Posterior-based Approach

Dr. Mathew Magimai Doss

Idiap Research Institute, Martigny, Switzerland

Posterior-based ASR Theory (1)

$$W^* = \arg\max_{W_k} P(W_k|X)$$

$$P(W_k|X) = \sum_Q P(W_K, Q|X)$$

$$= \sum_Q P(W_k|X,Q) \cdot P(Q|X)$$

$$\simeq \sum_Q P(Q|X) \cdot P(W_K|Q)$$

$$X = \{\mathbf{x}_1, \cdots \mathbf{x}_m, \cdots \mathbf{x}_M\}$$

$$Q = \{q_1, \cdots q_{m-1}, \cdots q_M\}$$

CONNECTIONIST SPEECH RECOGNITION - A Hybrid Approach, Hervé Bourlard and Nelson Morgan, KLUWER ACADEMIC PUBLISHERS, 1994

Posterior-based ASR Theory (2)

$$P(Q|X) = \prod_{m=1}^{M} P(q_m|X, Q_1^{m-1}) \quad \text{applying Gibbs rule}$$

$$Q_1^{m-1} = \{q_1, \cdots q_{m-1}\}$$

$$P(W_k|Q) = \frac{P(Q|W_k) \cdot P(W_k)}{P(Q)} \quad \text{applying Bayes' rule}$$

$$= \frac{(\prod_{m=1}^{M} P(q_m|Q_1^{m-1}, W_k) \cdot P(W_k)}{\prod_{m=1}^{M} P(q_m|Q_1^{m-1})} \quad \text{applying Gibbs rule}$$

$$P(W_k|X) \simeq \sum_{Q} \left(\prod_{m=1}^{M} P(q_m|X, Q_1^{m-1}) \right) \cdot \left(\frac{(\prod_{m=1}^{M} P(q_m|Q_1^{m-1}, W_k) \cdot P(W_k)}{\prod_{m=1}^{M} P(q_m|Q_1^{m-1})} \right)$$

$$P(W_k|X) \approx \max_{Q} \left(\prod_{m=1}^{M} P(q_m|X, Q_1^{m-1}) \right) \cdot \left(\frac{(\prod_{m=1}^{M} P(q_m|Q_1^{m-1}, W_k) \cdot P(W_k)}{\prod_{m=1}^{M} P(q_m|Q_1^{m-1})} \right) \quad \text{Viterbi approx.}$$

Posterior-based ASR Theory (3)

$$P(W_k|X) \simeq \sum_Q (\prod_{m=1}^{M} P(q_m|X, q_{m-1}) \cdot (\frac{(\prod_{m=1}^{M} P(q_m|q_{m-1}, W_k) \cdot P(W_k)}{\prod_{m=1}^{M} P(q_m|q_{m-1})})$$

$$P(W_k|X) \approx \max_Q (\prod_{m=1}^{M} P(q_m|X, q_{m-1}) \cdot (\frac{(\prod_{m=1}^{M} P(q_m|q_{m-1}, W_k) \cdot P(W_k)}{\prod_{m=1}^{M} P(q_m|q_{m-1})})$$

$$P(q_m|X, q_{m-1}) = \text{Conditional prior probabilities}$$
$$P(q_m|q_{m-1}, W_k) = \text{similar to HMM transition prob.}$$
$$P(q_m|q_{m-1}) = \text{Training data priors (time-invariant HMM)}$$

Acoustic modeling, language modeling and decoder are not independent any more

Hybrid HMM/ANN ASR

$$P(W_k|X) \approx \max_Q \Big(\prod_{m=1}^{M} P(q_m|X, q_{m-1}) \cdot \Big(\frac{(\prod_{m=1}^{M} P(q_m|q_{m-1}, W_k) \cdot P(W_k)}{\prod_{m=1}^{M} P(q_m|q_{m-1})} \Big)$$

$$\approx \max_Q \Big(\prod_{m=1}^{M} P(q_m|X_{m-c}^{m+c}) \Big) \cdot \Big(\frac{(\prod_{m=1}^{M} P(q_m|q_{m-1}, W_k) \cdot P(W_k)}{\prod_{m=1}^{M} P(q_m)} \Big)$$

$$\approx \max_Q \left(\prod_{m=1}^{M} \frac{P(q_m|X_{m-c}^{m+c})}{P(q_m)} \cdot P(q_m|q_{m-1}, W_k) \right) \cdot P(W_k)$$

$$X_{m-c}^{m+c} = \{\mathbf{x}_{m-c} \cdot \mathbf{x}_m, \cdots \mathbf{x}_{m+c}\} \quad \text{Limited context}$$

$$\frac{P(q_m|X_{m-c}^{m+c})}{P(q_m)} = \text{emission scaled-likelihood}$$

$$P(q_m|q_{m-1}, W_k) = \text{HMM transition prob.}$$

CTC

$$P(q_m|X)$$

- Drop dependency on previous states
- Model full acoustic context using transformers
- Frame shift: 40 ms (25 frames per second)
  - With 10 ms frame shift the approach can lead to inferior performance
- Introduce empty symbols to emulate string matching at the output

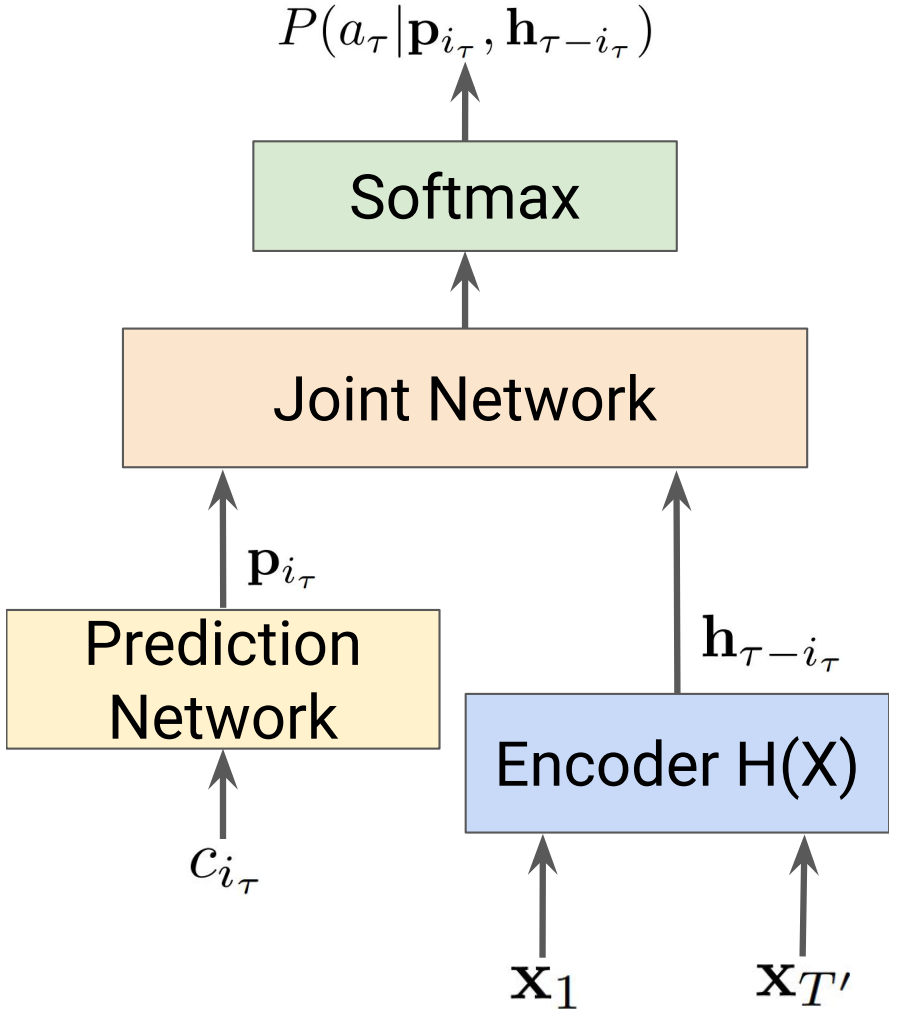
Connectionist Temporal Classification

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

RNN-T

$$\left(\prod_{m=1}^{M} P(q_m | X, Q_1^{m-1}) \right)$$

- Model dependency on previous states (Prediction Network)
- Model full acoustic context using transformers (Encoder)
- Frame shift: 40 ms (25 frames per second)
- Introduce empty symbols to emulate string matching at the output

