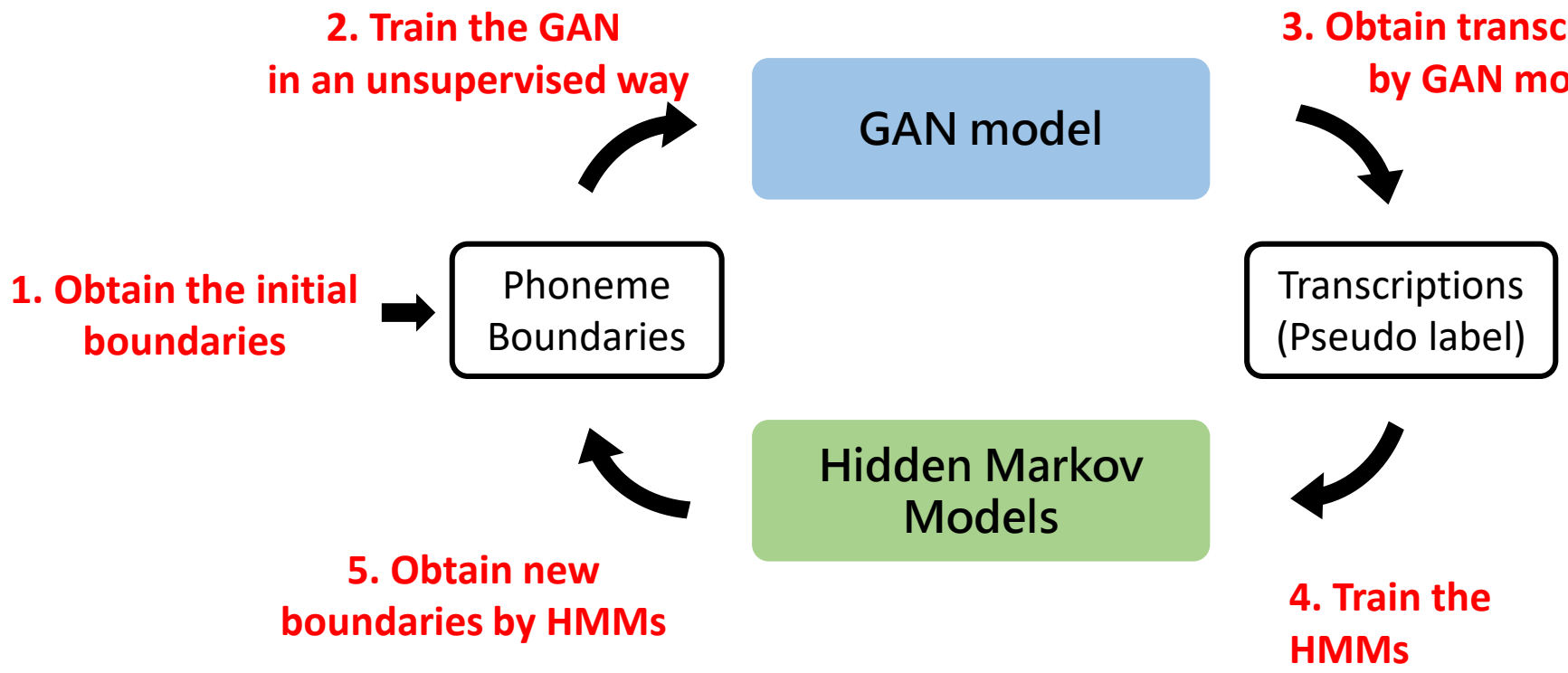
- **Discriminator** is a two layer 1-D CNN.



