Automatic Speech Recognition - Part I

Dr. Mathew Magimai Doss

Outline

Automatic speech recognition problem formulation

Language generation (Sequence generation)

String Matching (Sequence comparison)

Knowledge-based approach

Instance-based ASR approach

Next Part

Outline

Automatic speech recognition problem formulation

Language generation (Sequence generation)

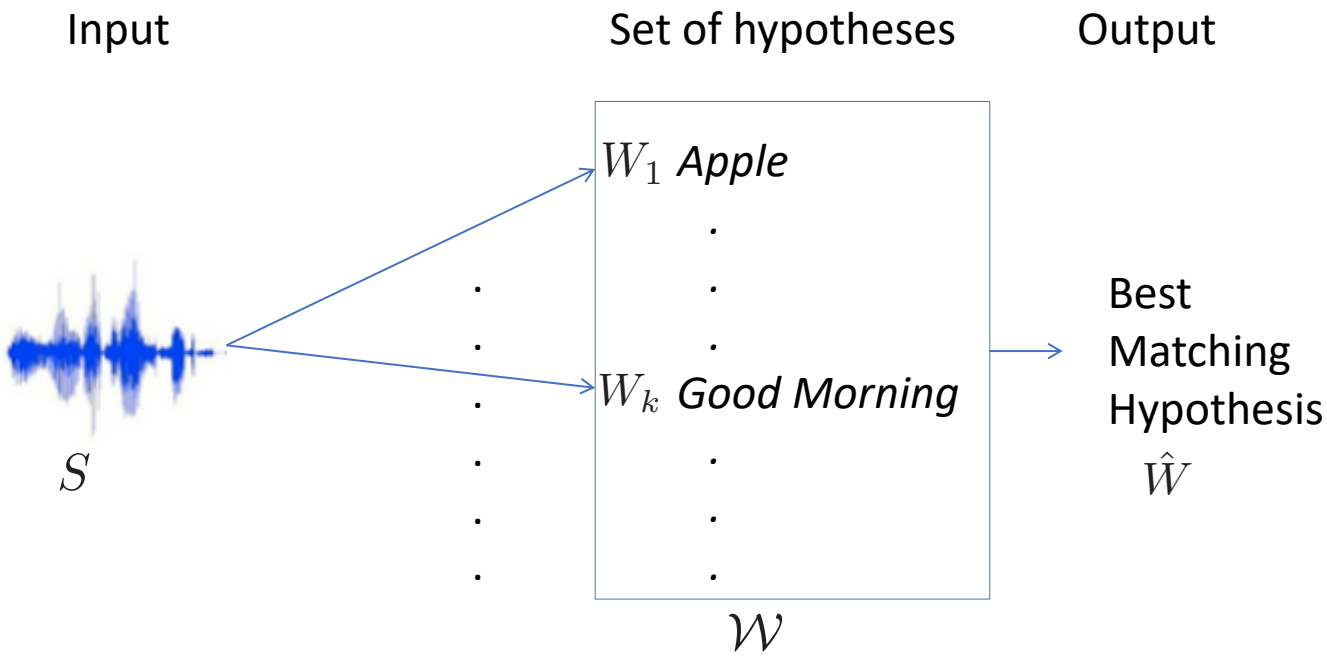
String Matching (Sequence comparison)

Knowledge-based approach

Instance-based ASR approach

Next Part

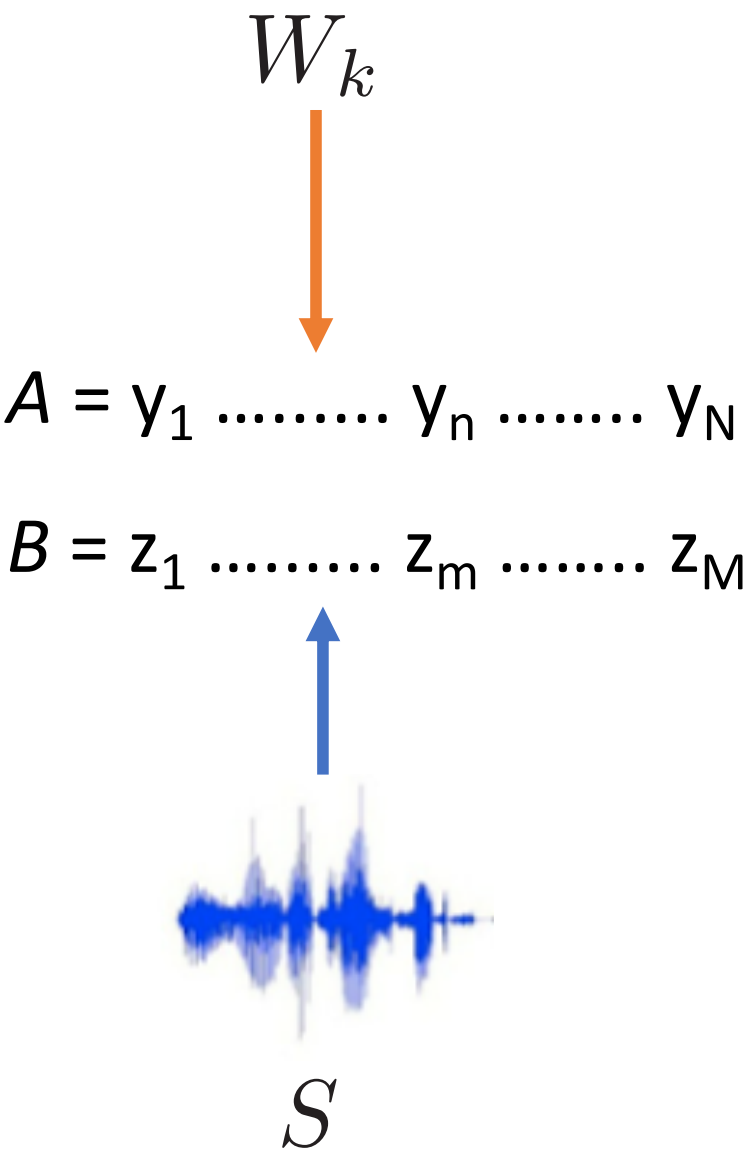
Automatic speech recognition (ASR)



$$\hat{W} = \underset{W_k \in \mathcal{W}}{\arg\max}\, \mathrm{Match}(W_k, S)$$

How to match an observed speech signal S with a word hypothesis W_k?

Abstract formulation for matching S and W_k



Core Idea

1. Map S and W_k to a shared latent symbol space
2. Match the resulting two latent symbol sequences A and B

Four sub questions

Q1: What is the shared latent symbol set?

Q2: How to map S to a latent symbol sequence B?

Q3: How to map W_k to a latent symbol sequence A?

Q4: How to match the two latent symbol sequences A and B?

Different ASR methods mainly differ on how these four sub questions are addressed.

ASR methods

1. Knowledge-based approach
2. Instance-based approach
3. Statistical sequence modeling-based approach

Outline

Automatic speech recognition problem formulation

Language generation (Sequence generation)

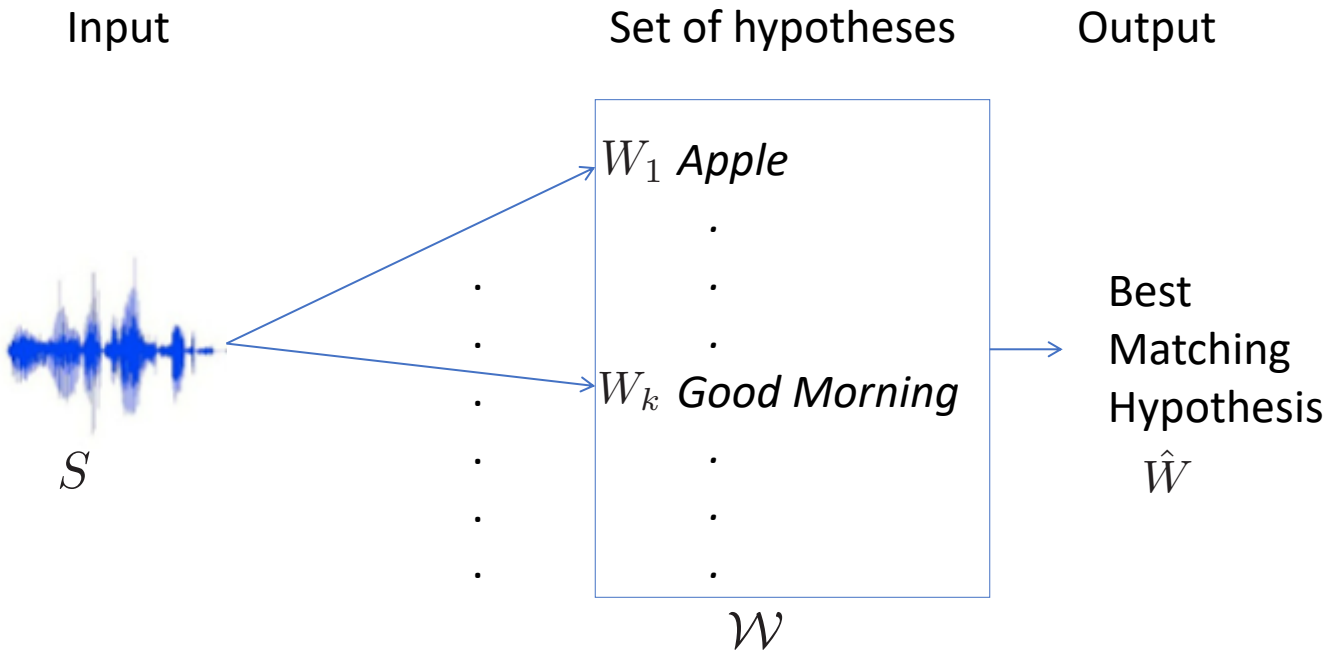
String Matching (Sequence comparison)

Knowledge-based approach

Instance-based ASR approach

Next Part

ASR formulation

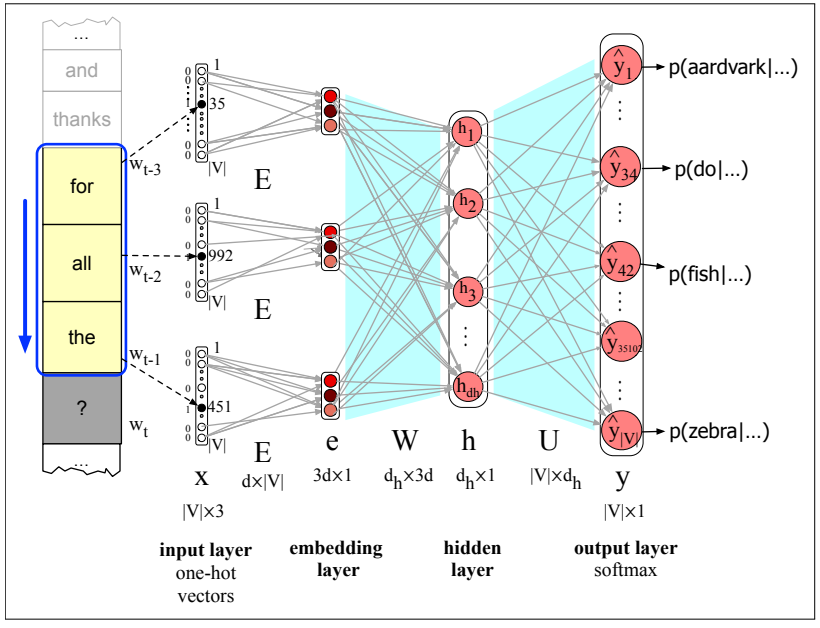


$$\hat{W} = \underset{W_k \in \mathcal{W}}{\arg\max}\, \mathrm{Match}(W_k, S)$$

How to model the set of word hypotheses $\mathcal{W}$?

Language modeling

- Syntactical rule-based method
- Discrete Markov model (DMM) based method
- Neural-based including large language models (LLMs)



Source: Dan Jurafsky and James H. Martin, Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Third Edition Draft.

Three Key Statistical Rules

1. Bayes's rule:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

2. If B_k $(k = 1, \ldots, K)$ are mutually exclusive and collectively exhaustive $(\sum_{k=1}^{K} P(B_k) = 1)$

$$P(A) = \sum_{k=1}^{K} P(A, B_k)$$

3. Gibbs sampler:

$$P(B_1, \ldots, B_k, \ldots, B_K) = \prod_{k=1}^{K} P(B_k|B_{k-1}, \ldots, B_1)$$

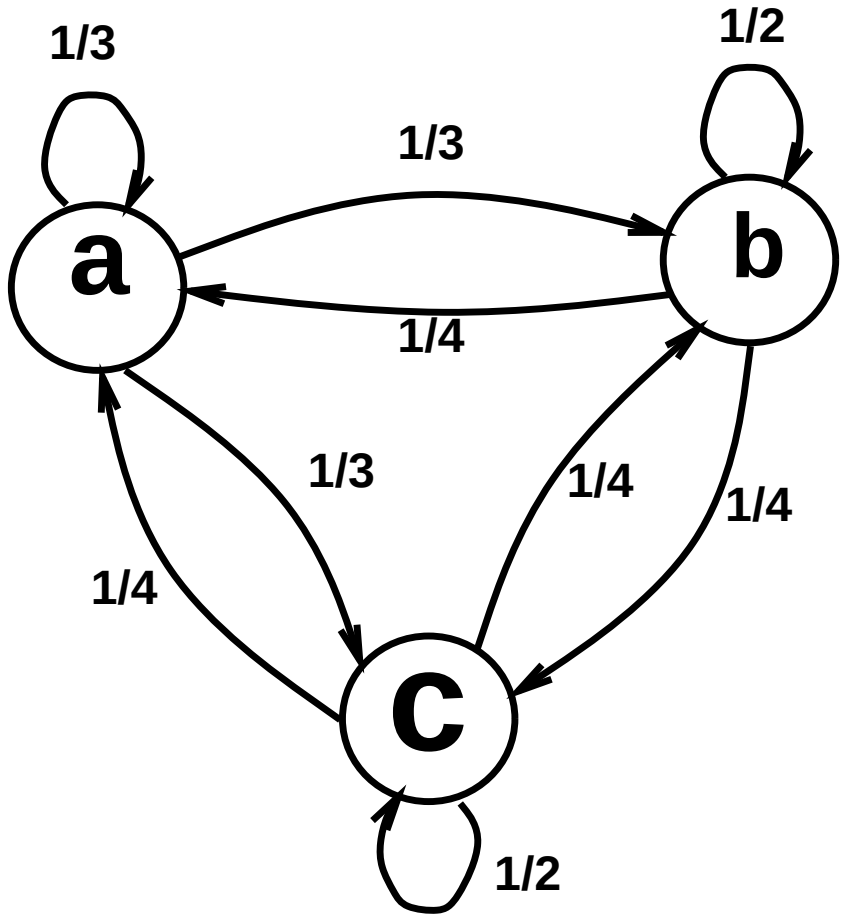
Discrete Markov Model (DMM)

- M built up from states q_ℓ from a set of classes (states) $\Omega = \{\omega_1, \ldots, \omega_k, \ldots, \omega_K\}$
- q^n particular state of M visited at time n,
- $q_\ell^n \equiv \{q^n = q_\ell\}$,
- Parametrized by:

$$\begin{aligned} P(q_\ell^n | q_k^{n-1}, q_j^{n-2}, \ldots) &\simeq P(q_\ell^n | q_k^{n-1}) \quad (\textit{1st order Markov}) \\ &\simeq P(q_\ell | q_k) = P_{k\ell} \quad (\textit{time independent}) \end{aligned}$$

Transition probability matrix: $A = \{P_{k,\ell}\}$.