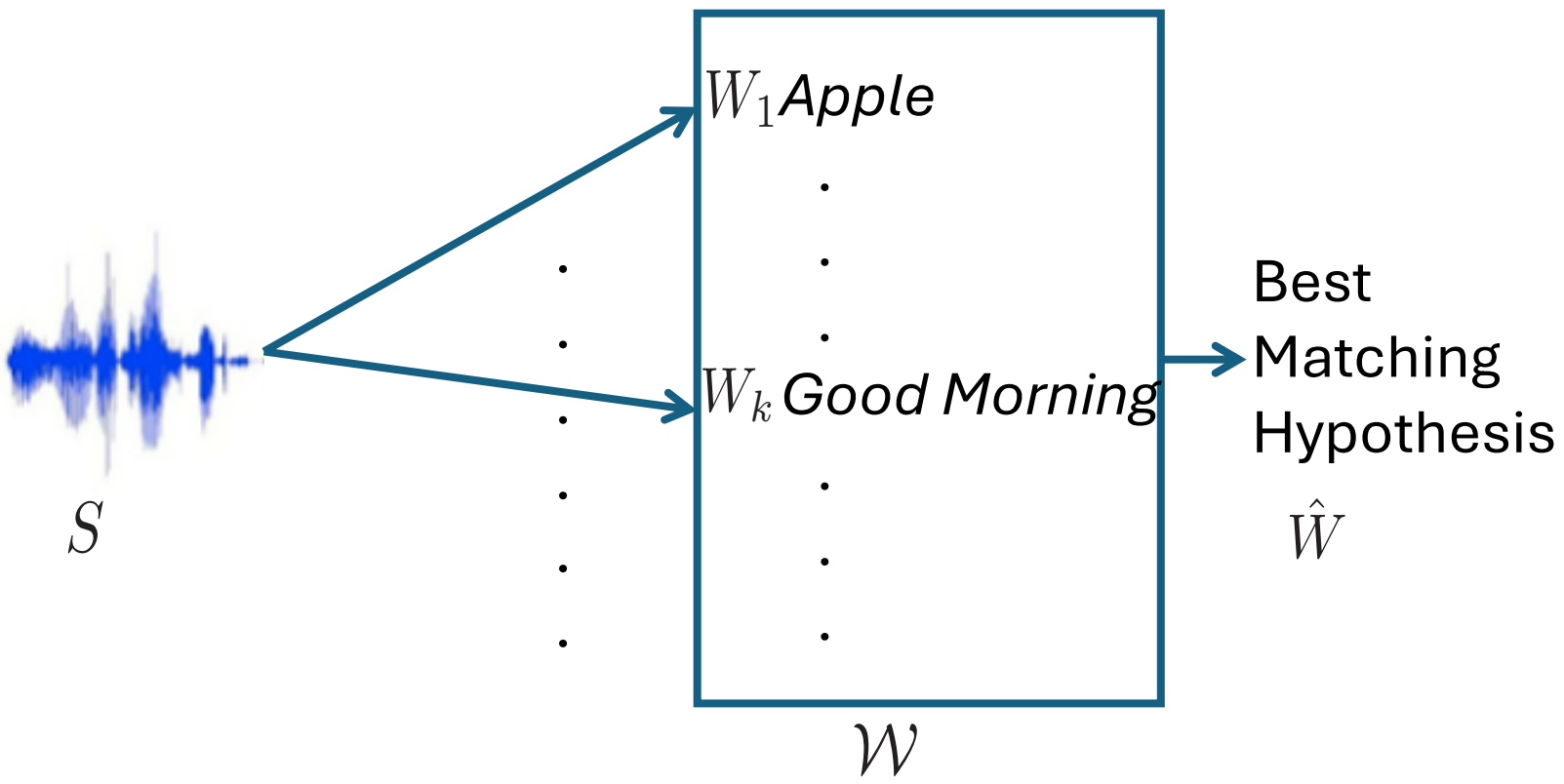


Recurrent Neural Network-Transducer

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

Sequence recognition

Automatic speech recognition (ASR)



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

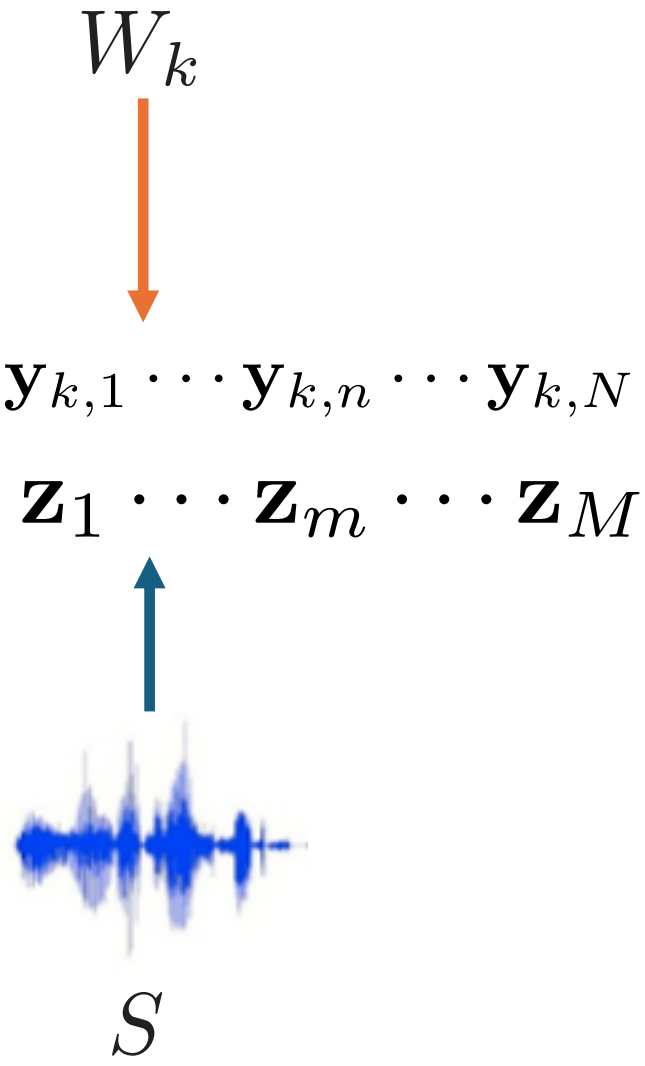
- **How to model the set of word hypothesis $\mathcal{W}$?**
  - Sequence generation

  Rule-based, discrete Markov model (DMM), neural LM, LLM
- **How to match W_k and S ?**
  - Sequence matching
- **How to search?**
  - Greedy search, beam search, stack decoding

Sequence matching: Four fundamental questions

Core idea

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



Q1: What is the shared latent symbol set $\{a^d\}_{d=1}^{D}$?

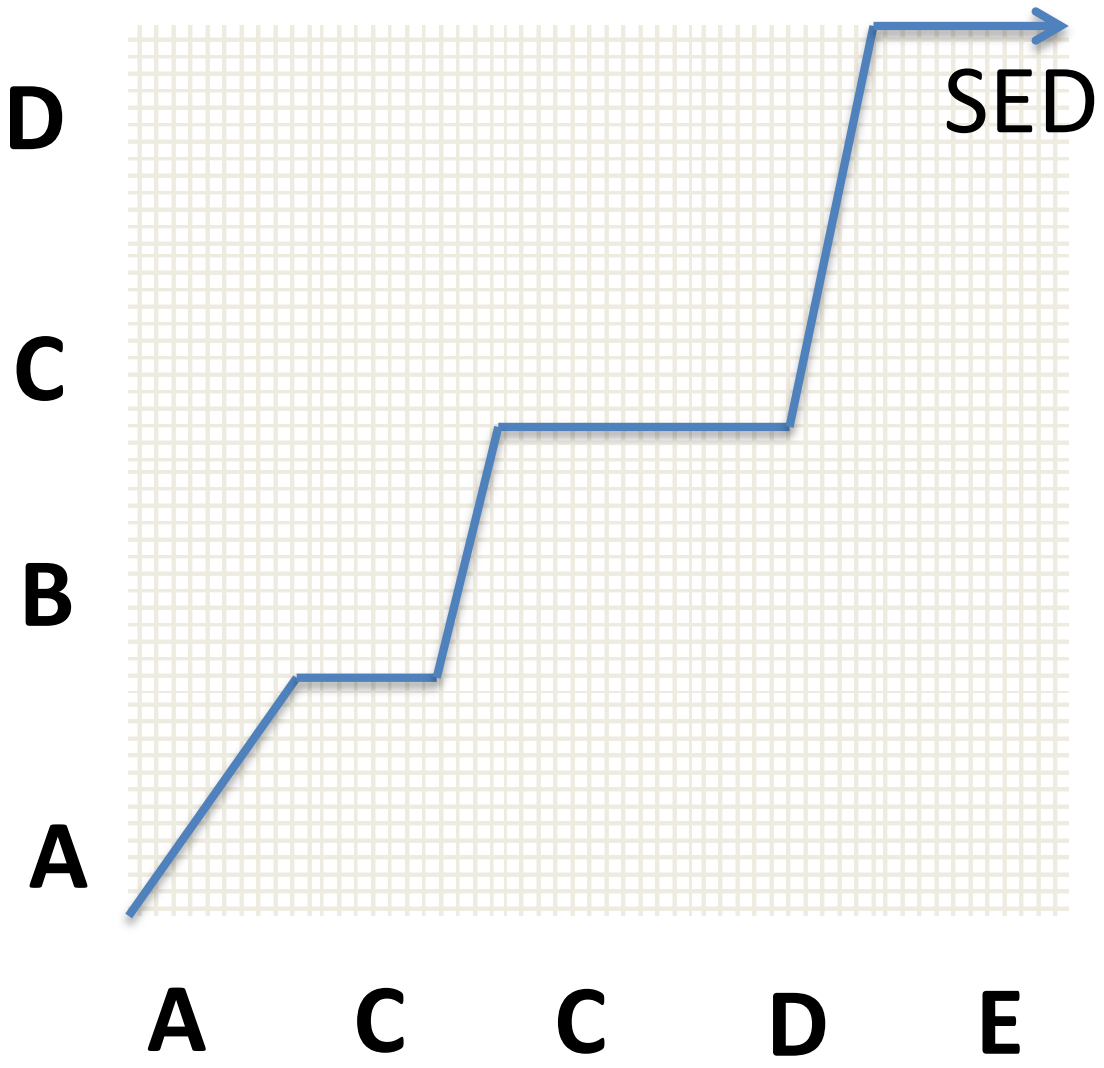
Q2: How to map S to latent symbol sequence Z?