[Interspeech 2019]

Model 2 (2019)

- **GAN iterated with HMMs**



[Interspeech 2019]

Experimental Results for Models 1, 2 on TIMIT

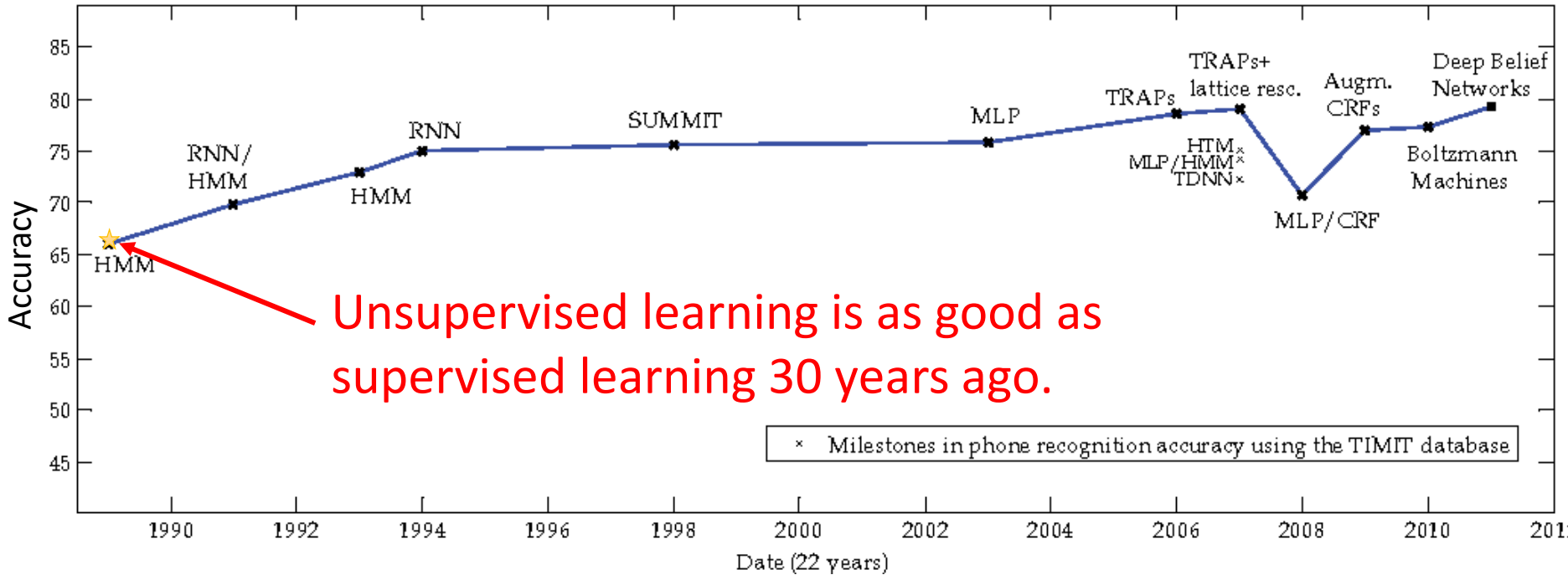
- **Phoneme Error Rate**

Approaches			PER	
			Matched	Unrelated
Supervised				
RNN Transducer			17.7	-
Standard HMMs			21.5	-
Completely unsupervised (no label at all)				
Model 1			76.0	-
Model 2	Iteration 1	GAN	48.6	50.0
		HMM	30.7	39.5
	Iteration 2	GAN	41.0	44.3
		HMM	27.0	35.5
	Iteration 3	GAN	38.4	44.2
		HMM	26.1	33.1

- **Model 1 cascaded clustering with GAN, while Model 2 did everything end-to-end with GAN** [Interspeech 2018] [Interspeech 2019]