DMM (2)



Example of fully connected discrete Markov model with $\Omega = \{a, b, c\}$. For example, in case of weather model: "a" = "cloudy", "b"= "rainy" and "c" = "sunny"

Typical Problems (1)

- **Probability of a particular path** $Q = \{q^1, \cdots q^n, \cdots, q^N\}$

$$\begin{aligned} P(Q|M) &= P(q^1|q^0_{initial})P(q^2|q^1)\ldots P(q^n|q^{n-1})\ldots P(q^N|q^{N-1}) \\ &= \prod_{n=1}^{N} P(q^n|q^{n-1}) \end{aligned}$$

$Q = \{sunny, sunny, sunny, rainy, cloudy, sunny, rainy\}$

- **State duration distribution**

  Probability to stay in state q_i for *exactly* d time steps?

$$Q = \{q_i^0, q_i^1, q_i^2, \ldots, q_i^d, q_j^{d+1}\}, \text{ with } j \neq i$$

and:

$$P(Q|M) = (P_{ii})^{d-1}(1 - P_{ii})$$

$Q = \{sunny, sunny, sunny, sunny, sunny, rainy\}$

Typical Problems (2)

- **Probability to go from state q_i to q_j in N steps**

$$P(q_j^N|q_i^0) = \sum_{n=0}^{N} \sum_{\ell=1}^{L} P(q_j^N, q_\ell^n|q_i^0)$$

Defining

$$\alpha(\ell, n) = P(q_\ell^n|q_i^0, N)$$

We have:

$$\alpha(\ell, n+1) = \sum_k \alpha(k, n) P_{k\ell}$$

$$P(q_j^N|q_i^0) = \alpha(j, N)$$

Suppose N=4, $q^1 = \textit{rainy}$ and $q^4 = \textit{cloudy}$

rainy, rainy, rainy, cloudy

rainy, rainy, cloudy, cloudy

rainy, rainy, sunny, cloudy

.

.

Typical Problems (3)

- **Probability of best path of length N between q_i and q_j**
  If $\overline{P}(k, n)$ is probability of best path to go from q_i to q_k in n steps:

$$\overline{P}(\ell, n+1) = \max_k \overline{P}(k, n) P_{k\ell}$$

  and

$$\overline{P}(q_j^N | q_i^0) = P(j, N)$$

  Generalization:

$$A^n(i, j) = P(q_j^n | q_i^0)$$

Estimation of $P(W_k)$

Let $W_k = \{w_{k,1}, \ldots w_{k,j}, \ldots w_{k,J}\}$ (sequence of words)

$$\begin{aligned} P(W_k) &= P(w_{1,k}, \ldots w_{j,k}, \ldots w_{J,k}) \\ &= P(w_{J,k}|w_{J-1,k}, \ldots, w_{1,k}) \cdot P(w_{J-1,k}, \ldots w_{1,k}) \\ &= P(w_{J,k}|w_{J-1,k}, \ldots, w_{1,k}) \cdot P(w_{J-1,k}|w_{J-2,k}, \ldots, w_{1,k}) \cdots \\ &\quad P(w_{3,k}|w_{2,k}, w_{1,k}) \cdot P(w_{2,k}|w_{1,k}) \cdot P(w_{1,k}) \end{aligned}$$

Usually $P(w_{1,k}) = P(w_{1,k}|\text{initial state})$
Example: $W_k = \{\text{my}, \text{name}, \text{is}, \text{bond}\}$

$$\begin{aligned} P(\text{my}, \text{name}, \text{is}, \text{bond}) &= P(\text{bond}|\text{is}, \text{name}, \text{my}) \cdot P(\text{is}, \text{name}, \text{my}) \\ &= P(\text{bond}|\text{is}, \text{name}, \text{my}) \cdot P(\text{is}|\text{name}, \text{my}) \cdot \\ &\quad P(\text{name}|\text{my}) \cdot P(\text{my}) \end{aligned}$$

Estimation of $P(W_k)$: n-gram

- Challenge: variable history length
- Solution: Markov assumption
- n-gram is a $(n-1)$ order Markov model
  - bigram language model: first order Markov model

$$P(\text{my}, \text{name}, \text{is}, \text{bond}) = P(\text{bond}|\text{is}) \cdot P(\text{is}|\text{name}) \cdot P(\text{name}|\text{my}) \cdot P(\text{my}|\text{initial state})$$

  - trigram language model: second order Markov model

$$P(\text{my}, \text{name}, \text{is}, \text{bond}) = P(\text{bond}|\text{is}, \text{name}) \cdot P(\text{is}|\text{name}, \text{my}) \cdot P(\text{name}|\text{my}) \cdot P(\text{my}|\text{initial state})$$

Suppose the vocabulary size is O then the number of parameters or transition probabilities to estimate are:

- Bigram language model: $O \times O + O$
- Trigram language model: $O \times O \times O + O \times O + O$

n-gram parameter estimation

- Requires large amount of text e.g. books, web
- Parameter estimation through counting, e.g. trigram

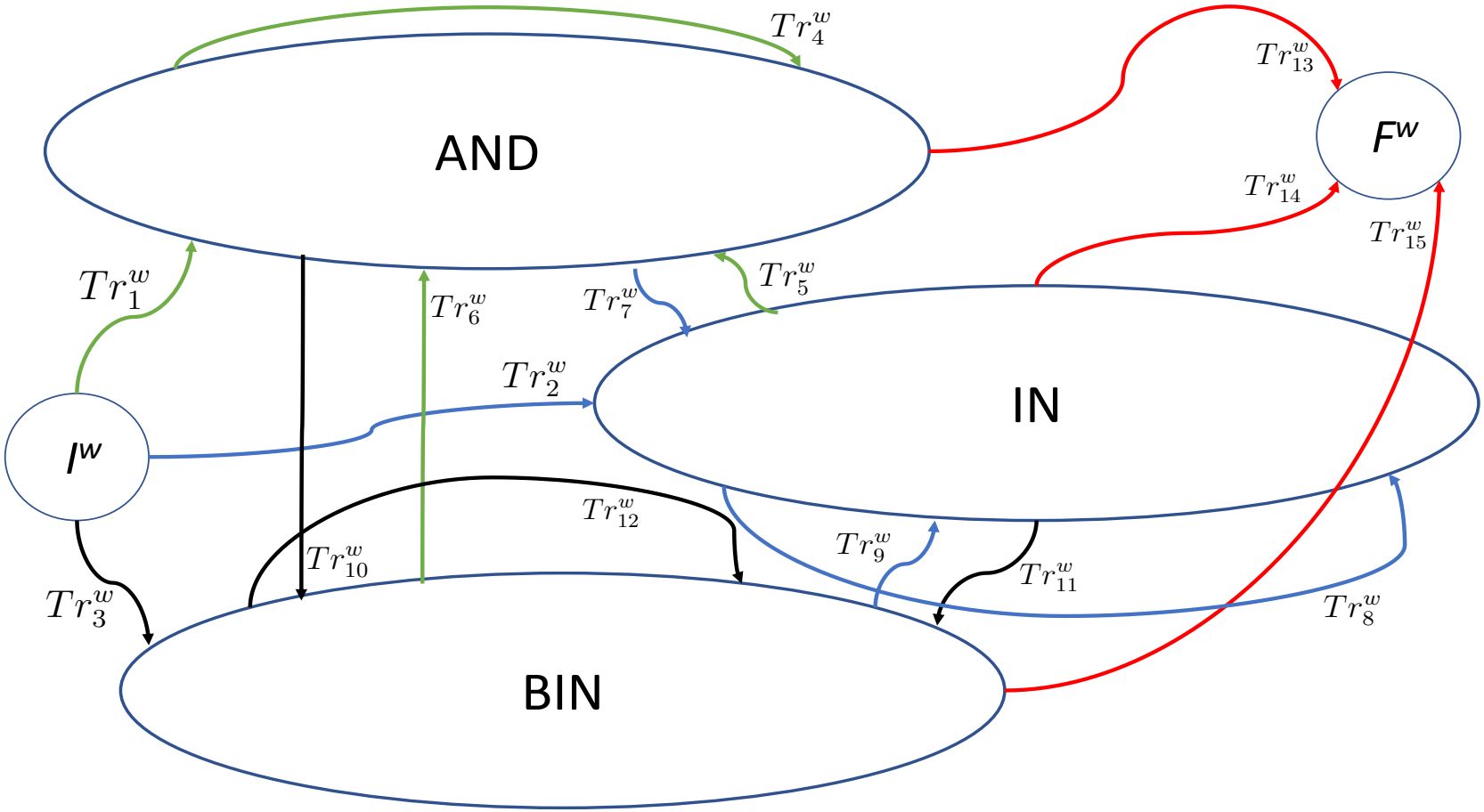
$$P(w_n|w_{n-1}, w_{n-2}) = \frac{\text{Count}(w_n, w_{n-1}, w_{n-2})}{\sum_w \text{Count}(w, w_{n-1}, w_{n-2})}$$

Example:

$$P(\text{bond}|\text{is}, \text{name}) = \frac{\text{Count}(\text{bond}, \text{is}, \text{name})}{\sum_w \text{Count}(w, \text{is}, \text{name})}$$

- Small probability estimation problems, e.g. not enough examples, unseen word contexts
  - Interpolation with lower n-gram probabilities
  - Discounting: assign a small probability mass
  - Back-off to lower order n-gram probabilities
    $P(w_n|w_{n-1}, w_{n-2}) \Rightarrow P(w_n|w_{n-1})$
- Written text versus spoken language (conversational speech)

DMM bigram illustration



$$Tr_1^w = P(\text{AND}|I^w)\ Tr_2^w = P(\text{IN}|I^w)\ Tr_3^w = P(\text{BID}|I^w)$$
$$Tr_4^w = P(\text{AND}|\text{AND})\ Tr_5^w = P(\text{AND}|\text{IN})\ Tr_6^w = P(\text{AND}|\text{BID})$$
$$Tr_7^w = P(\text{IN}|\text{AND})\ Tr_8^w = P(\text{IN}|\text{IN})\ Tr_9^w = P(\text{IN}|\text{BID})$$
$$Tr_{10}^w = P(\text{BID}|\text{AND})\ Tr_{11}^w = P(\text{BID}|\text{IN})\ Tr_{12}^w = P(\text{BID}|\text{BID})$$
$$Tr_{13}^w = P(F^w|\text{AND})\ Tr_{14}^w = P(F^w|\text{IN})\ Tr_{15}^w = P(F^w|\text{BID})$$

Outline

Automatic speech recognition problem formulation

Language generation (Sequence generation)

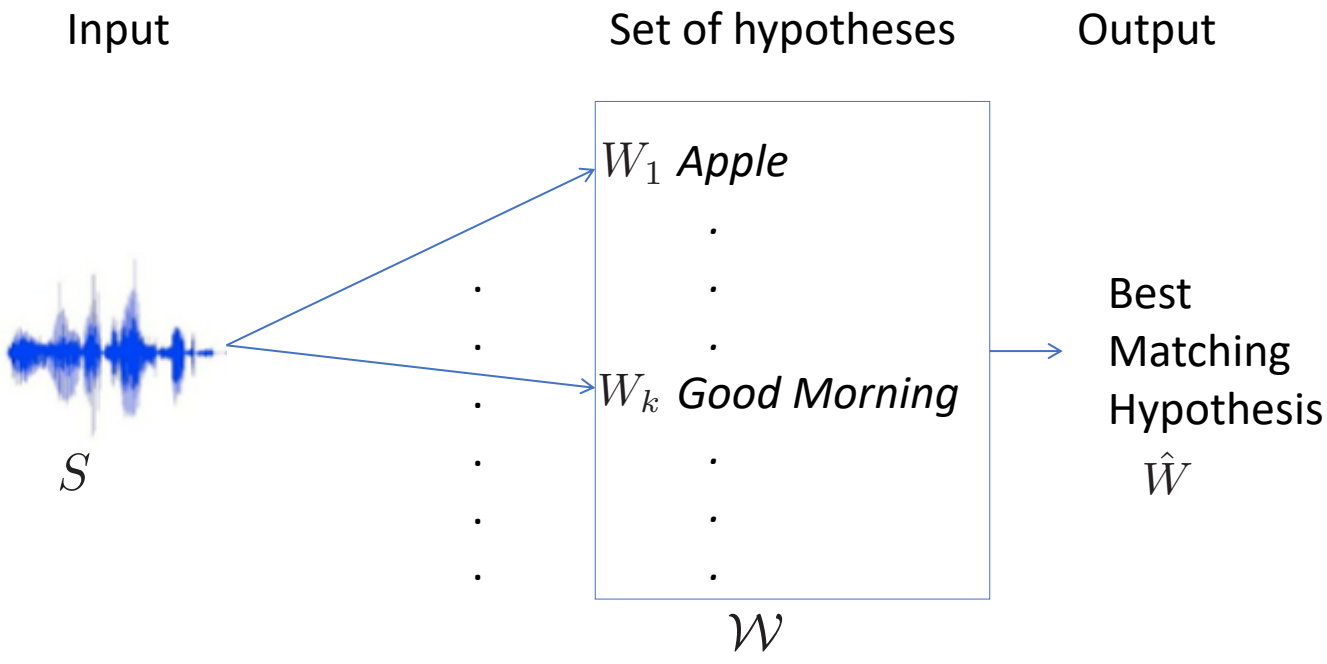
String Matching (Sequence comparison)