Q3: How to map W_k to latent symbol sequence Y_k?

Q4: How to match Z and Y_k to estimate $\mathrm{Match}(W_k, S)$?

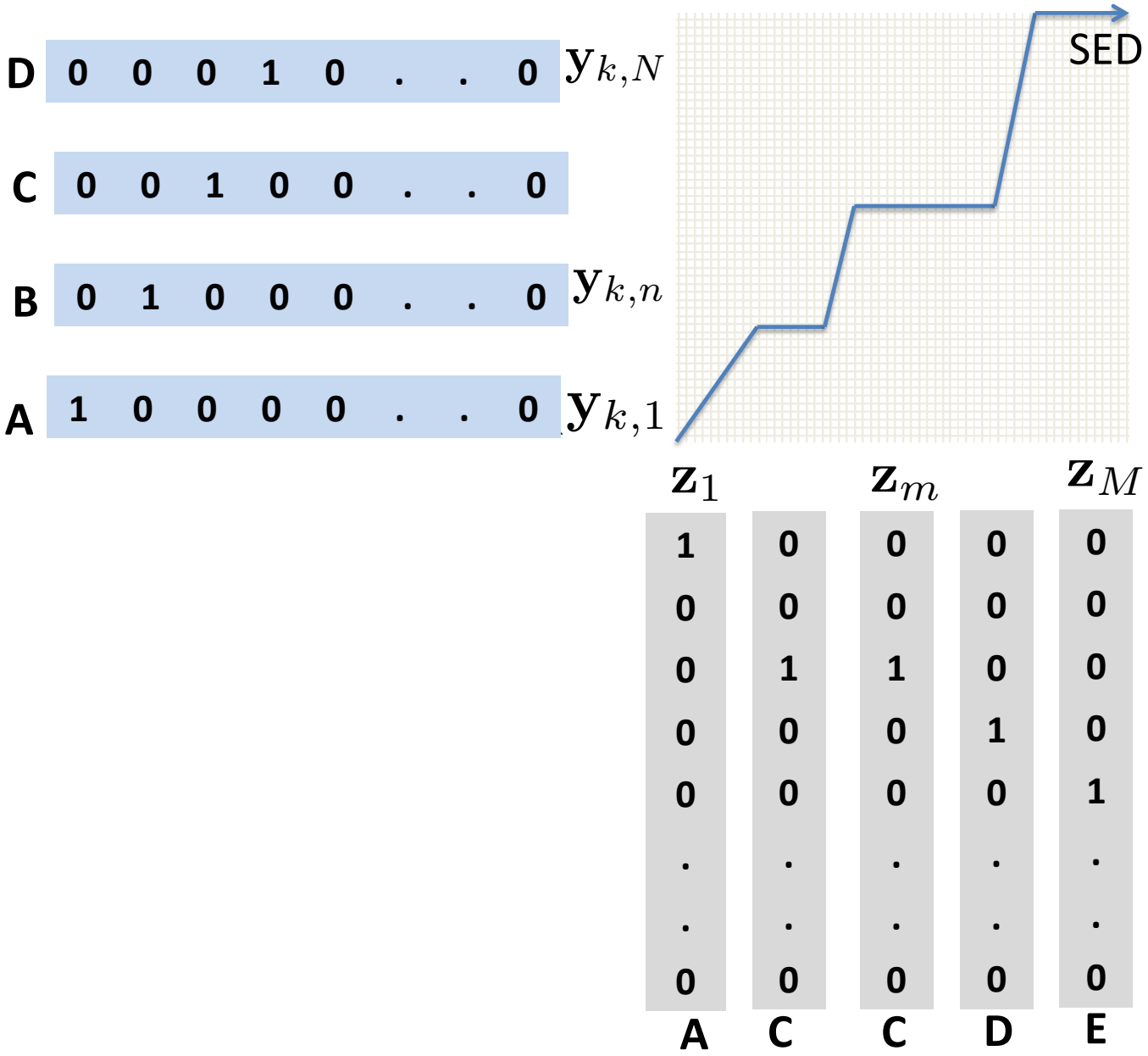
On Modeling Context-Dependent Clustered States: Comparing HMM/GMM, Hybrid HMM/ANN and KL-HMM Approaches, Marzieh Razavi, Ramya Rasipuram and Mathew Magimai-Doss, in Proc. of ICASSP, 2014.

String matching Q4




```
Initial conditions
for n from 1 to N
  for m from 1 to M
    if str[m] = str[n]
       d[m,n] = 0      // Local score
    else
        d[m,n] = 1
    done
  D[m, n] = minimum(
                  D[m-1, n] + 1, // deletion
                  D[m, n-1] + 1,  //insertion
                  D[m-1,n-1] + d[m,n]  //substitution
                              )    // Global score
  done
done
        String edit distance SED = D[M,N]
```

Deterministic symbol sequence matching



- Set of symbols

$$\{a^d\}_{d=1}^{D} = \{\mathrm{A}, \cdots, \mathrm{Z}\}$$

- Deterministic symbols

$$\mathbf{y}_{k,n} = \delta^d \quad \mathbf{z}_m = \delta^{d'}$$

- Local score d[m,n]

$$\mathrm{if}(\mathrm{KL}(\mathbf{y}_{k,n}, \mathrm{z}_m) \leq \Delta) \quad \Delta = 0$$

$$\mathrm{d}[\mathrm{m}, \mathrm{n}] = 0$$

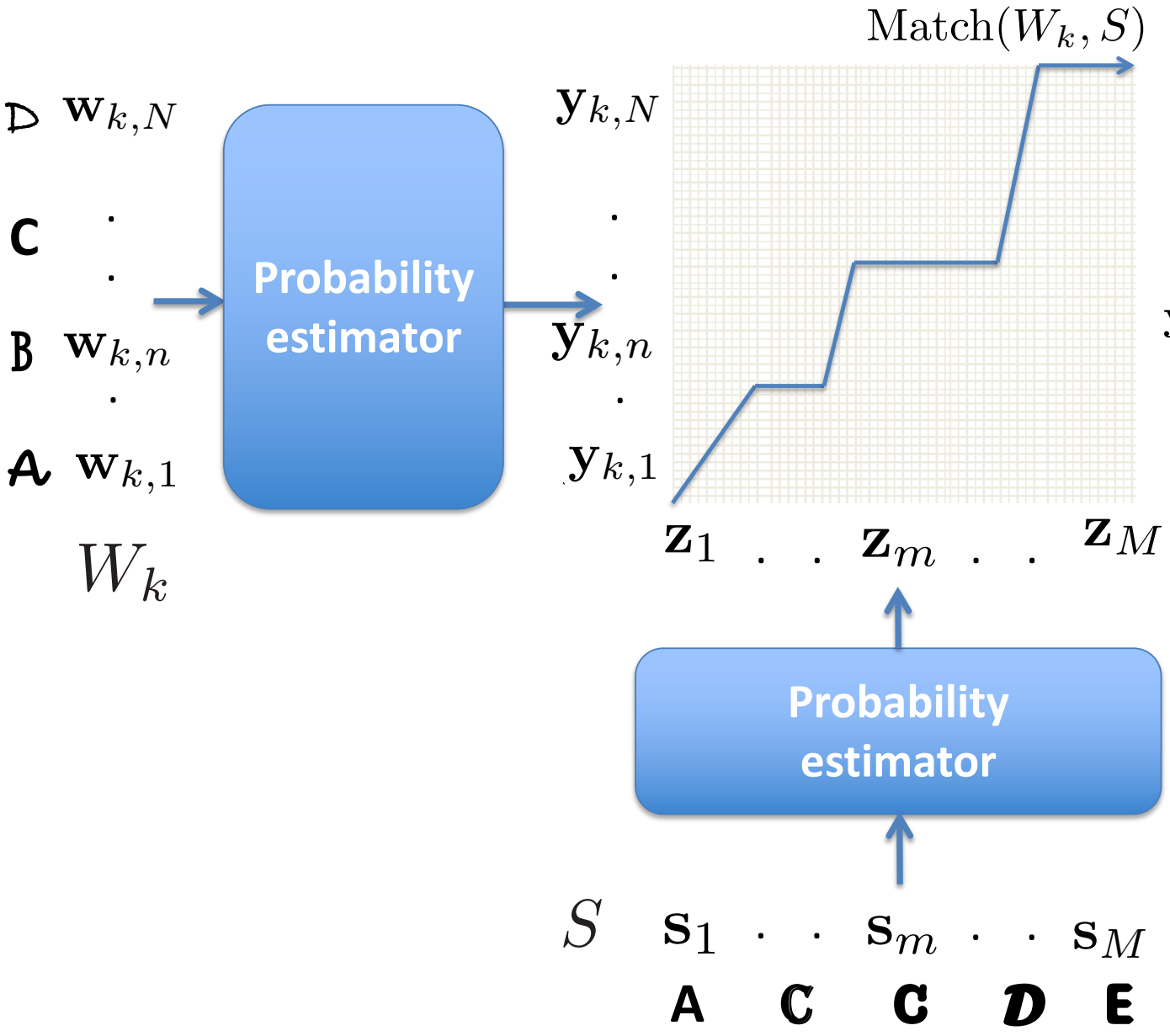
$$\mathrm{else}$$

$$\mathrm{d}[\mathrm{m}, \mathrm{n}] = 1$$

Kullback-Leibler (KL) divergence

$$\mathrm{KL}(\mathbf{y}_{k,n}, \mathbf{z}_m) = \sum_{d=1}^{D} y_{k,n}^d \cdot \log(\frac{y_{k,n}^d}{z_m^d})$$