The Progress of Supervised Learning on TIMIT

- **Milestones in phone recognition accuracy**



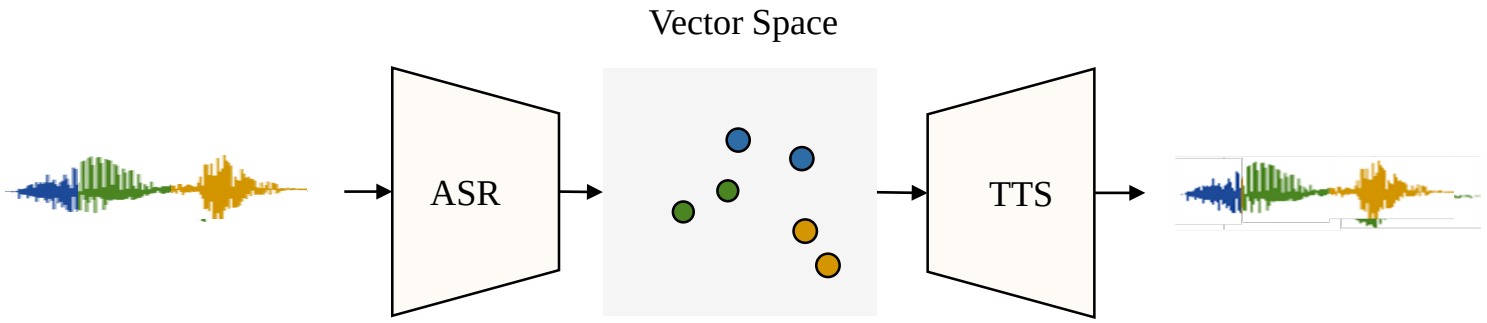
  - Will it take another 30 years for unsupervised learning to achieve the performance of supervised learning today ?

[Phone recognition on the TIMIT database, Lopes, C. and Perdigão, F., 2011. Speech Technologies, Vol 1, pp. 285--302.]

Model 3 (2019)

- **A Cascading Recognition-Synthesis Framework**

  - Learning basic sound units by listening and reproducing
  - Babies learn to speak from parents

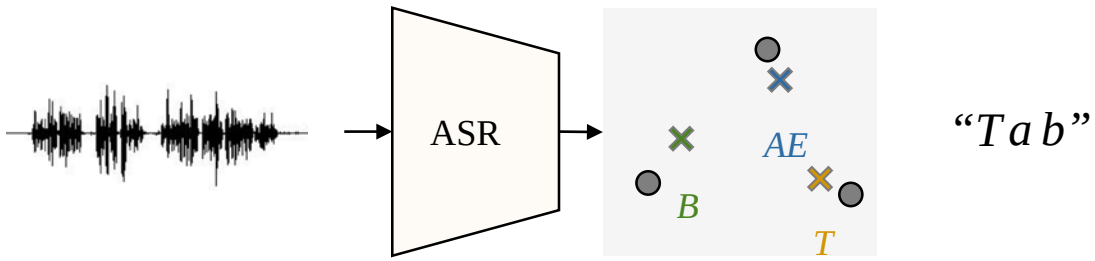


[ICASSP 2020]

Model 3 (2019)

- **A Cascading Recognition-Synthesis Framework**

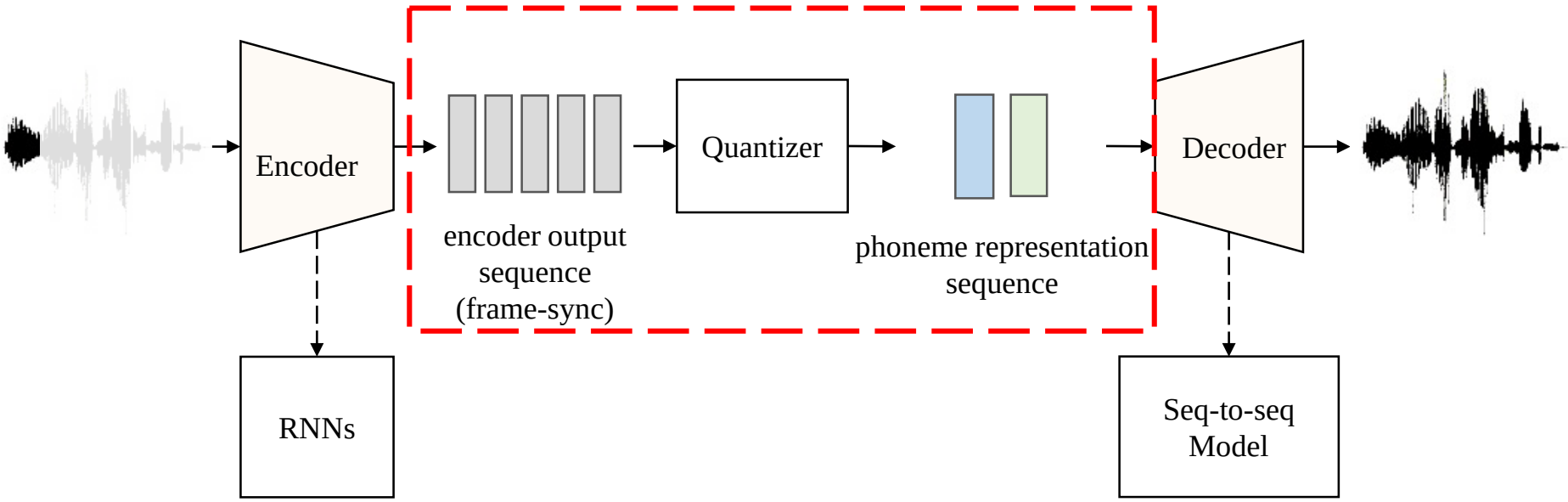
  - Learning basic sound units by listening and reproducing
  - Babies learn to speak from parents
  - ASR ! (Phoneme recognition)
  - Audio data only (no text) plus a small set of paired data