Knowledge-based approach

Instance-based ASR approach

Next Part

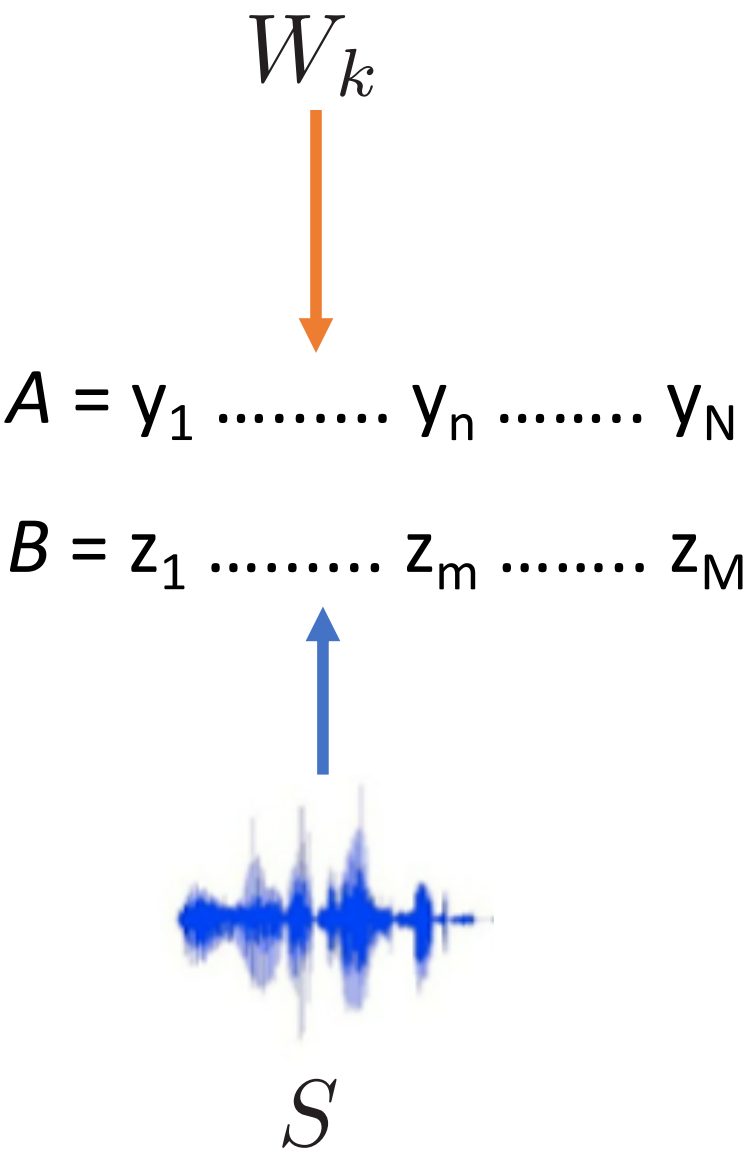
Automatic speech recognition (ASR)



$$\hat{W} = \underset{W_k \in \mathcal{W}}{\arg\max}\, \mathrm{Match}(W_k, S)$$

How to match an observed speech signal S with a word hypothesis W_k?

Abstract formulation for matching S and W_k



Core Idea

1. Map S and W_k to a shared latent symbol space
2. Match the resulting two latent symbol sequences A and B

Four sub questions

Q1: What is the shared latent symbol set?

Q2: How to map S to a latent symbol sequence B?

Q3: How to map W_k to a latent symbol sequence A?

Q4: How to match the two latent symbol sequences A and B?

Different ASR methods mainly differ on how these four sub questions are addressed.

Sequence comparison

ASR system evaluation

ASR System

THE FROG BARKED AT AT THE RACING CAR

Human Transcriber

THE BIG DOG BARKED AT THE PASSING CAR

Compare the two strings to compute errors

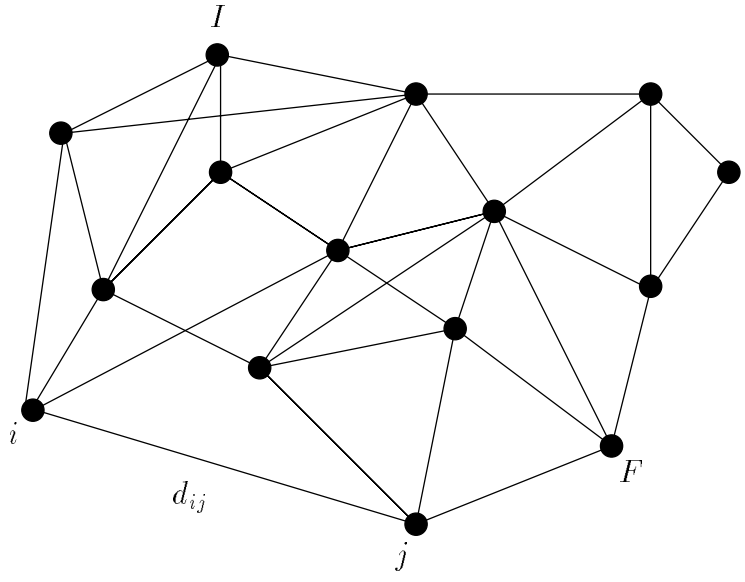
Dynamic Programming (DP)

Bellman, 1960

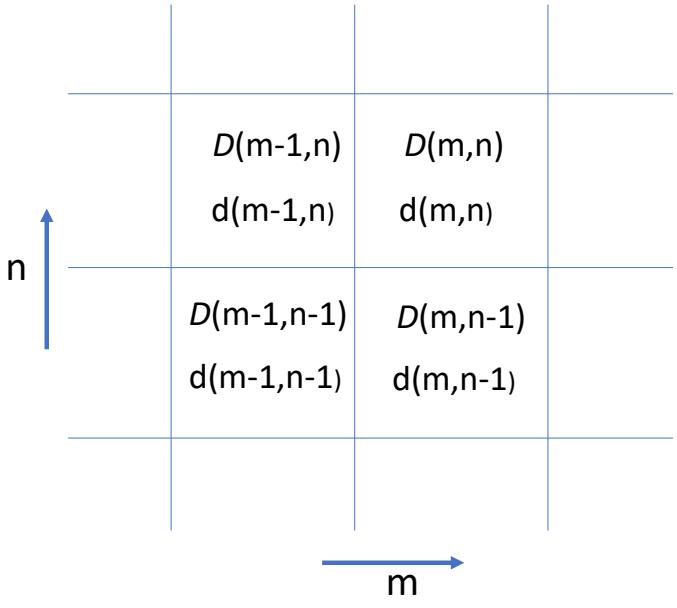
"*Optimal policy is composed of optimal sub-policies*".

Other Applications:

- Cargo loading problem, VLSI design, etc...
- Finding the shortest path between two points in a graph



String matching using DP



local score $d(m, n)$:
if $\mathrm{str}(m) = \mathrm{str}(n)$
$d(m, n) = 0$
else
$d(m, n) = 1$

1. Initial condition: path starts at $(1, 1)$
2. Recursion:

$$D(m, n) = d(m, n) + min[D(m-1, n), D(m-1, n-1), D(m, n-1)]$$

$$\mathrm{Path}(m, n) = \arg\min[D(m-1, n), D(m-1, n-1), D(m, n-1)]$$

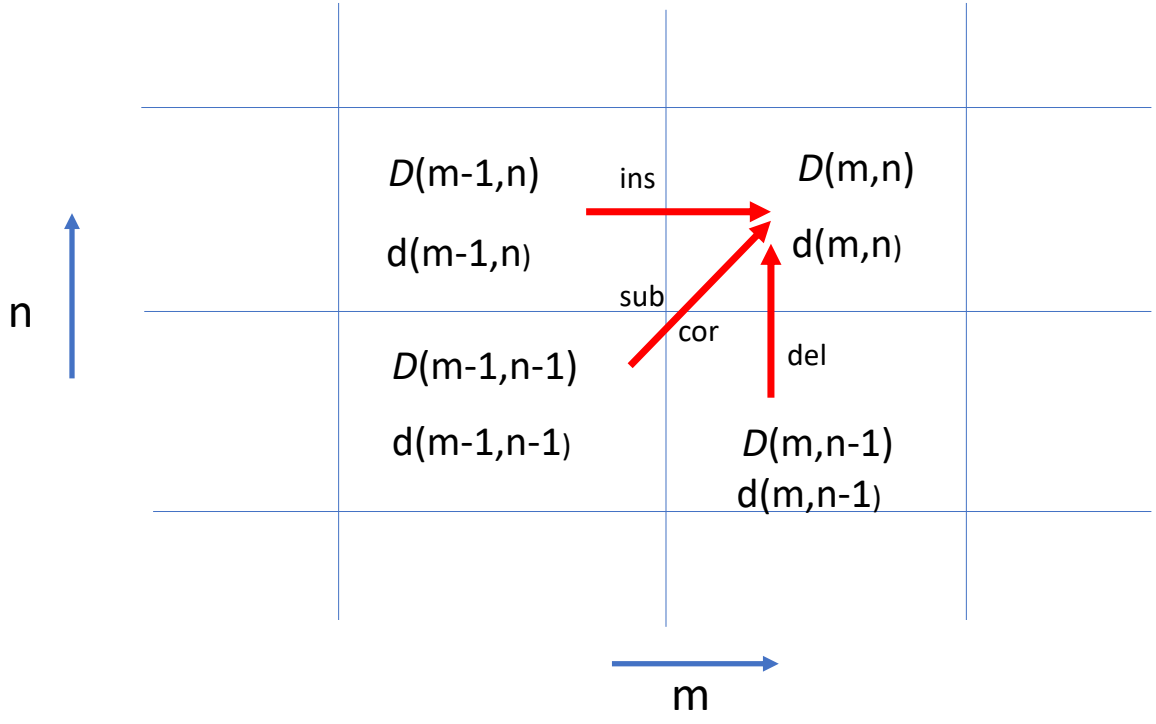
$\forall m \in \{1 \ldots M\}$ and $n \in \{1, \ldots N\}$

3. Final condition: path ends at (M, N) and $D(M, N)$ is the global score

$\mathrm{Path}(m, n)$ denotes the path index. Path can be traced back from $\mathrm{Path}(M, N)$

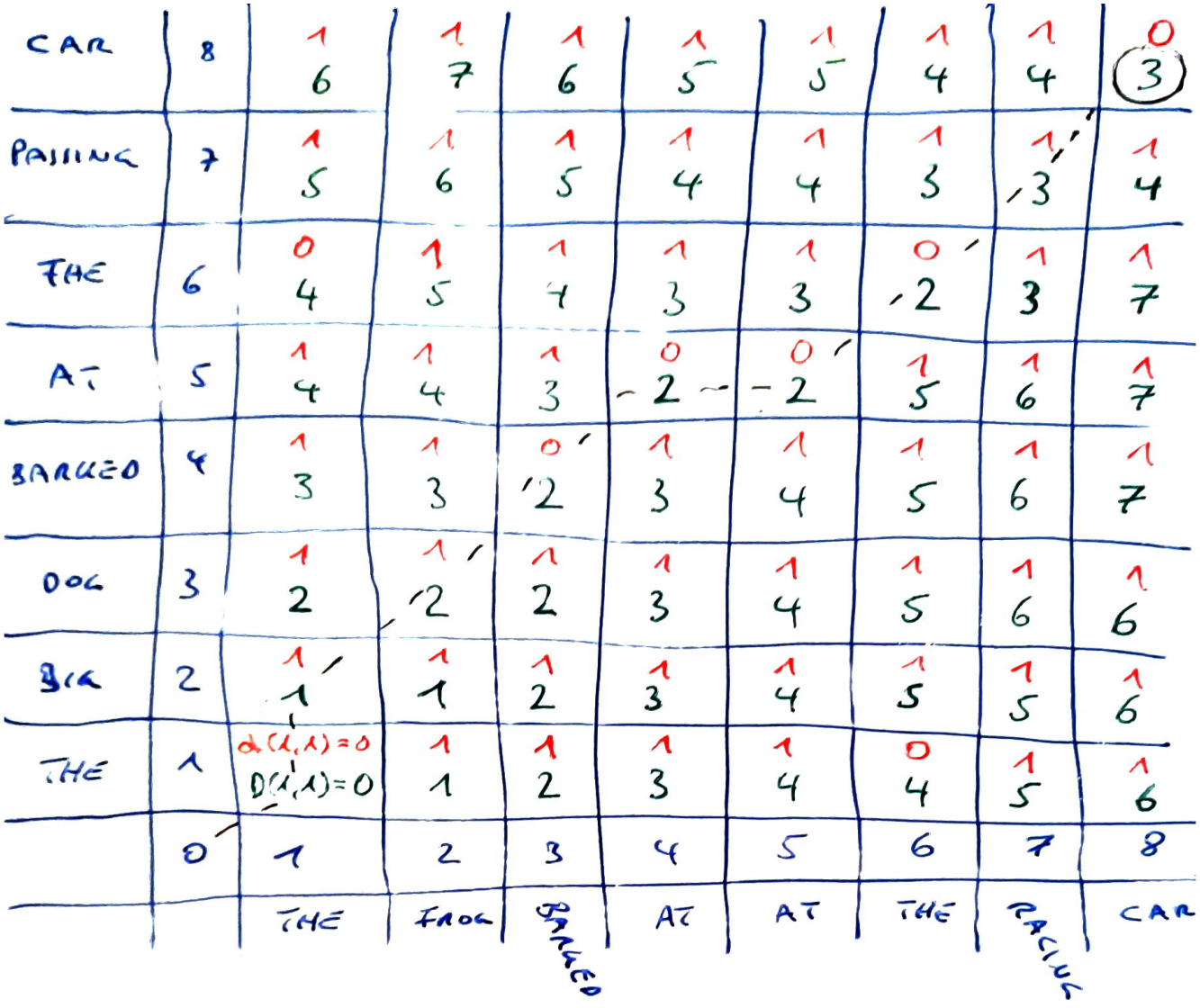
Local constraints

$$D(m,n) = d(m,n) + min[D(m-1,n), D(m-1,n-1), D(m,n-1)]$$



ins: insertion, cor: correct, sub: substitution, del: deletion

Word error rate (WER) calculation



$$\text{WER} = \frac{\text{ins}+\text{del}+\text{sub}}{N_{ref}}$$

N_{ref}: number of words in the reference transcript

$$\text{WER} = \frac{1+1+2}{8} = 0.5$$

Outline

Automatic speech recognition problem formulation

Language generation (Sequence generation)