Probabilistic symbol sequence matching



- Set of symbols

$$\{a^d\}_{d=1}^{D} = \{\mathrm{A}, \cdots, \mathrm{Z}\}$$

- ***Posterior features*** (Prob. symbols)

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

$$\mathbf{z}_m = [P(a^1|\mathbf{s}_1) \cdots P(a^d|\mathbf{s}_m) \cdots P(a^D|\mathbf{s}_M)]^{\mathrm{T}}$$

- Local score d[m,n]

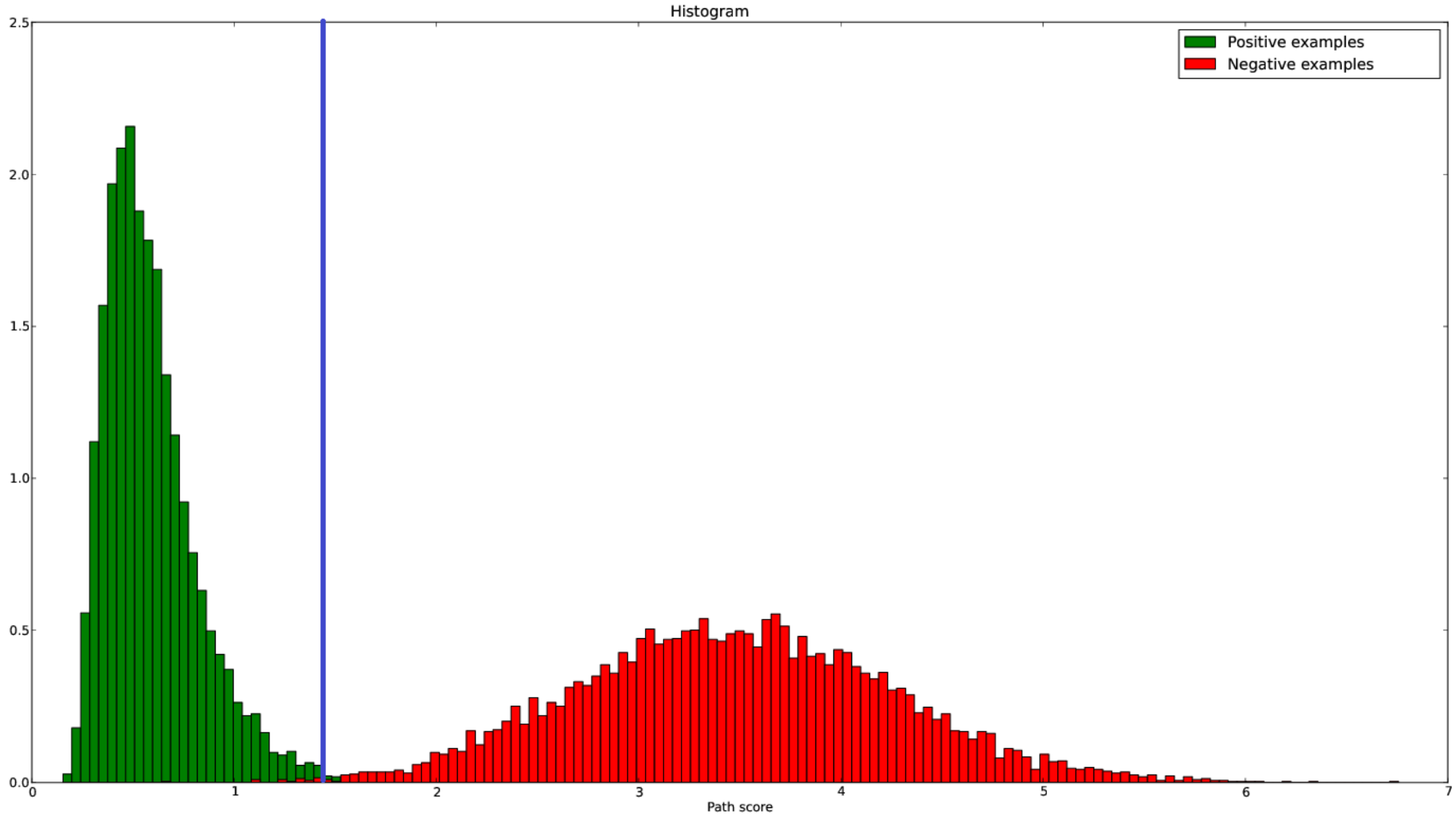
$$\mathrm{d}[\mathrm{m}, \mathrm{n}] = \mathrm{KL}(\mathbf{y}_{k,n}, \mathbf{z}_m)$$

- Global score recursion

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

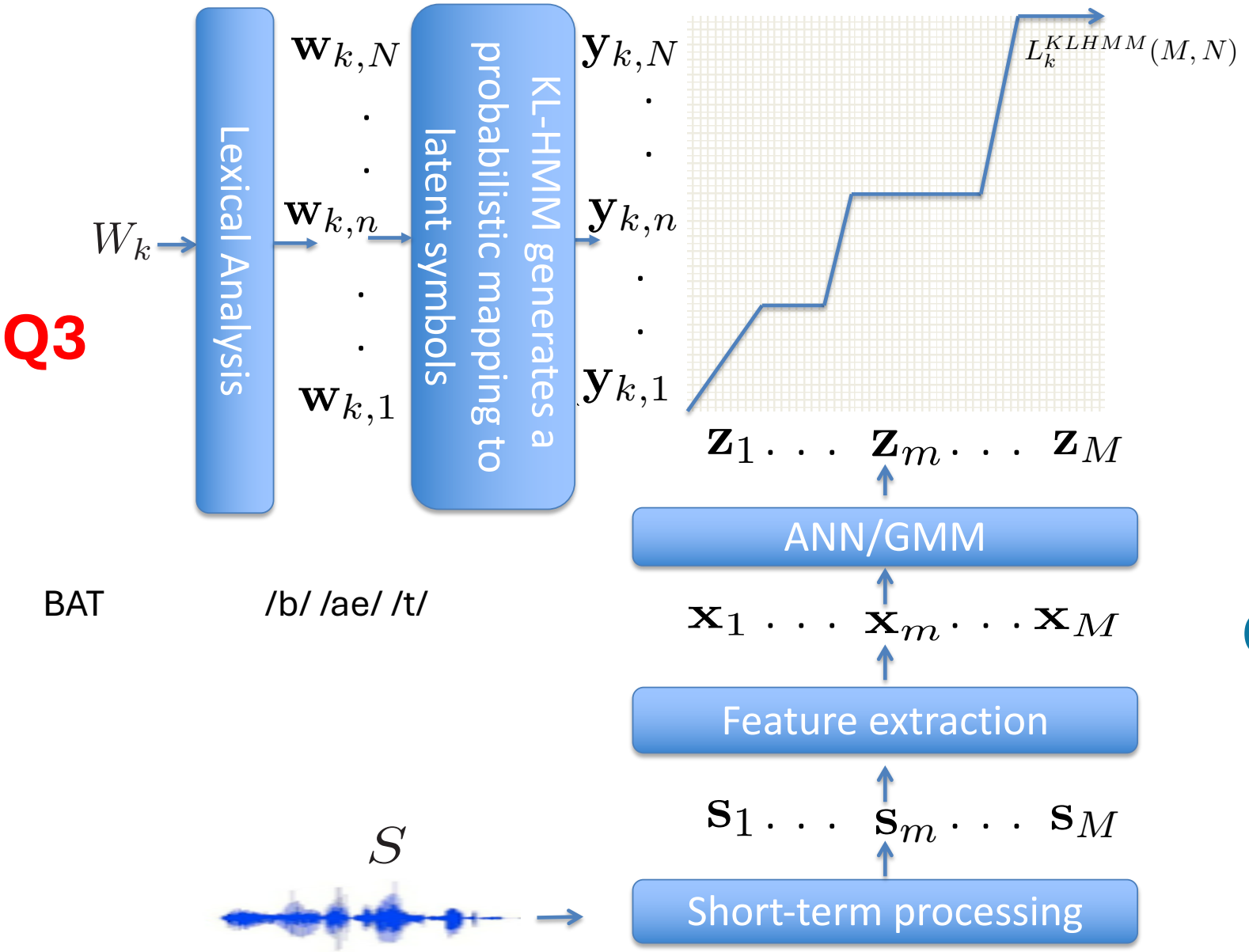
$$\mathrm{Match}(W_k, S) = -\mathrm{D}[\mathrm{M}, \mathrm{N}]$$