[ICASSP 2020]

Model 3 (2019)

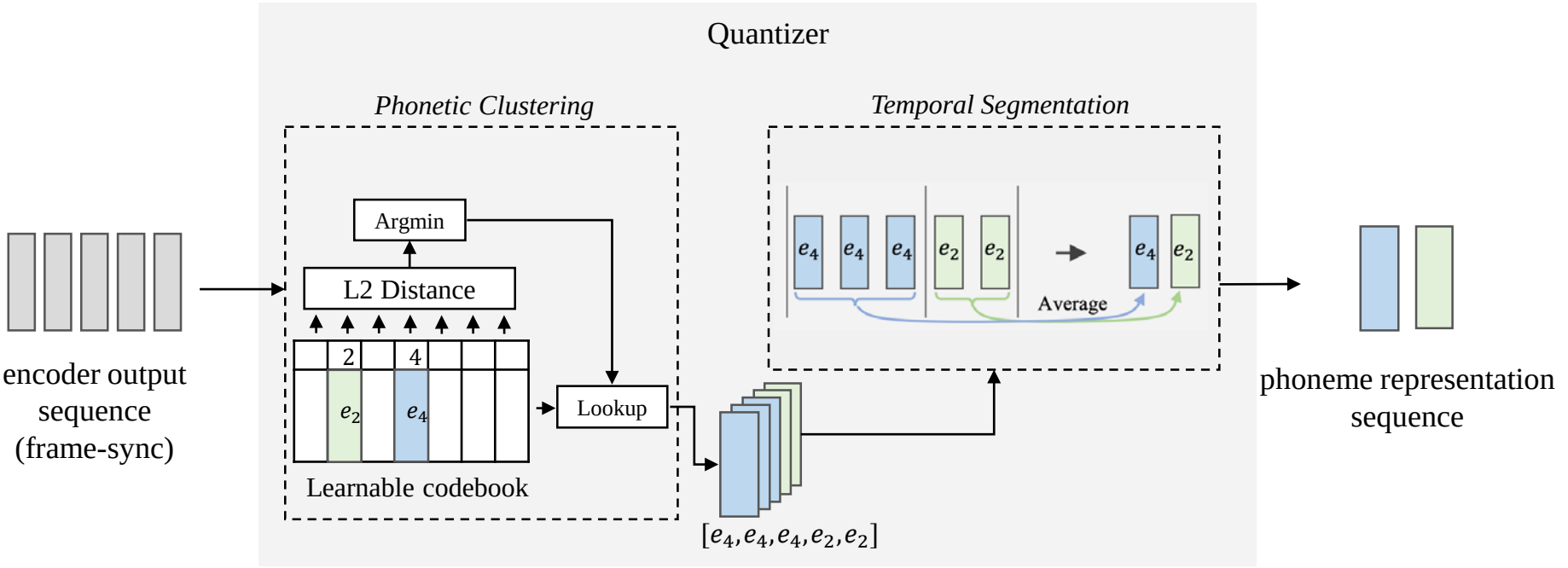
- **Implemented with a Sequential Quantization AutoEncoder**