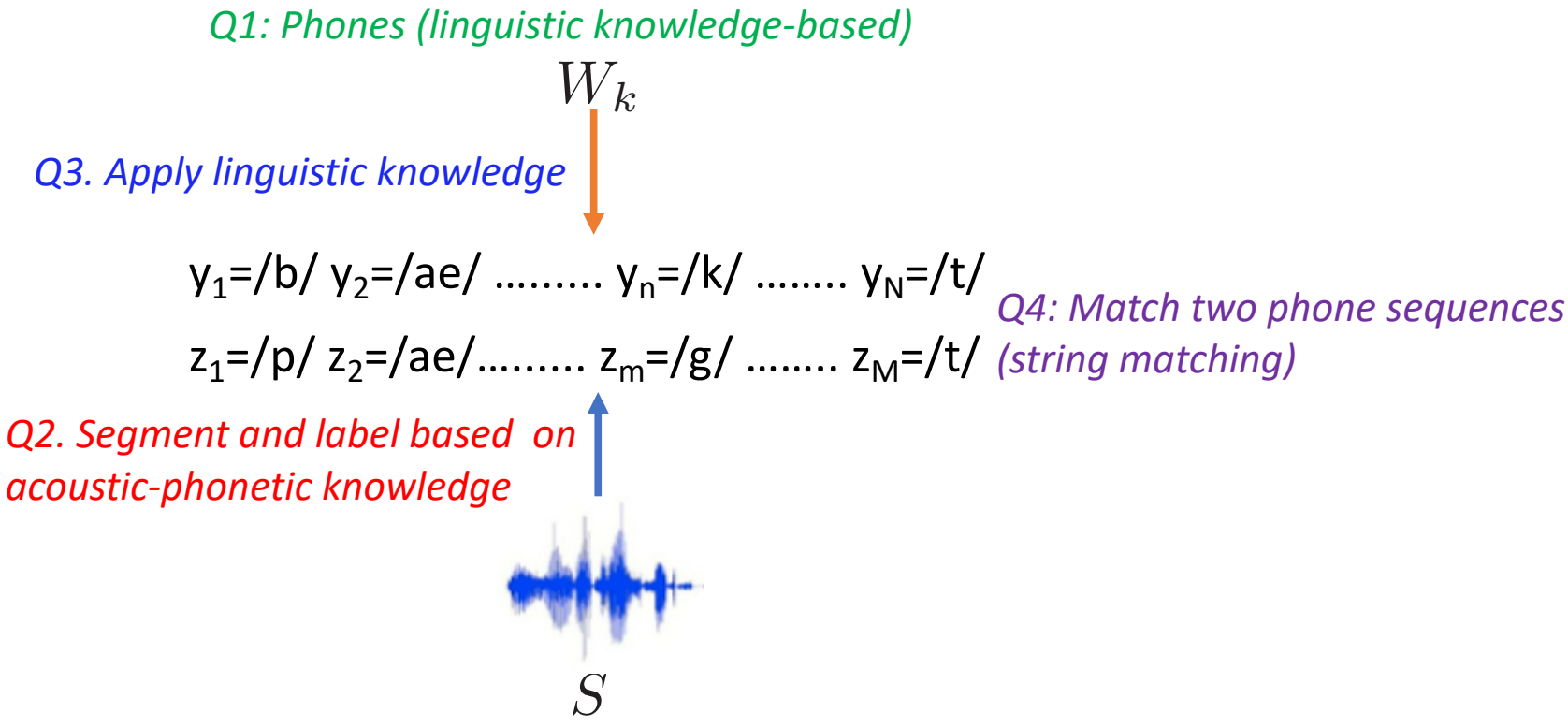
String Matching (Sequence comparison)

Knowledge-based approach

Instance-based ASR approach

Next Part

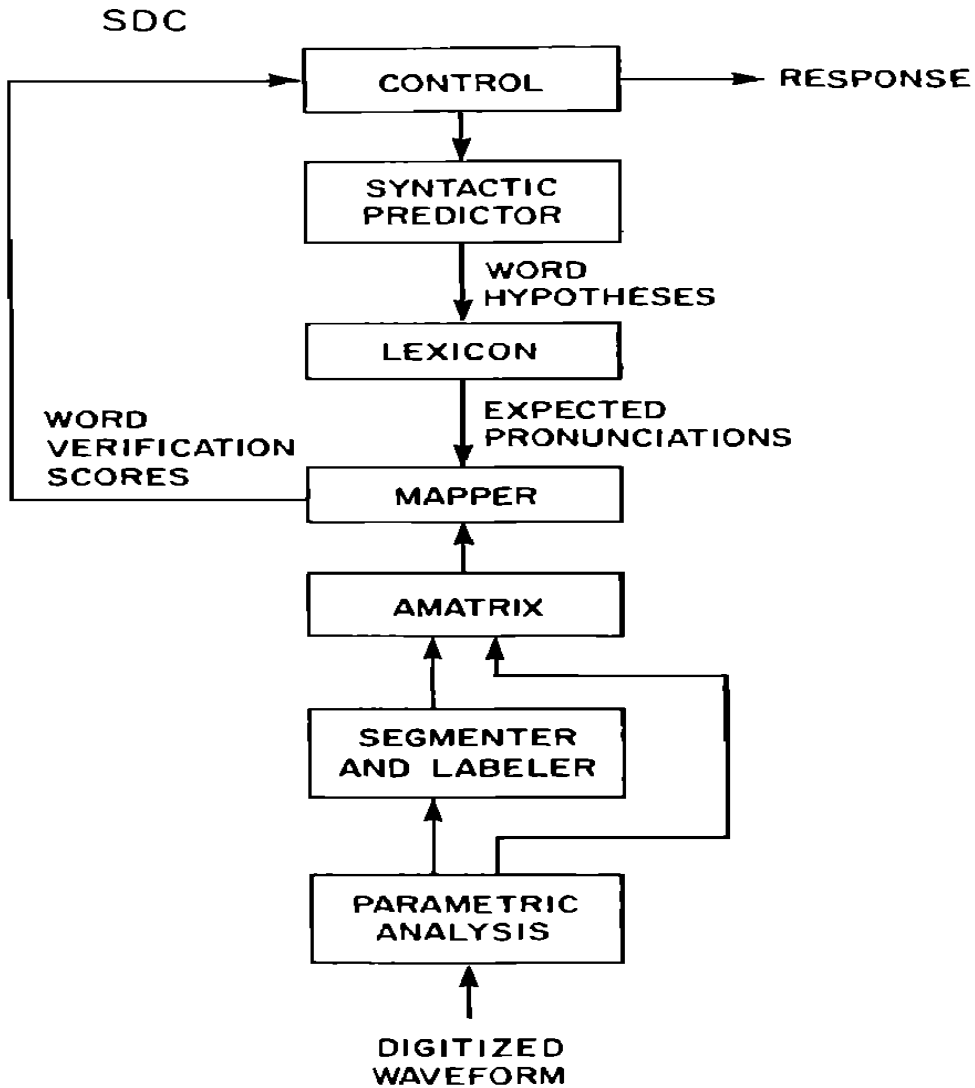
Knowledge-based ASR approach



Limitations:

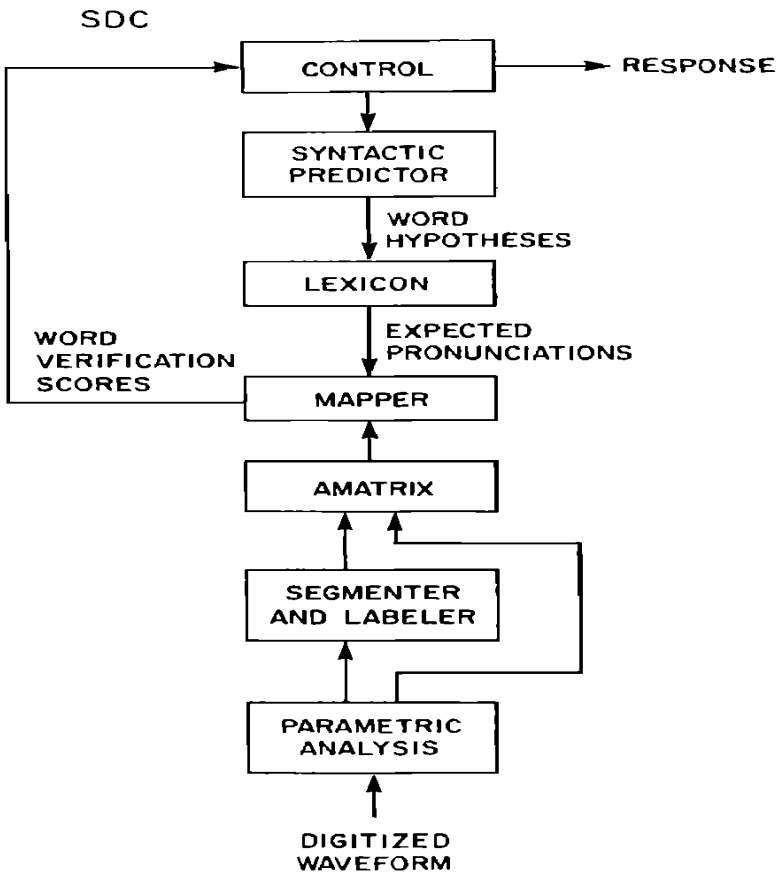
- Overly relies on knowledge
- Makes early decision so difficult to recover from errors such as, segmentation and labeling errors

Knowledge-based ASR system (1)



Source: D. H. Klatt. Review of the ARPA speech understanding project. J. Acoust. Soc. Amer., 62(6):1345-1366, December 1977.

Contrast to current approaches



Source: D. H. Klatt. Review of the ARPA speech understanding project. J. Acoust. Soc. Amer., 62(6):1345-1366, December 1977.

Connectionist Temporal Classification

We start with an input sequence, like a spectrogram of audio.

The input is fed into an RNN, for example.

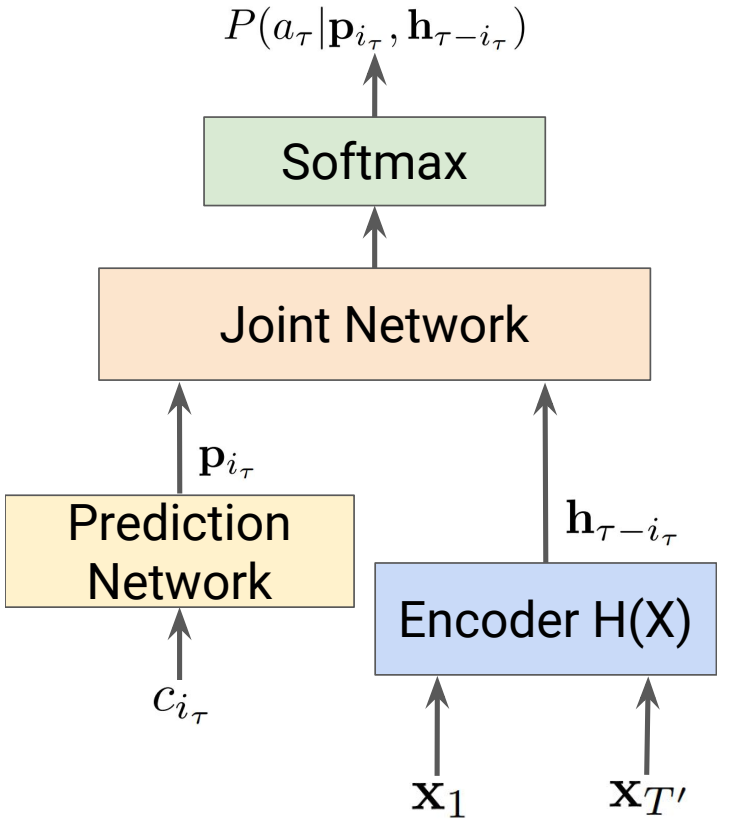
The network gives $p_t(a \mid X)$, a distribution over the outputs $\{h, e, l, o, \epsilon\}$ for each input step.

With the per time-step output distribution, we compute the probability of different sequences

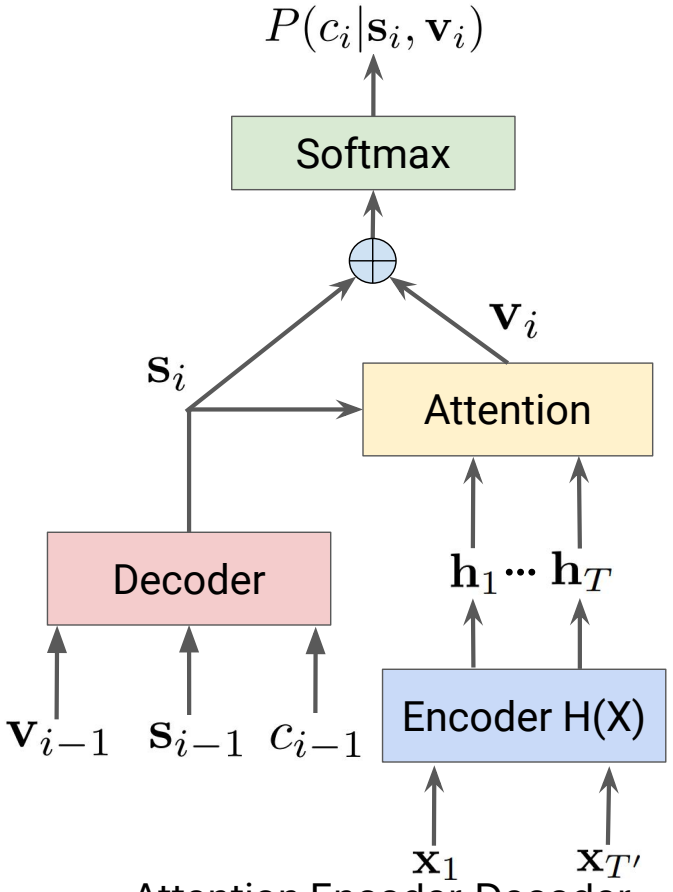
By marginalizing over alignments, we get a distribution over outputs.

Source: Hannun, "Sequence Modeling with CTC", Distill, 2017.