Probabilistic symbols sequence matching



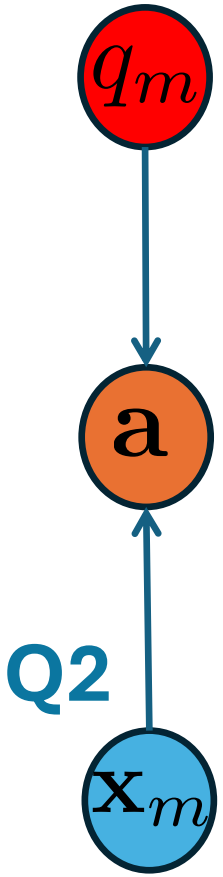
$$\mathrm{Match}(W_k, S)$$

Posterior feature-based speech processing (1)



Q3

BAT /b/ /ae/ /t/

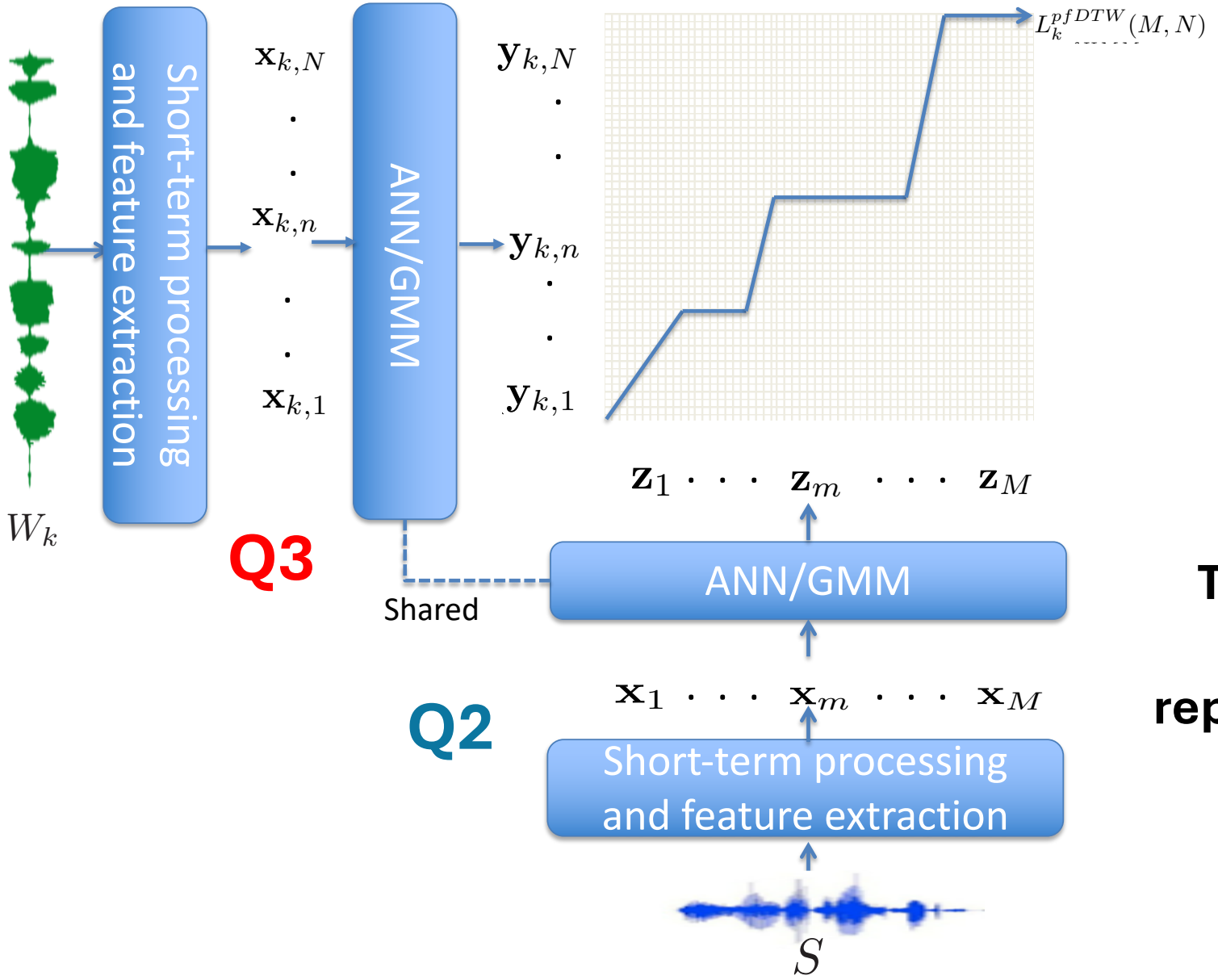


Q2

Represention of word hypothesis by “text”

Assumption: Target language has a writing system

Posterior feature-based speech processing (2)

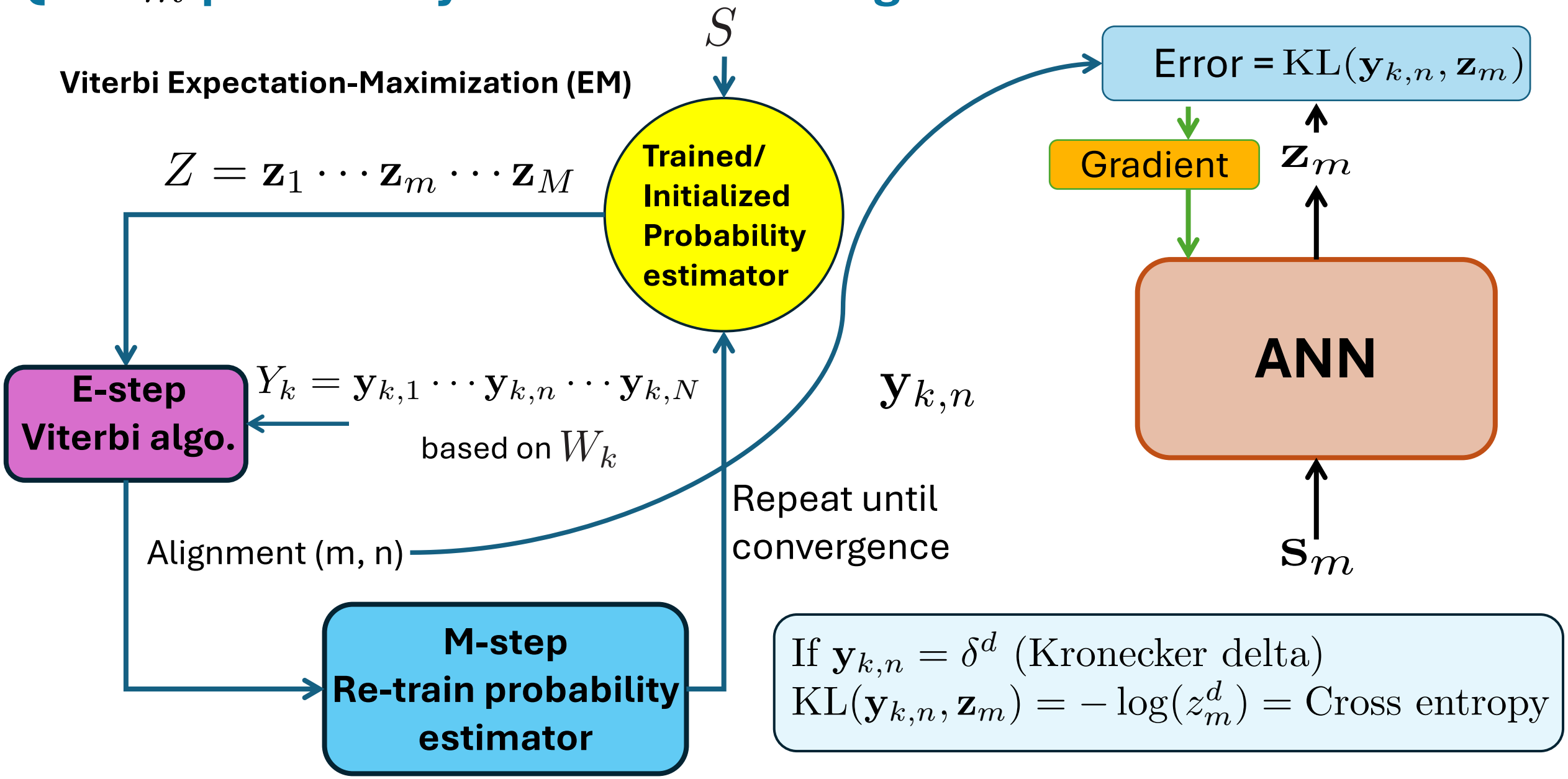


Representation of word hypothesis by acoustic instance(s)

Each time frame can be regarded as an HMM state

Text-based representation and acoustic instance-based representation of word hypothesis are interchangeable