[ICASSP 2020]

Model 3 (2019)

- **Implemented with a Sequential Quantization AutoEncoder**



[ICASSP 2020]

Model 3 (2019)

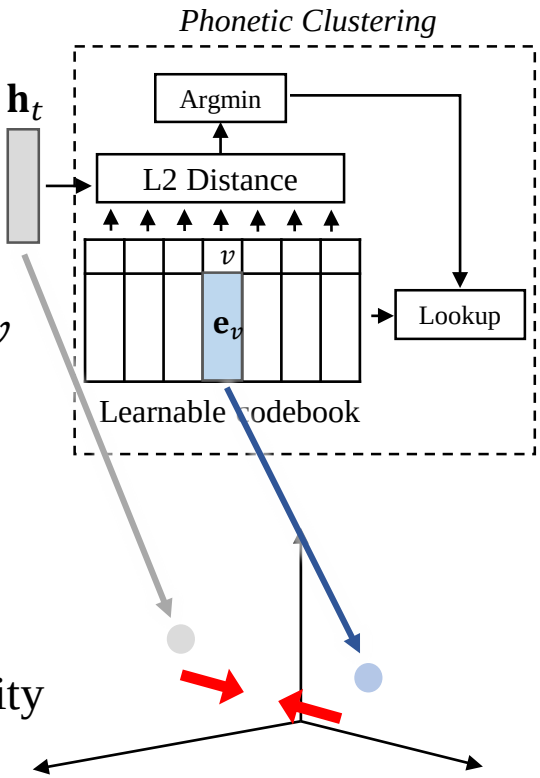
- **Sequential Quantization AutoEncoder**
  - With small quantity of paired data, train codebook to match real phonemes

1. Assign one phoneme to each codeword

2. Define the probability that vector $\mathbf{h}_t$ belongs to phoneme v

$$\Pr(\mathbf{h}_t \text{ belongs to } v) = \frac{\exp(-\|\mathbf{h}_t - \mathbf{e}_v\|_2)}{\sum_u \exp(-\|\mathbf{h}_t - \mathbf{e}_u\|_2)}$$

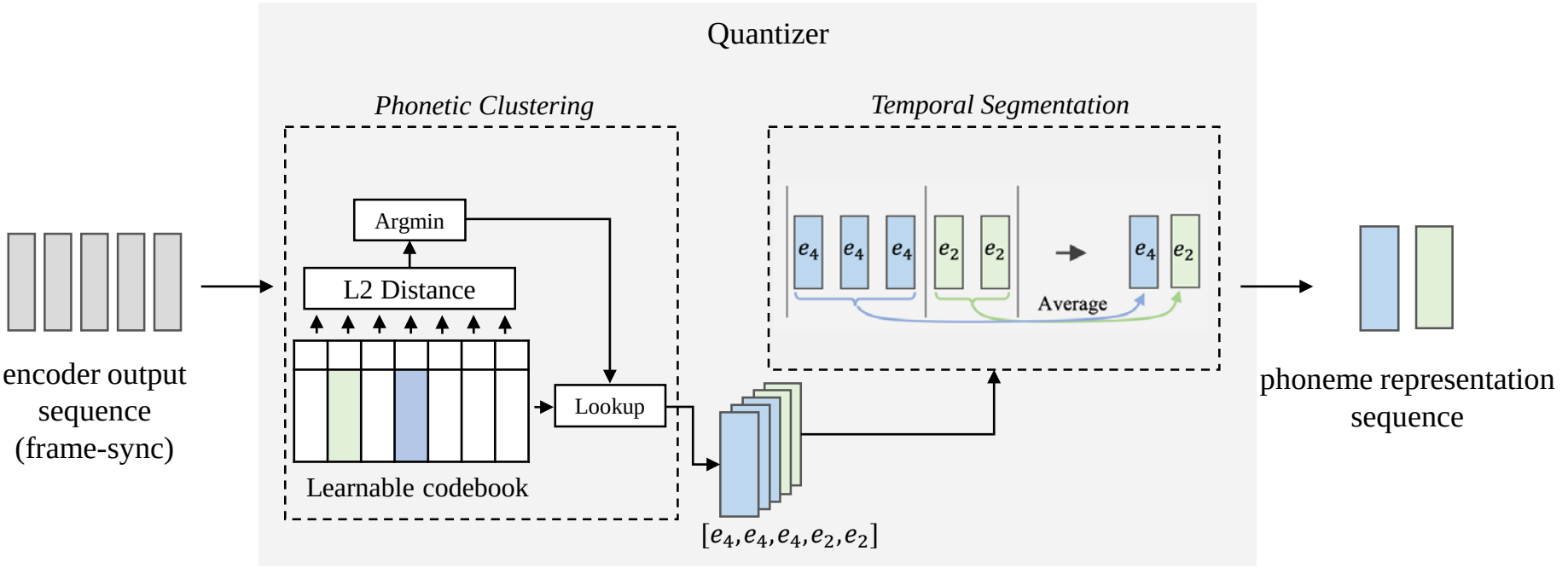
3. Matching with the paired data by maximizing the probability