Contrast to current approaches (2)



Recurrent Neural Network-Transducer



Attention Encoder-Decoder

Source: R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter and S. Watanabe, End-to-End Speech Recognition: A Survey, arXiv:2303.03329, 2023

Outline

Automatic speech recognition problem formulation

Language generation (Sequence generation)

String Matching (Sequence comparison)

Knowledge-based approach

Instance-based ASR approach

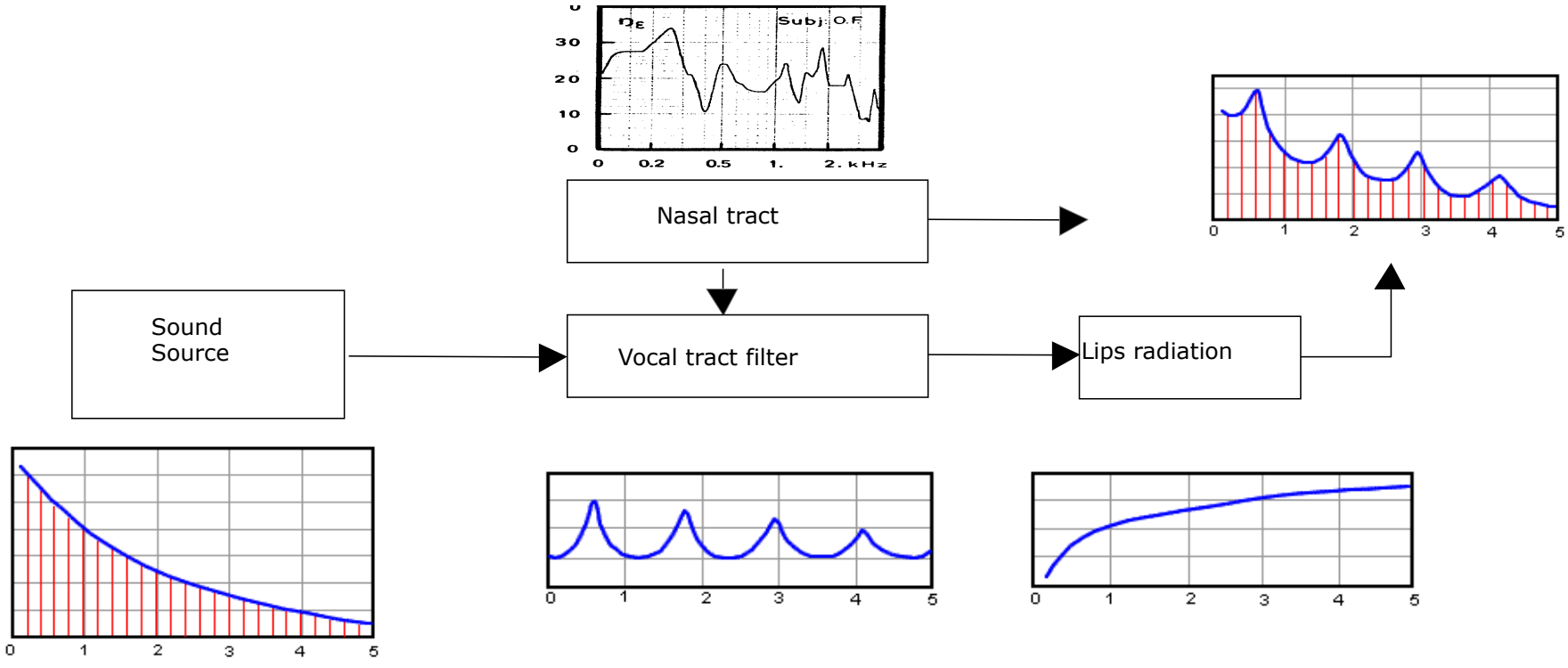
Next Part

Instance-based ASR approach

In instance-based (also called template-based) approach W_k is represented by a speech signal

For example, record each word

Motivation (1)

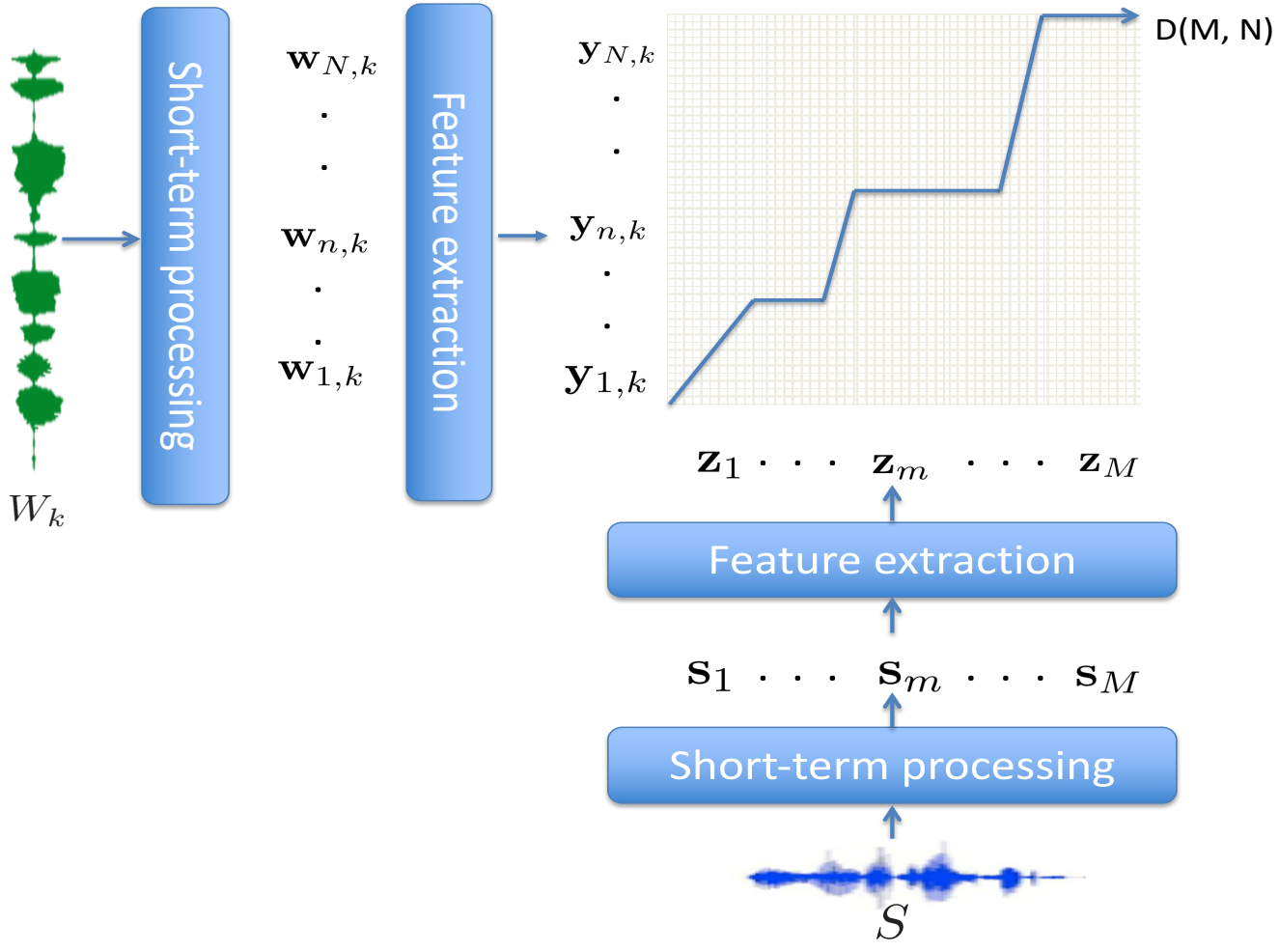


Credits: Lindqvist-Gauffin, Sundberg, Stevens, Mannel

Motivation (2)

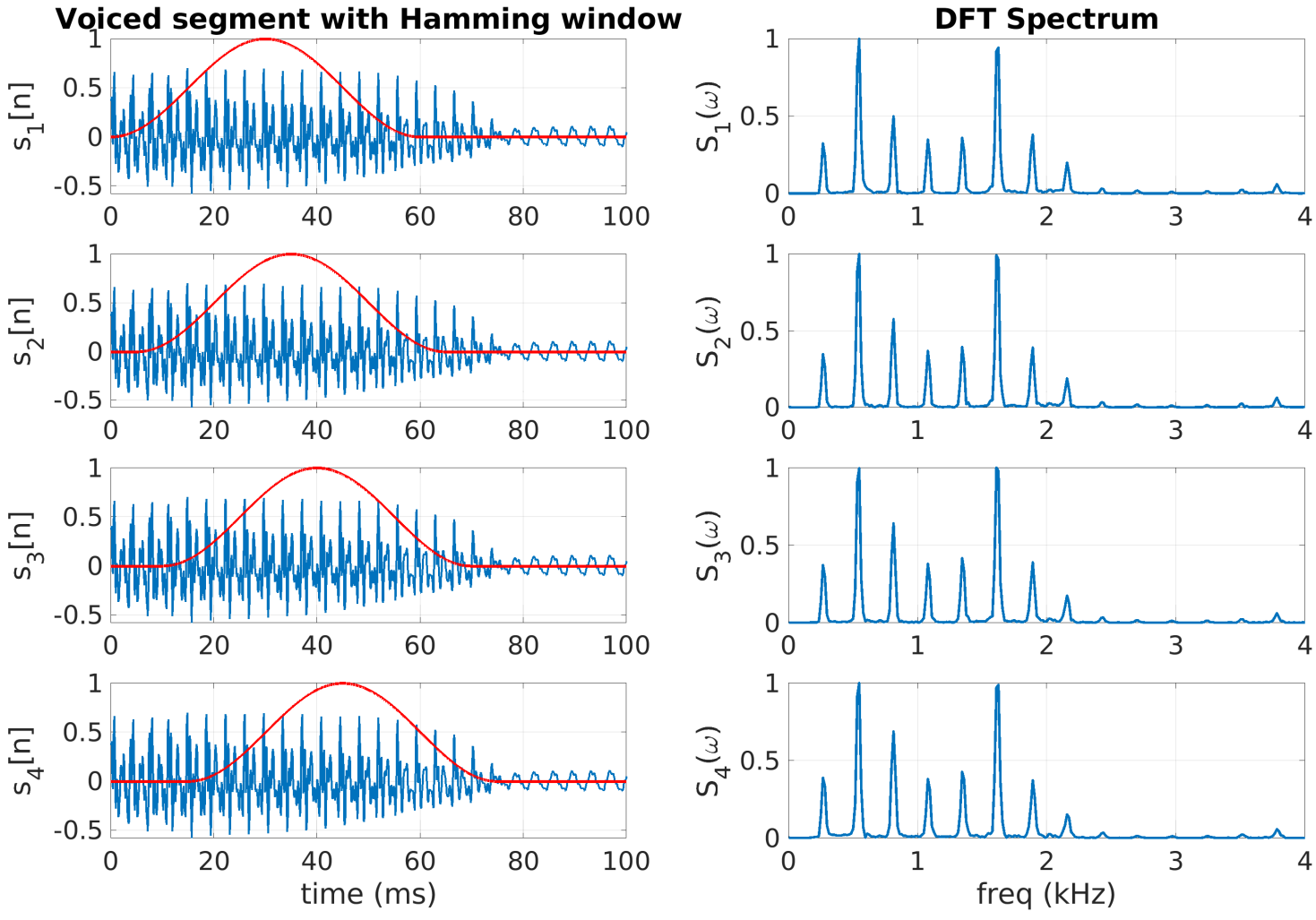
- Speech signal can be deconvolved into source and system components and synthesized back by putting these components. (e.g., linear prediction, cepstral analysis)
- Vocal tract shape is different for different sounds (caution: there are pair of sounds that differ mainly in terms of voicing, e.g., /p/ and /b/)
- Parametrize the vocal tract system information integrating speech perception knowledge (e.g. MFCCs, PLP cepstral coefficients) and compare S and W_k .

Matching S and W_k

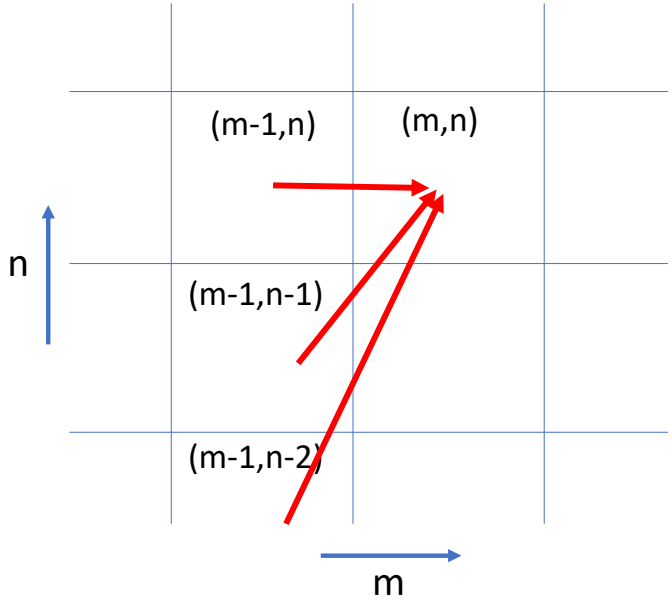


- s_m and $w_{n,k}$ denote of frame of speech signal
- z_m and $y_{n,k}$ denote the corresponding feature vectors

Short-term spectral processing



Dynamic Time Warping (DTW)



local score $d(m, n)$:

- Cepstral features: Euclidean distance between z_m and $y_{n,k}$
- Linear prediction coefficients: Itakura distance between z_m and $y_{n,k}$
- Spectral information: Itakura-Saito distance between z_m and $y_{n,k}$

1. Initial condition: path starts at $(1, 1)$
2. Recursion:

$$\begin{aligned} D(m,n) =& d(m,n) + min[D(m-1,n), \\ & D(m-1,n-1), \\ & D(m-1,n-2)] \end{aligned}$$

$$\begin{aligned} \mathrm{Path}(m,n) =& \arg\min[D(m-1,n), \\ & D(m-1,n-1), \\ & D(m,n-2)] \end{aligned}$$