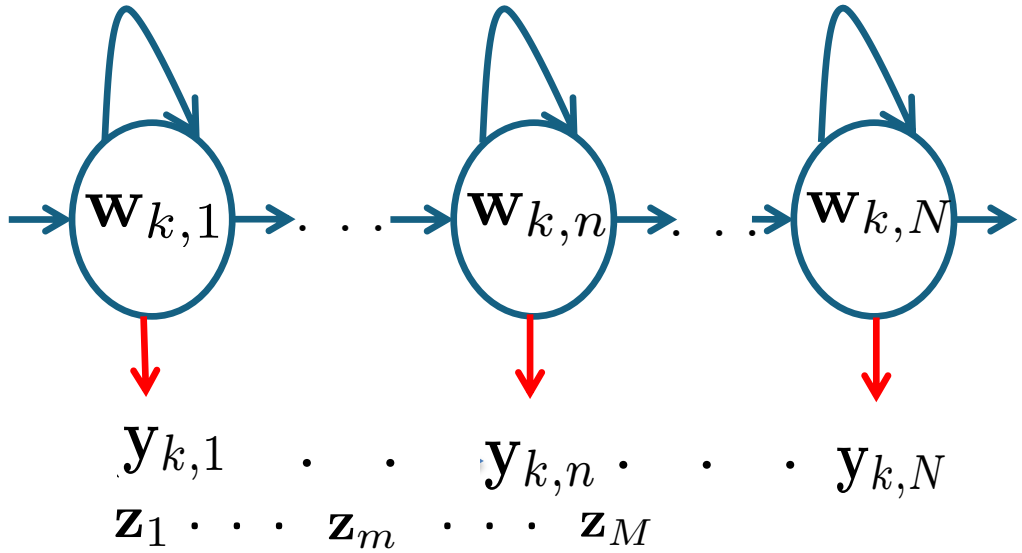
Q2: $\mathbf{z}_m$ probability estimator training

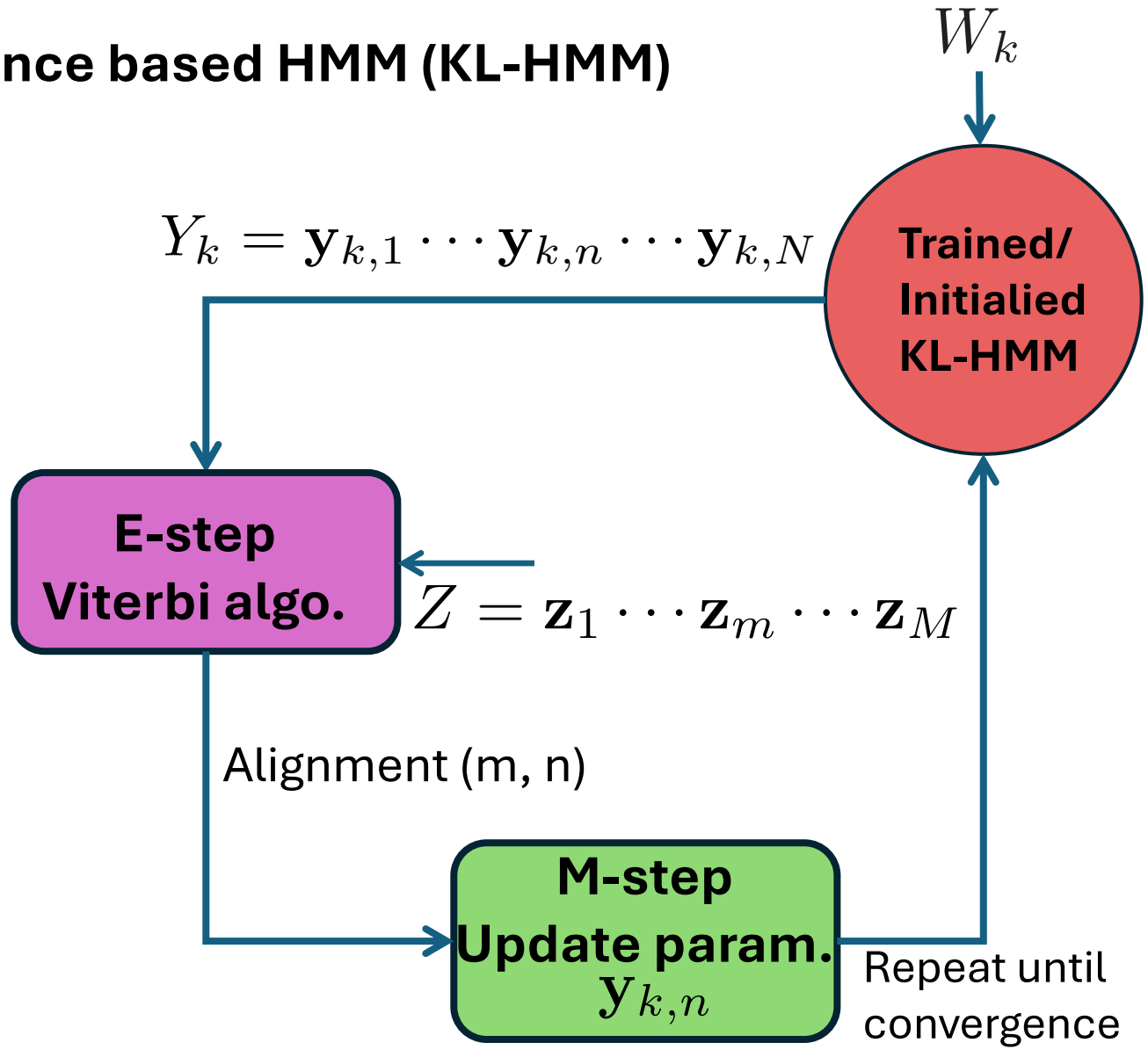


Segment-level training of ANNs based on acoustic confidence measures for hybrid HMM/ANN Speech Recognition, S. P. Dubagunta and M. Magimai-Doss, Proc. of ICASSP, 2019

Q3: $\mathbf{y}_{k,n}$ parameters estimation

Kullback-Leibler divergence based HMM (KL-HMM)





G. Aradilla, J. Vepa and H. Bourlard, An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features, Proc. Of ICASSP. 2007

G. Aradilla, H. Bourlard and M. Magimai.-Doss, Using KL-based Acoustic Models for Large Vocabulary Recognition task, Proc. of Interspeech, 2008.

Comparison of approaches

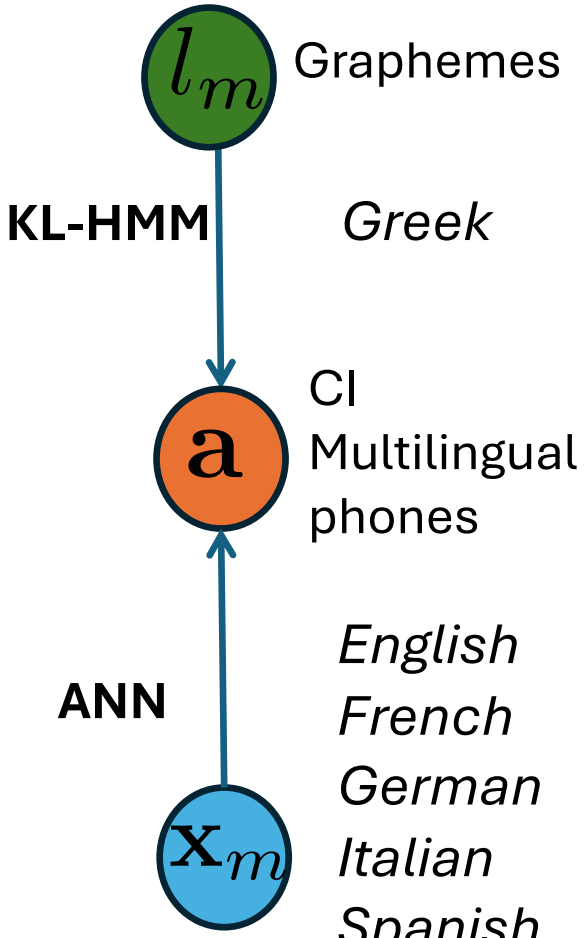
	Q1	Q2	Q3	Q4	Match( )
HMM-based Text only W_k	Phones or Graphemes	Likelihood estimation GMM or ANN	Deterministic	Dyn. Prog.	$p(W_k, S)$
End-to-end Text only W_k	Phones or Graphemes	Probability estimation ANN	Deterministic	Dyn. Prog.	$P(W_k \mid S)$
Posterior feature-based	Different possibilities	Probability estimation Many possible estimators	Probabilistic	Dyn. Prog.	$\frac{p(S \mid W_k)}{p(S \mid \neg W_k)}$
Instance-based (spec.)	Spectral feat. based	Spec. feat. vector extract.	Spec. feat. vector extract.	Dyn. Prog.	$\frac{p(S \mid W_k)}{p(S \mid \neg W_k)}$