[ICASSP 2020]

Model 3 (2019)

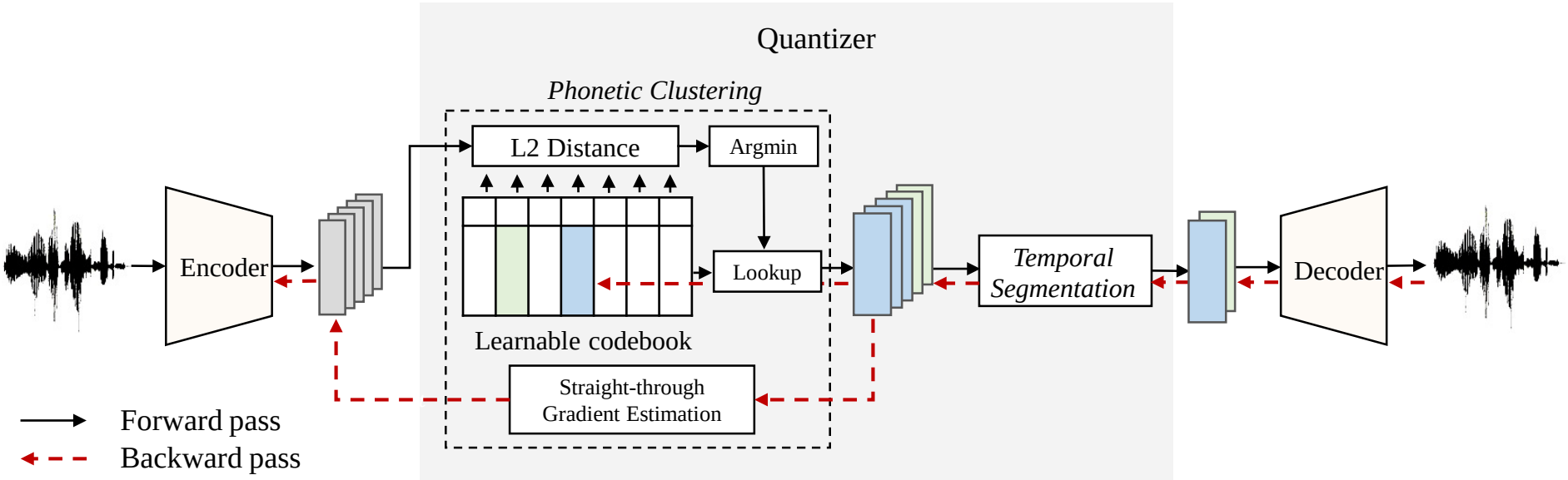
- **Implemented with a Sequential Quantization AutoEncoder**



[ICASSP 2020]

Model 3 (2019)