$\forall m \in \{1 \ldots M\}$ and $n \in \{1, \ldots N\}$

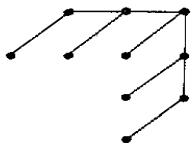
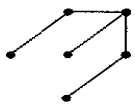
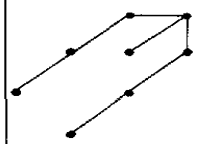
3. Final condition: path ends at (M, N) and $D(M, N)$ is the global score

$\mathrm{Path}(m, n)$ denotes the path index. Path can be traced back from $\mathrm{Path}(M, N)$

DTW local constraints

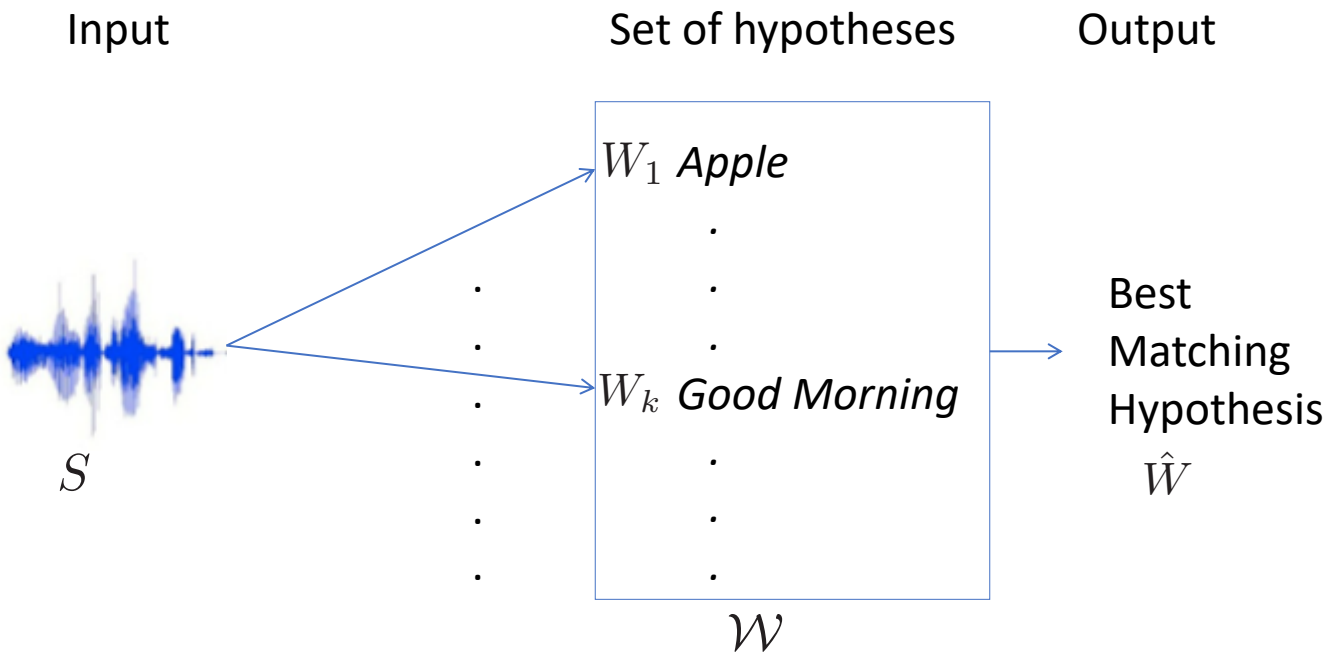
TABLE I

SYMMETRIC AND ASYMMETRIC DP-ALGORITHMS WITH SLOPE CONSTRAINT CONDITION $P = 0, \frac{1}{2}, 1,$ AND 2

P	Schematic explanation	Symmetric / Asymmetric	DP-equation $g(i, j) =$
0		Symmetric	$\min \begin{bmatrix} g(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-1,j)+d(i,j) \end{bmatrix}$
		Asymmetric	$\min \begin{bmatrix} g(i,j-1) \\ g(i-1,j-1)+d(i,j) \\ g(i-1,j)+d(i,j) \end{bmatrix}$
1/2		Symmetric	$\min \begin{bmatrix} g(i-1,j-3)+2d(i,j-2)+d(i,j-1)+d(i,j) \\ g(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-2,j-1)+2d(i-1,j)+d(i,j) \\ g(i-3,j-1)+2d(i-2,j)+d(i-1,j)+d(i,j) \end{bmatrix}$
		Asymmetric	$\min \begin{bmatrix} g(i-1,j-3)+(d(i,j-2)+d(i,j-1)+d(i,j))/3 \\ g(i-1,j-2)+(d(i,j-1)+d(i,j))/2 \\ g(i-1,j-1)+d(i,j) \\ g(i-2,j-1)+d(i-1,j)+d(i,j) \\ g(i-3,j-1)+d(i-2,j)+d(i-1,j)+d(i,j) \end{bmatrix}$
1		Symmetric	$\min \begin{bmatrix} g(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-2,j-1)+2d(i-1,j)+d(i,j) \end{bmatrix}$
		Asymmetric	$\min \begin{bmatrix} g(i-1,j-2)+(d(i,j-1)+d(i,j))/2 \\ g(i-1,j-1)+d(i,j) \\ g(i-2,j-1)+d(i-1,j)+d(i,j) \end{bmatrix}$
2		Symmetric	$\min \begin{bmatrix} g(i-2,j-3)+2d(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-3,j-2)+2d(i-2,j-1)+2d(i-1,j)+d(i,j) \end{bmatrix}$
		Asymmetric	$\min \begin{bmatrix} g(i-2,j-3)+2(d(i-1,j-2)+d(i,j-1)+d(i,j))/3 \\ g(i-1,j-1)+d(i,j) \\ g(i-3,j-2)+d(i-2,j-1)+d(i-1,j)+d(i,j) \end{bmatrix}$

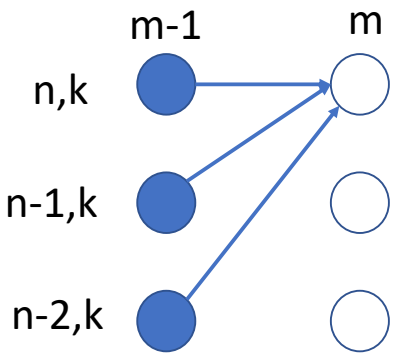
Source: H. Sakoe and S. Chiba,"Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on Acoustics, Speech, and Signal Processing, 26(1), 1978.

Instance-based ASR

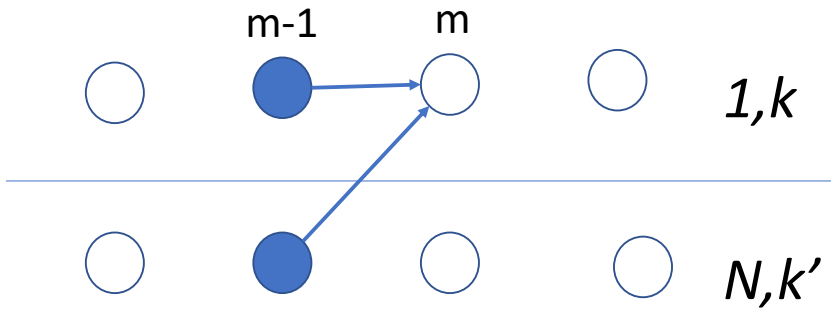


$$\hat{W} = \underset{W_k \in \mathcal{W}}{\arg\min}\, \mathrm{DTW}(W_k, S)$$

Across word local constraint



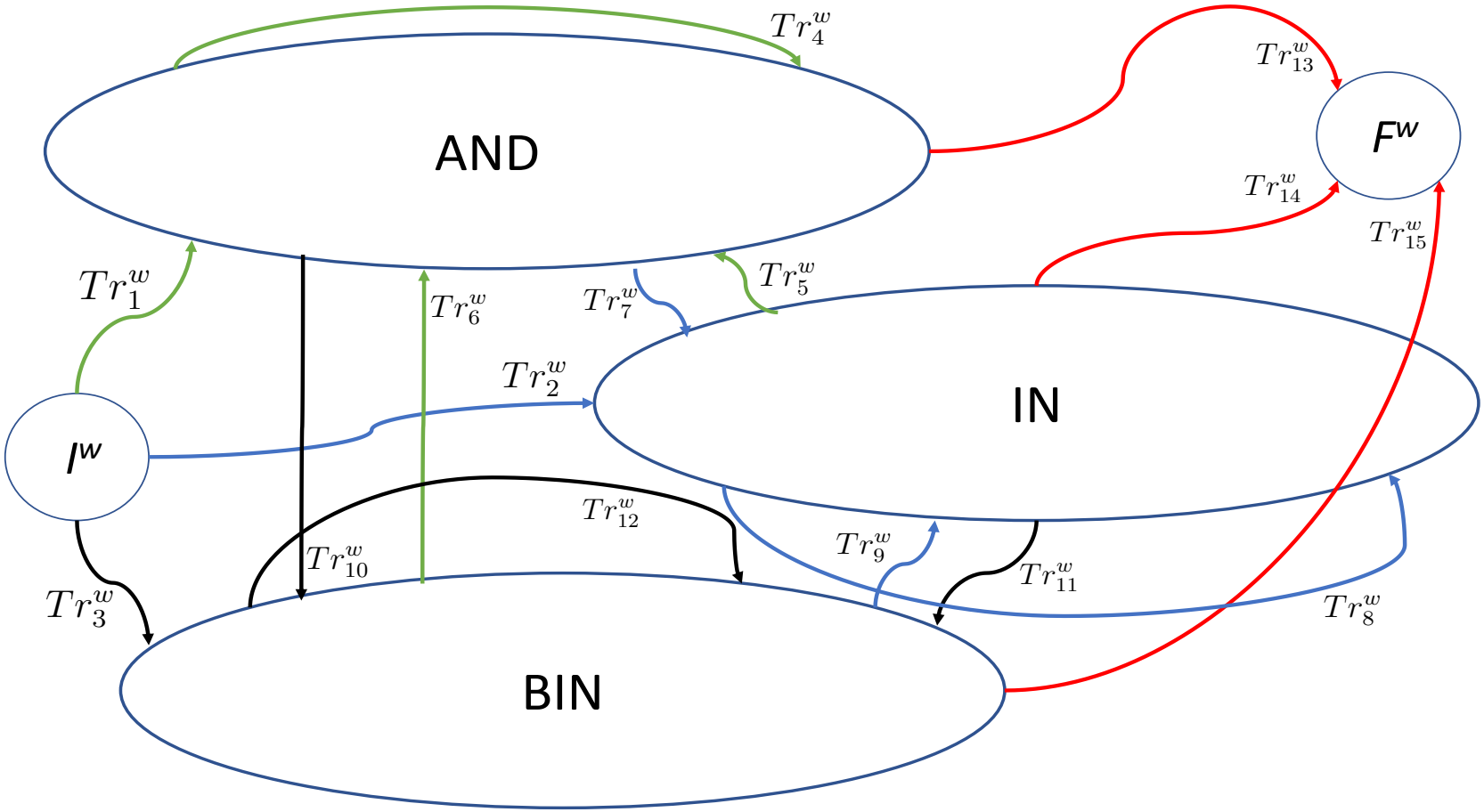
Within word constraint



Across word constraint

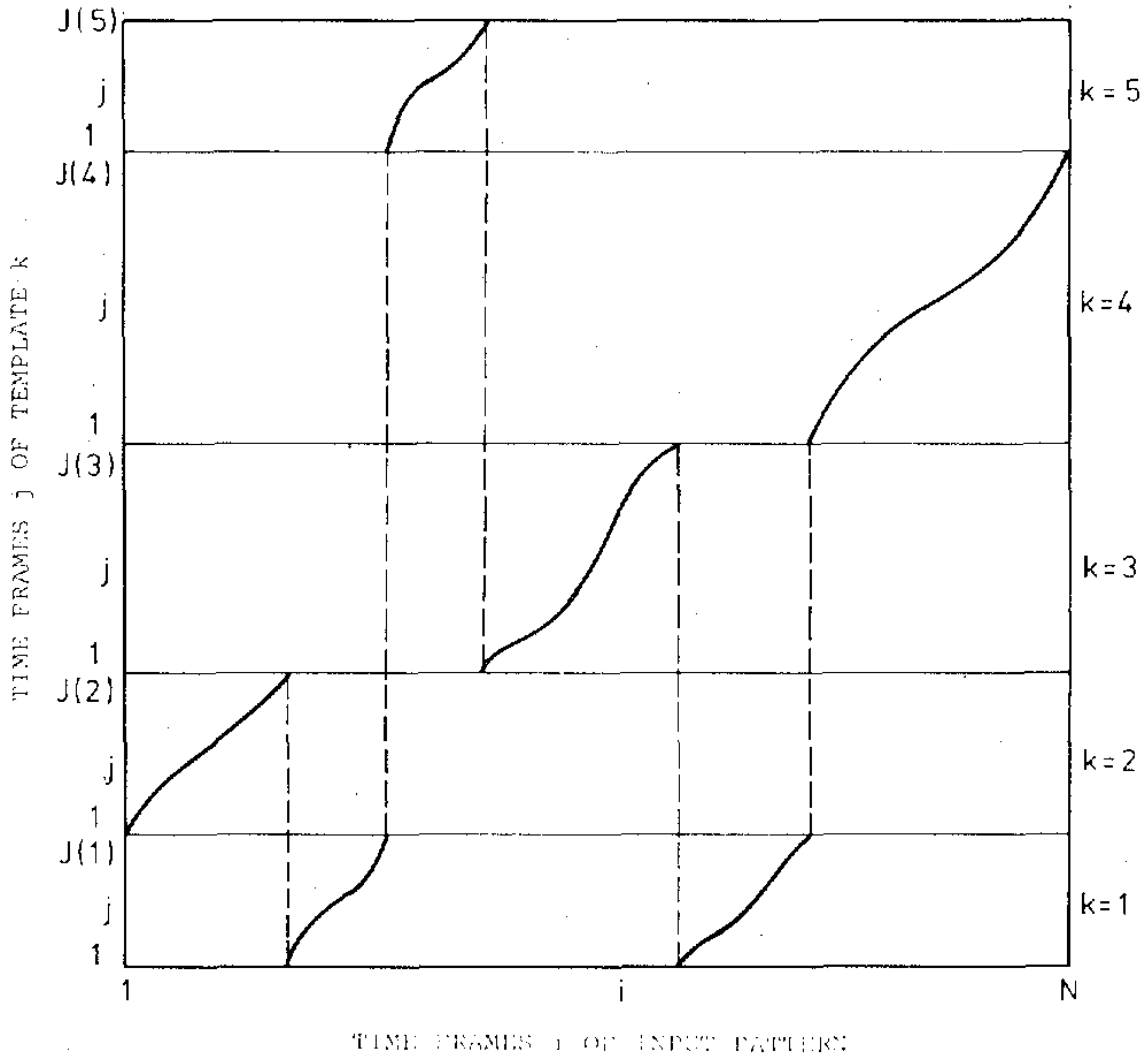
$$\forall k' \in \{1, \cdots K\}$$

DMM bigram illustration



$$Tr_1^w = P(\text{AND}|I^w)\ Tr_2^w = P(\text{IN}|I^w)\ Tr_3^w = P(\text{BID}|I^w)$$
$$Tr_4^w = P(\text{AND}|\text{AND})\ Tr_5^w = P(\text{AND}|\text{IN})\ Tr_6^w = P(\text{AND}|\text{BID})$$
$$Tr_7^w = P(\text{IN}|\text{AND})\ Tr_8^w = P(\text{IN}|\text{IN})\ Tr_9^w = P(\text{IN}|\text{BID})$$
$$Tr_{10}^w = P(\text{BID}|\text{AND})\ Tr_{11}^w = P(\text{BID}|\text{IN})\ Tr_{12}^w = P(\text{BID}|\text{BID})$$
$$Tr_{13}^w = P(F^w|\text{AND})\ Tr_{14}^w = P(F^w|\text{IN})\ Tr_{15}^w = P(F^w|\text{BID})$$

Continuous speech recognition



Source: H. Ney,"The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition ", IEEE Trans. on Acoustics, Speech, and Signal Processing, 32(2), 1984.