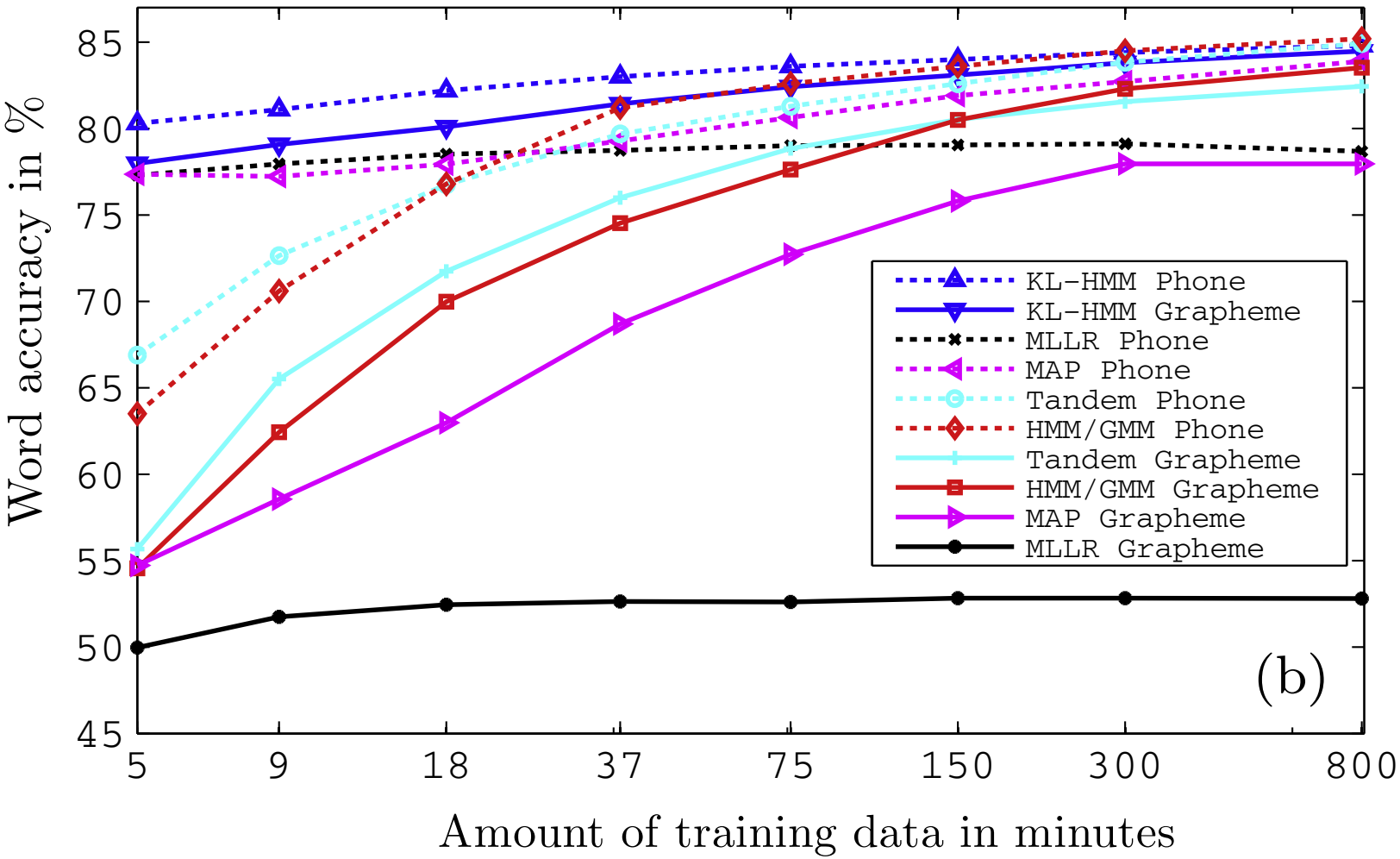
Posterior feature-based ASR Demo

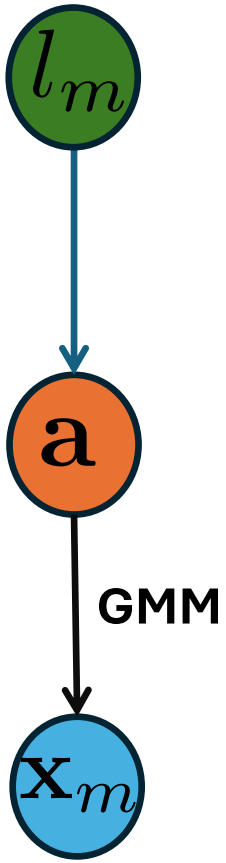
- Instance-based approach
- Posterior feature extractor
  - Two hidden layer MLP trained to classify context-independent phones
  - Input: nine frames of 39 dimensional cepstral feature vector (13 static + 13 delta + 13 delta-delta)
  - Cross entropy loss
  - Stochastic gradient descent

Posterior Features for Template-based ASR, Serena Soldo, Mathew Magimai-Doss, Joel Praveen Pinto and Hervé Bourlard, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Prague, 2011
Synthetic References for Template-based ASR using Posterior Features, Serena Soldo, Mathew Magimai-Doss and Hervé Bourlard, in: Proceedings of Interspeech, 2012

Handling acoustic and lexical resource constraints





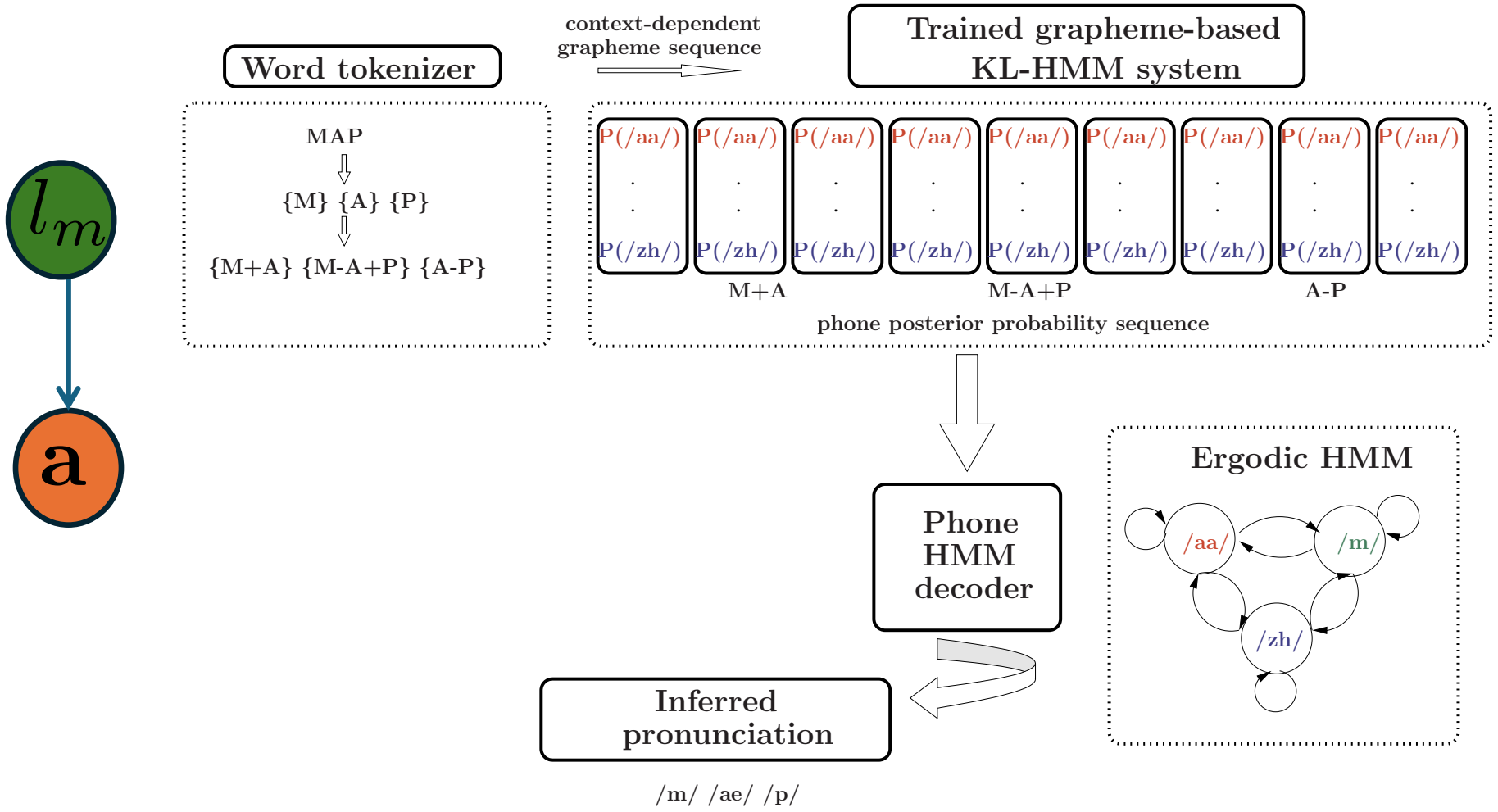


No transcription available: start training by initializing KL-HMM parameters based on prior G2P knowledge

Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model, Ramya Rasipuram and Mathew Magimai.-Doss, in: Speech Communication, 68:23–40, 2015

Probabilistic Lexical Modeling and Unsupervised Training for Zero-Resourced ASR, Ramya Rasipuram, Marzieh Razavi and Mathew Magimai.-Doss, in: Proc. of ASRU, 2013