- **Implemented with a Sequential Quantization AutoEncoder**

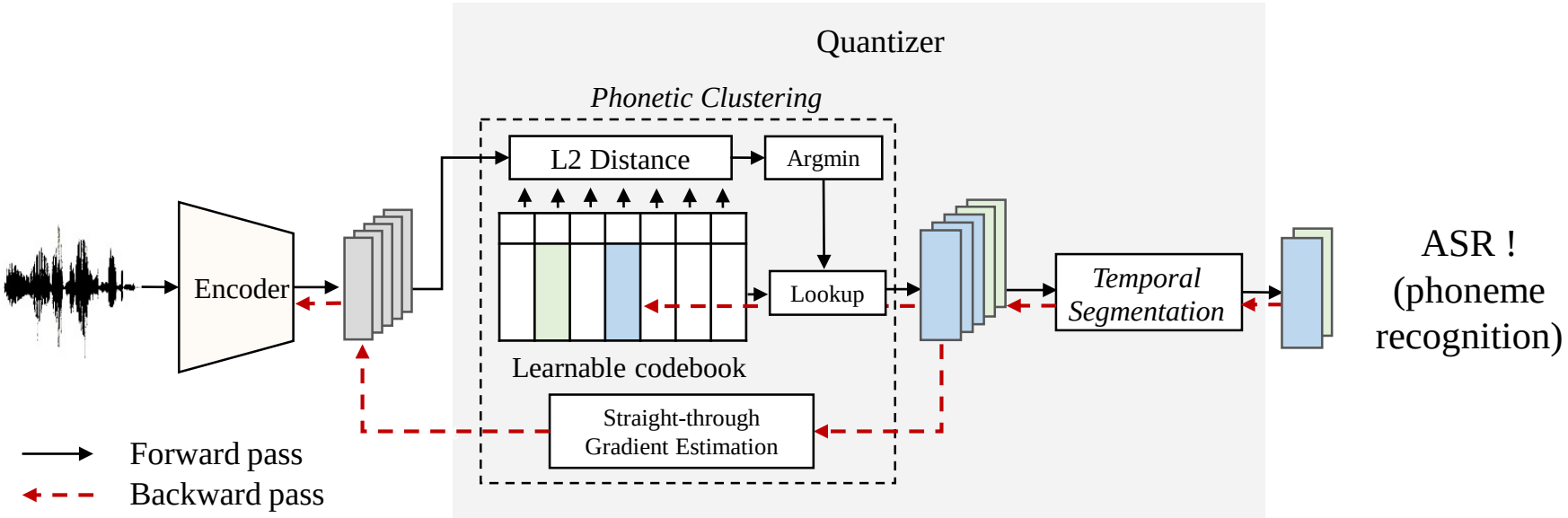


  - End-to-end trained (with enough audio data plus small paired data)

[ICASSP 2020]

Model 3 (2019)

- **Implemented with a Sequential Quantization AutoEncoder**



  - End-to-end trained (with enough audio data plus small paired data)

[ICASSP 2020]

Experiment Results for Model 3 on LJSpeech

- **Phoneme Error Rate (PER)**
  - when different amount of paired data is available (single speaker)

	Unlabeled Audio	20 min. Paired	15 min. Paired	10 min. Paired	5 min. Paired
Proposed Method	22 hrs.	25.5	29.0	35.2	49.3

  - initial recognition observed, although 20 min of paired data used
  - related work reported recently

[ICASSP 2020]

Mini-Remarks

- **Deep learning makes tasks impossible before realizable today**
- **Much more crazy concepts are yet to be explored !**
- **Unpaired data may be more useful than we thought**
- **End-to-end training is more attractive than cascading in the context of deep learning**
  - overall performance is optimized globally as compared to locally

Concluding Remarks

- **Doing something we never could is interesting and exciting !**
- **Though challenging, the difficulties may be reduced by the fast advancing technologies including deep learning**
- **15 years ago we never knew what kind of technologies we could have today**
  - today we never know what kind of technologies we may have 15 years from now
  - anything in our mind could be possible
- **This may be the golden age we never had for research**
  - very deep learning, very big data, very powerful machines, very strong industry
  - which we never had before
  - possible to do something we Never could !