Four sub questions for instance-based approach

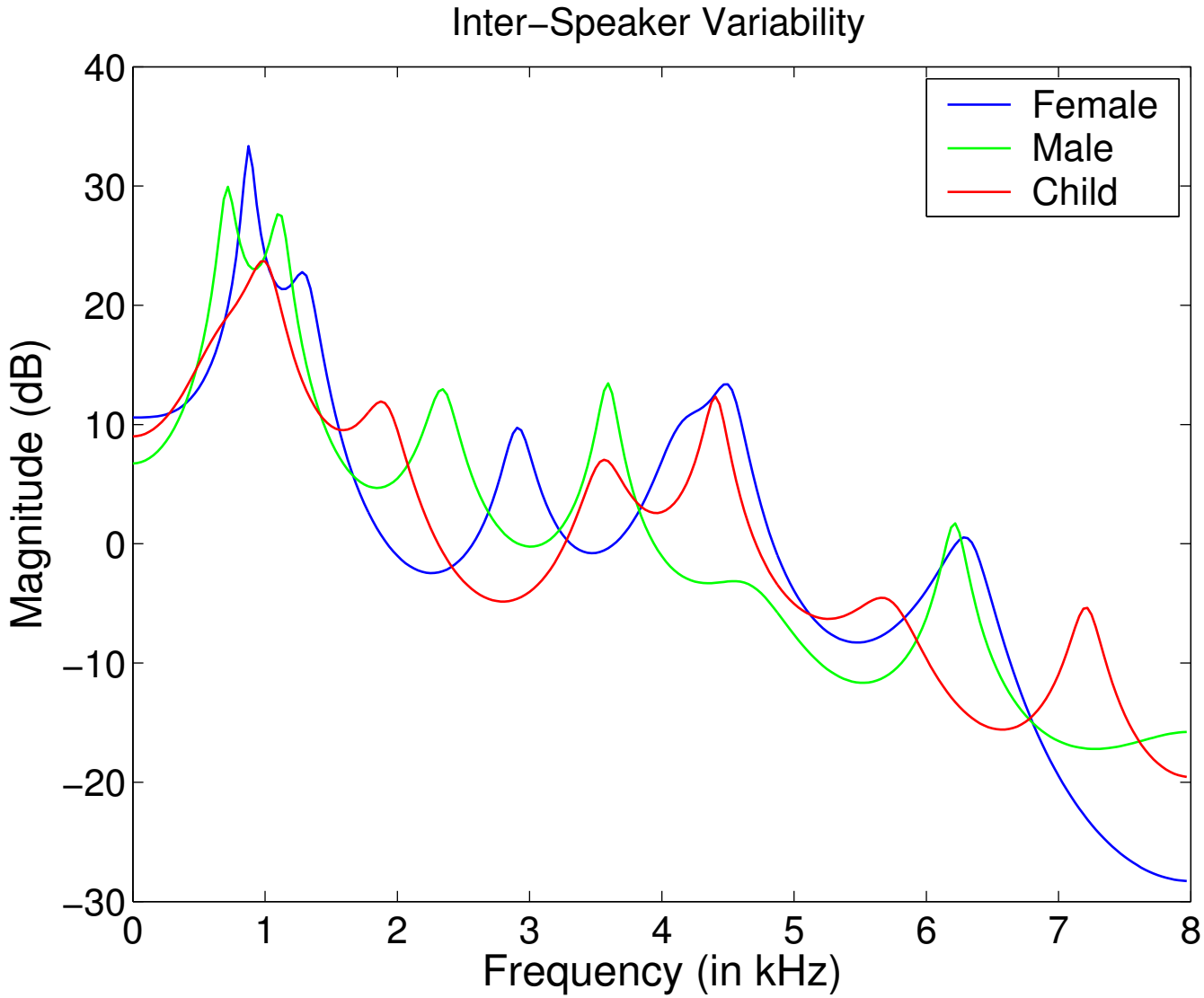
Q1 : Short-term spectral feature vectors are the latent symbols. The set of symbols is undefined, as there is no unique feature vector representation for speech sounds due to variabilities.

Q2 : Short-term speech processing-based feature extraction

Q3 : Short-term speech processing-based feature extraction

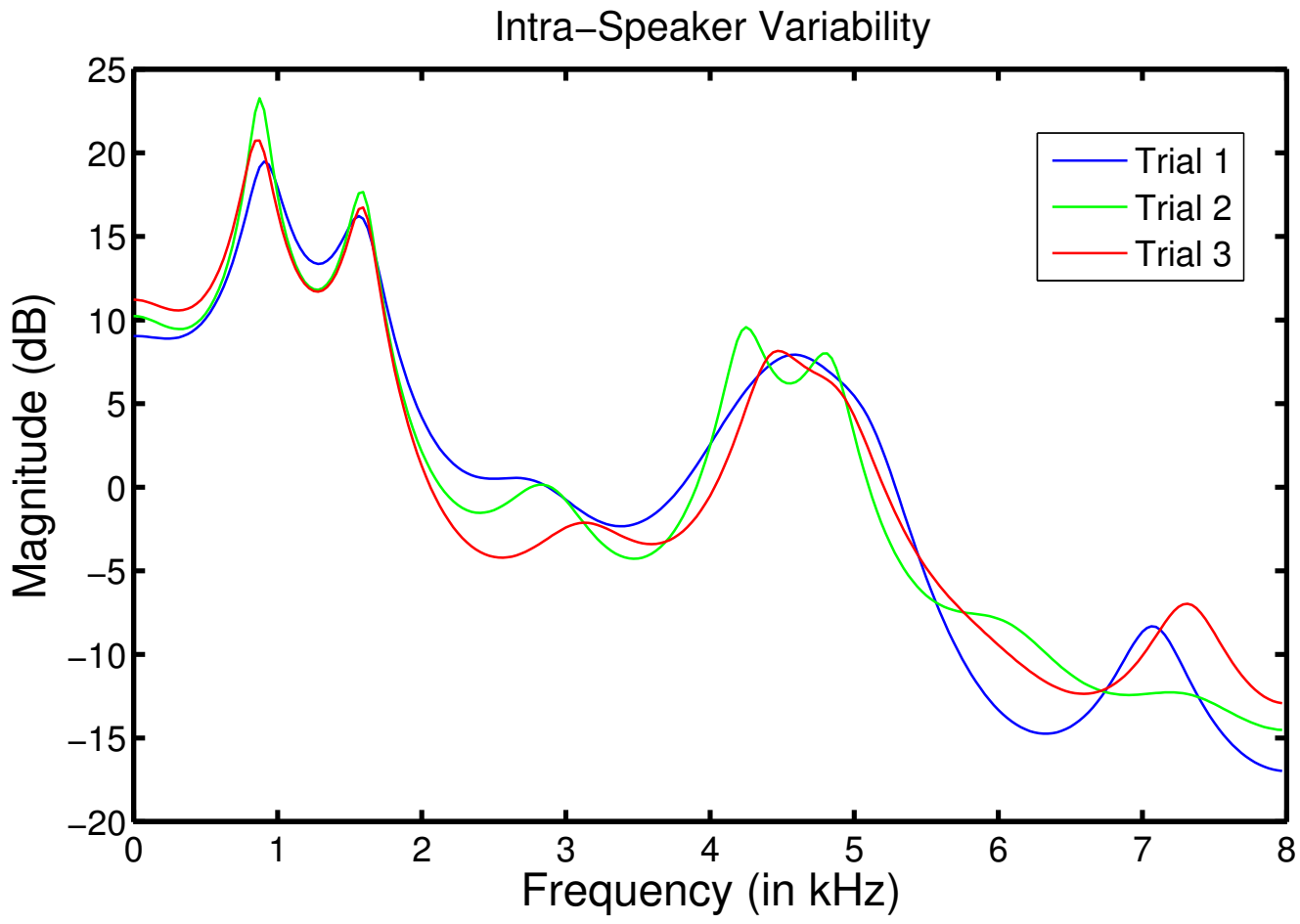
Q4 : Dynamic programming, i.e. DTW, with appropriate local score and local constraints

Inter-speaker variability



Linear prediction spectrum of a frame of sustained vowel /aa/ from different speakers

Intra-speaker variability



Linear prediction spectrum of a frame of sustained vowel /aa/ from the same speaker

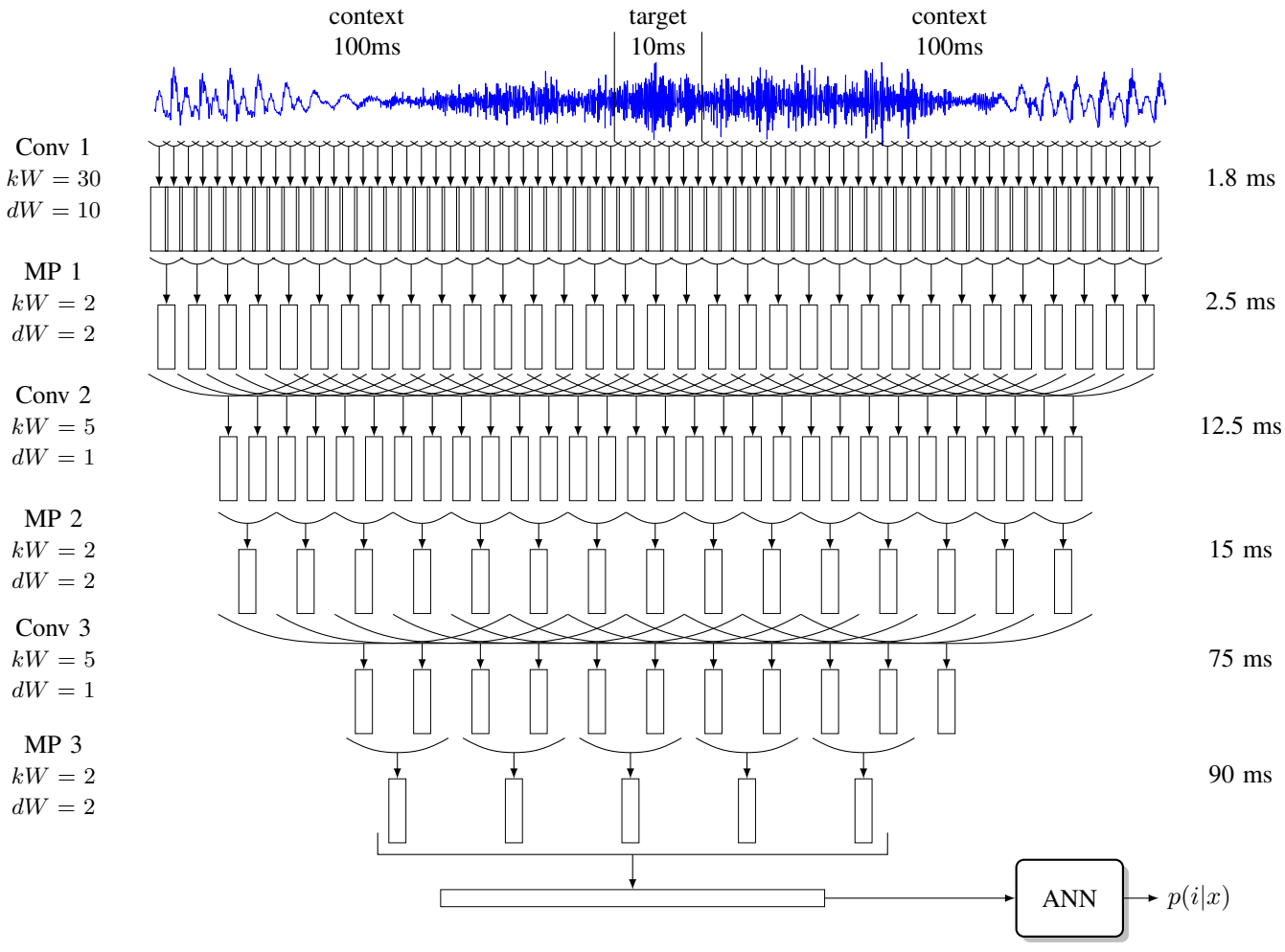
Limitations

- Works well for speaker-dependent, clean and controlled conditions
  - Late 1990s name dialing on mobile phones
- Generalization across speakers and conditions is a highly challenging problem
- Reference templates typically represent word units. Every new word needs a new reference template.
- Getting phone-based reference templates is a non-trivial task
- Large amount of CPU and memory requirements

Pro: No training needed

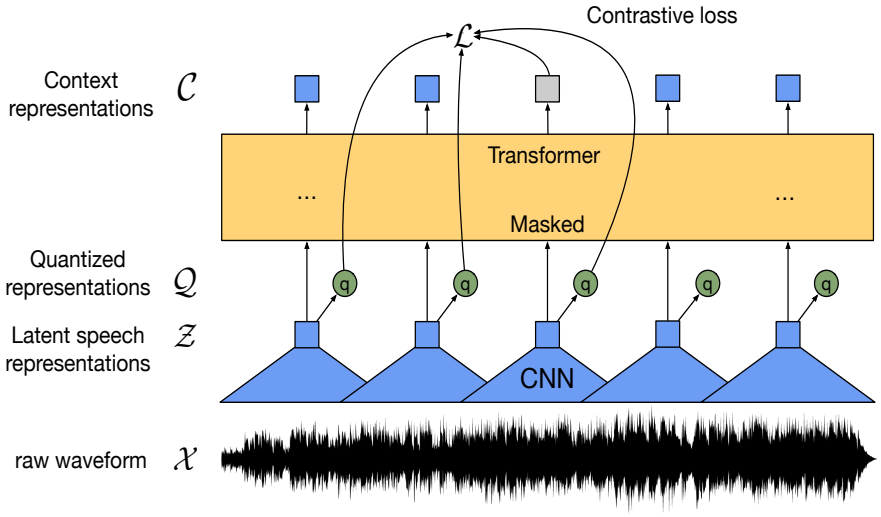
Holy grail: Find the short-term speech processing based feature representation that carries linguistic unit (phone/syllable) related information and is robust to undesirable variabilities.

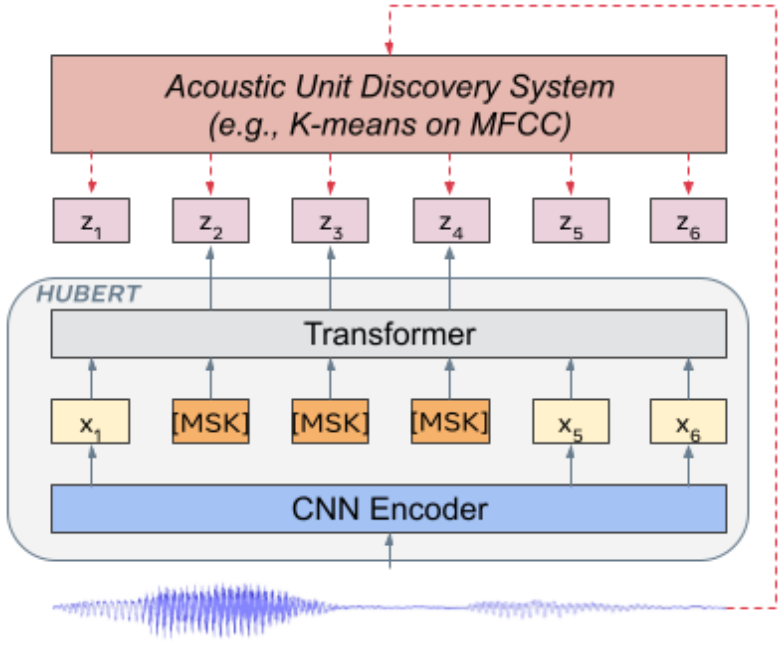
Representation learning (1)



Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," Speech Communication, Vol. 108, April 2019, Pages 15–32.

Representation learning (2)





wav2vec2

HuBERT

wav2vec2: Alexei Baveski et al., wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, arXiv:2006.11477v3, 2020

HuBERT: Wei-Ning Hsu et al., HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, arXiv:2106.07447, 2021

Outline

Automatic speech recognition problem formulation

Language generation (Sequence generation)

String Matching (Sequence comparison)

Knowledge-based approach

Instance-based ASR approach

Next Part

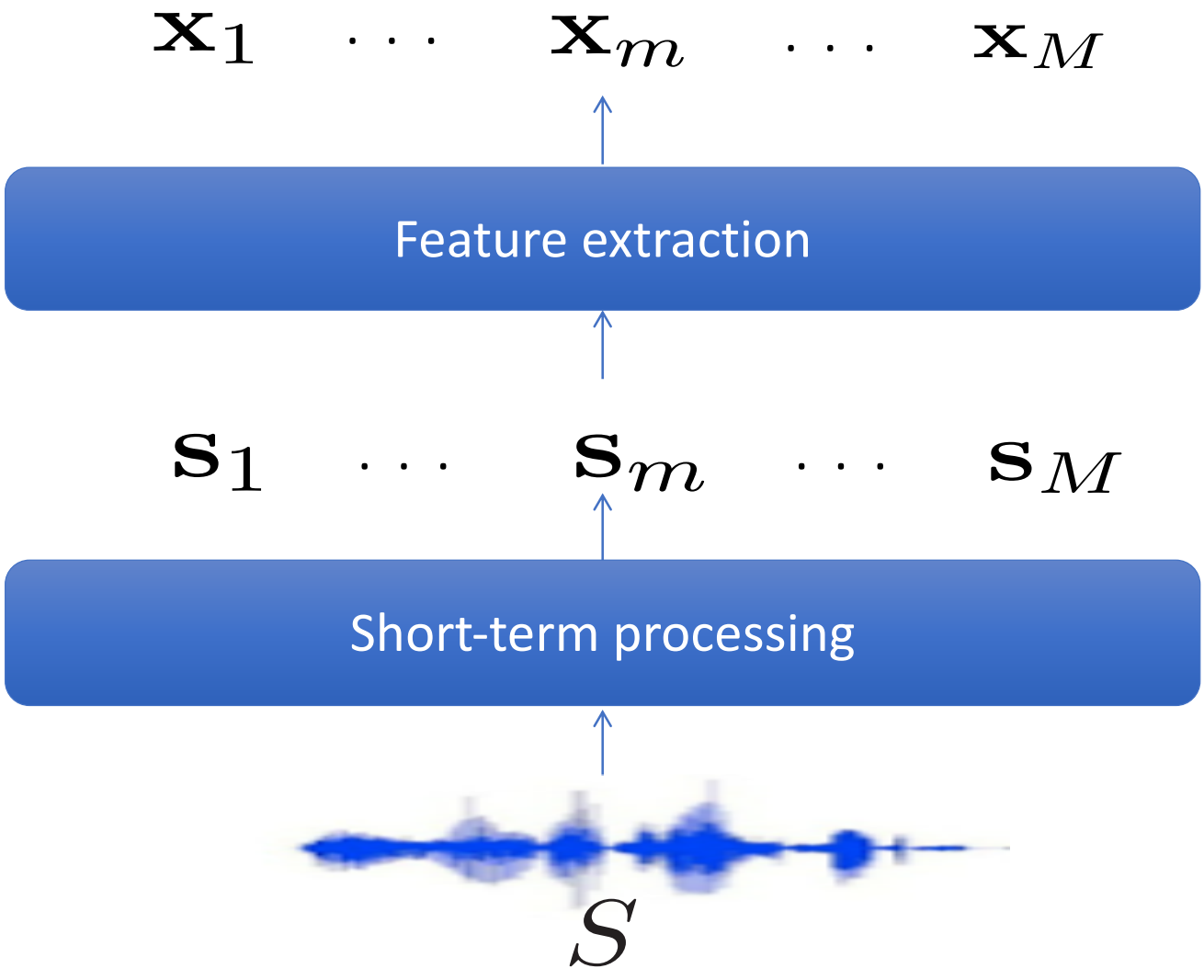
Statistical formulation for matching S and W_k

$$\hat{W} = \arg\max_{W_k \in \mathcal{W}} P(W_k|S) = \arg\max_{W_k \in \mathcal{W}} \frac{p(W_k, S)}{p(S)}$$

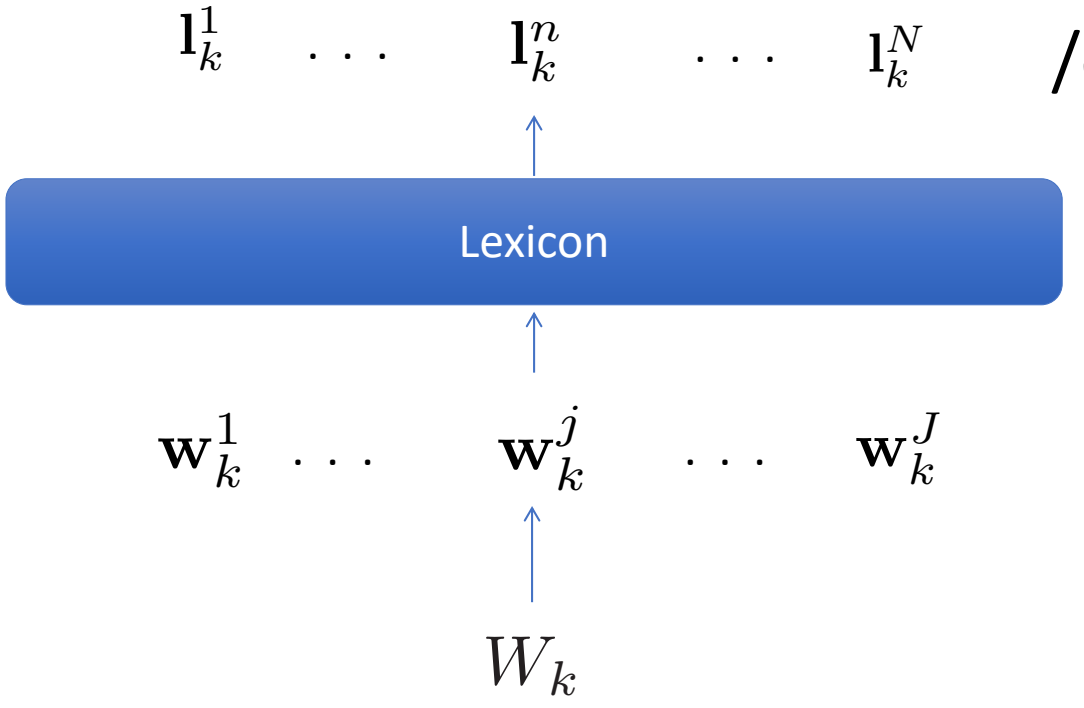
- Likelihood-based approach
  $p(W_k, S)$
- Posterior-based approach
  $P(W_k|S)$

Model-based approach

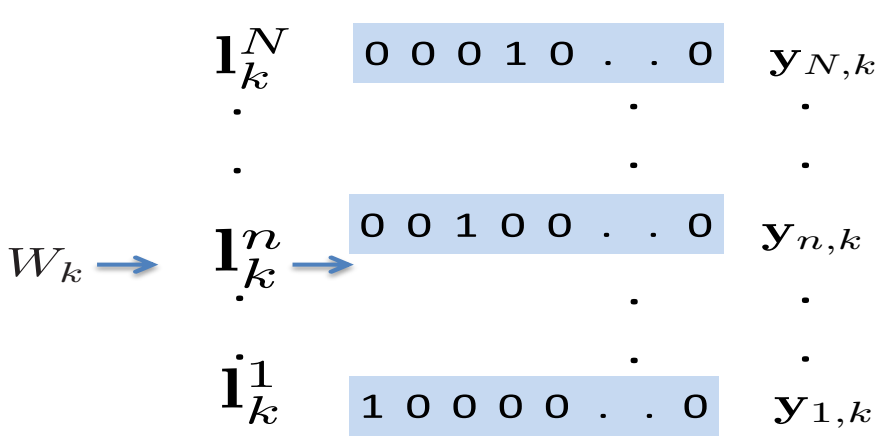
Speech signal S representation

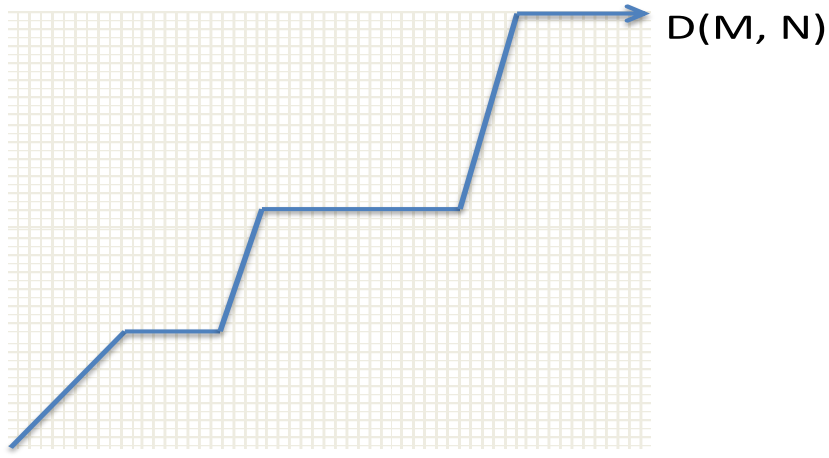


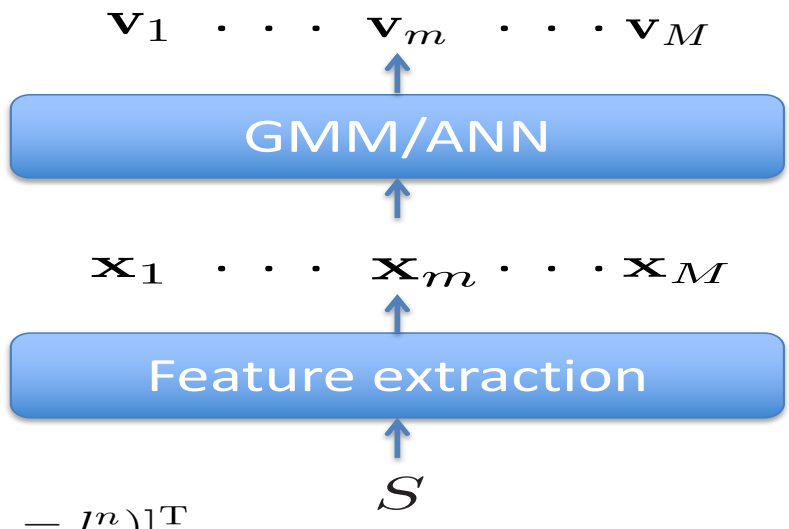
Word hypothesis W_k representation



Matching S and W_k: $P(W_k, S)$







$$\mathbf{v}_m = [p(\mathbf{x}_m|a^1) \cdots p(\mathbf{x}_m|a^D)]^{\mathrm{T}}$$

$$\mathbf{y}_{n,k} = [P(a^1|q_m = l_k^n) \cdots P(a^D|q_m = l_k^n)]^{\mathrm{T}}$$

Thank you for your attention!

Dr. Mathew Magimai Doss

Idiap Research Institute, Martigny, Switzerland