Lexical resource development



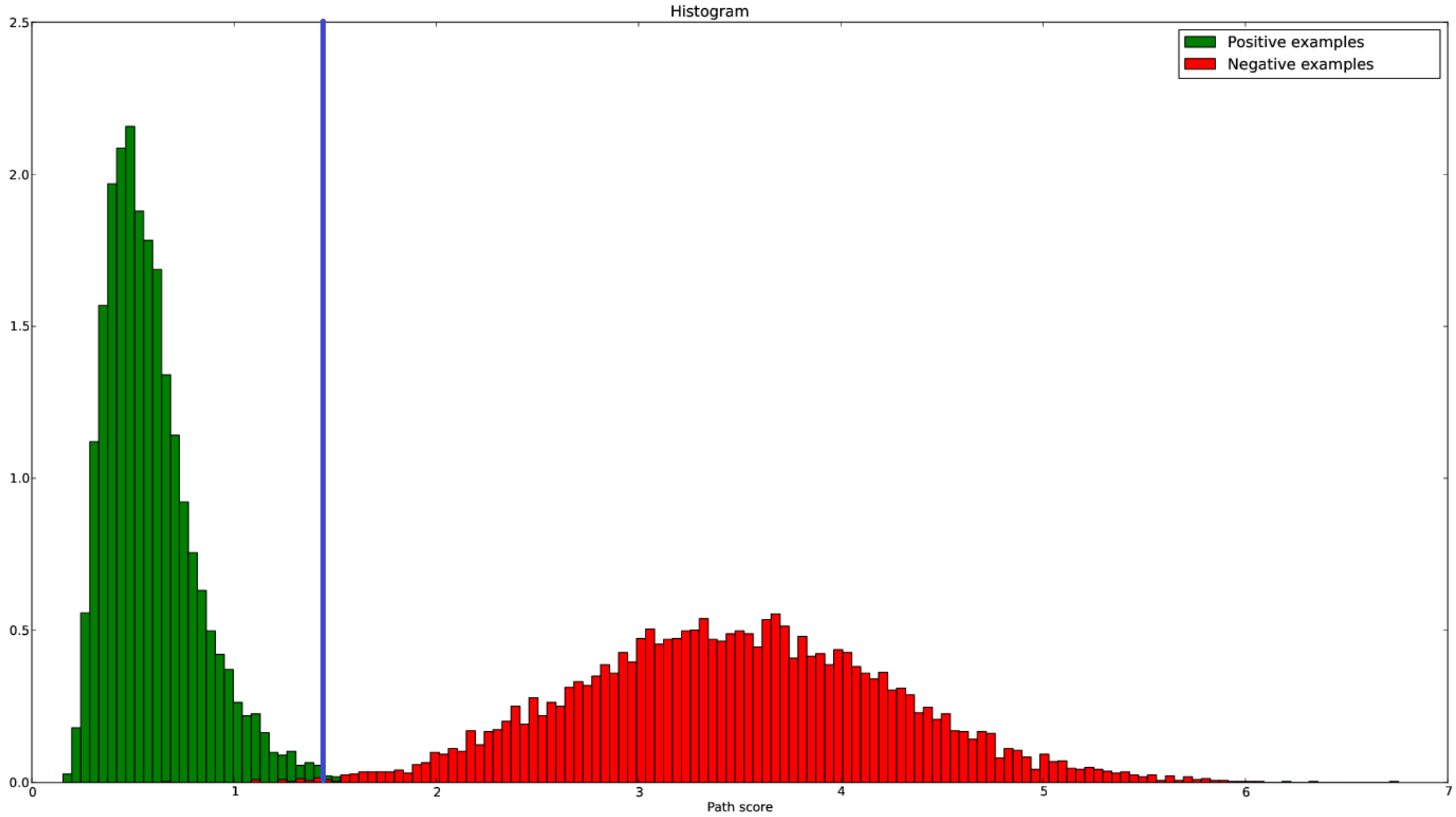
- Can be extended to automatic subword derivation and lexicon development
- Combination of multiple G2P converters
- Unified framework for acoustic-to-phone conversion and G2P conversion

Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework, M. Razavi, R. Rasipuram and M. Magimai.-Doss, in: Speech Communication, 80, 2016

Towards Weakly Supervised Acoustic Subword Unit Discovery and Lexicon Development Using Hidden Markov Models, M. Razavi, R. Rasipuram and M. Magimai-Doss, in: Speech Communication, 96:168-183, 2018

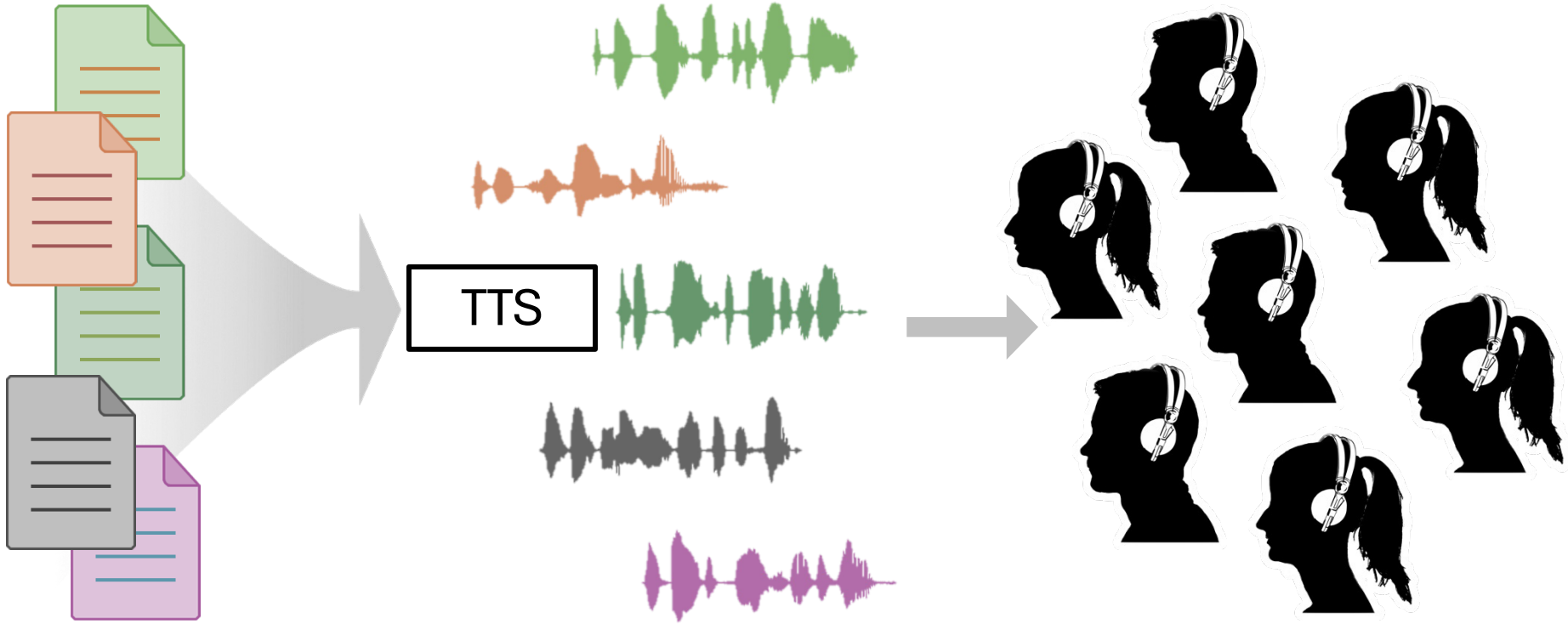
M. Razavi, On Modeling the Synergy Between Acoustic and Lexical Information for Pronunciation Lexicon Development, PhD Thesis 7851, EPFL, 2017

Probabilistic symbols sequence matching

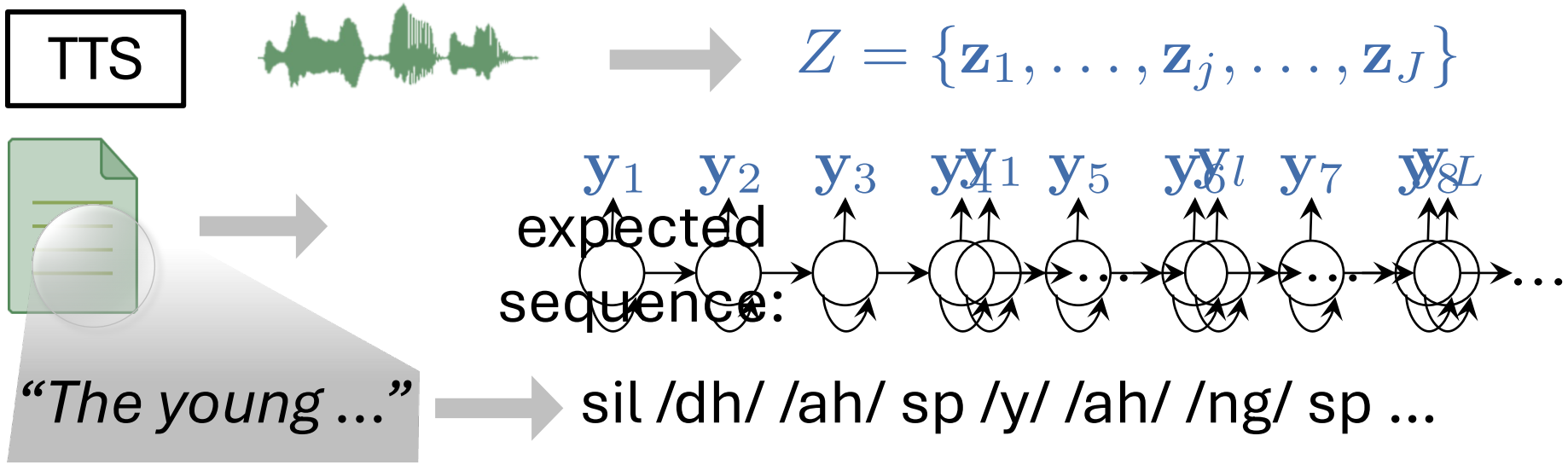


$$\mathrm{Match}(W_k, S)$$

Subjective intelligibility assessment (1)

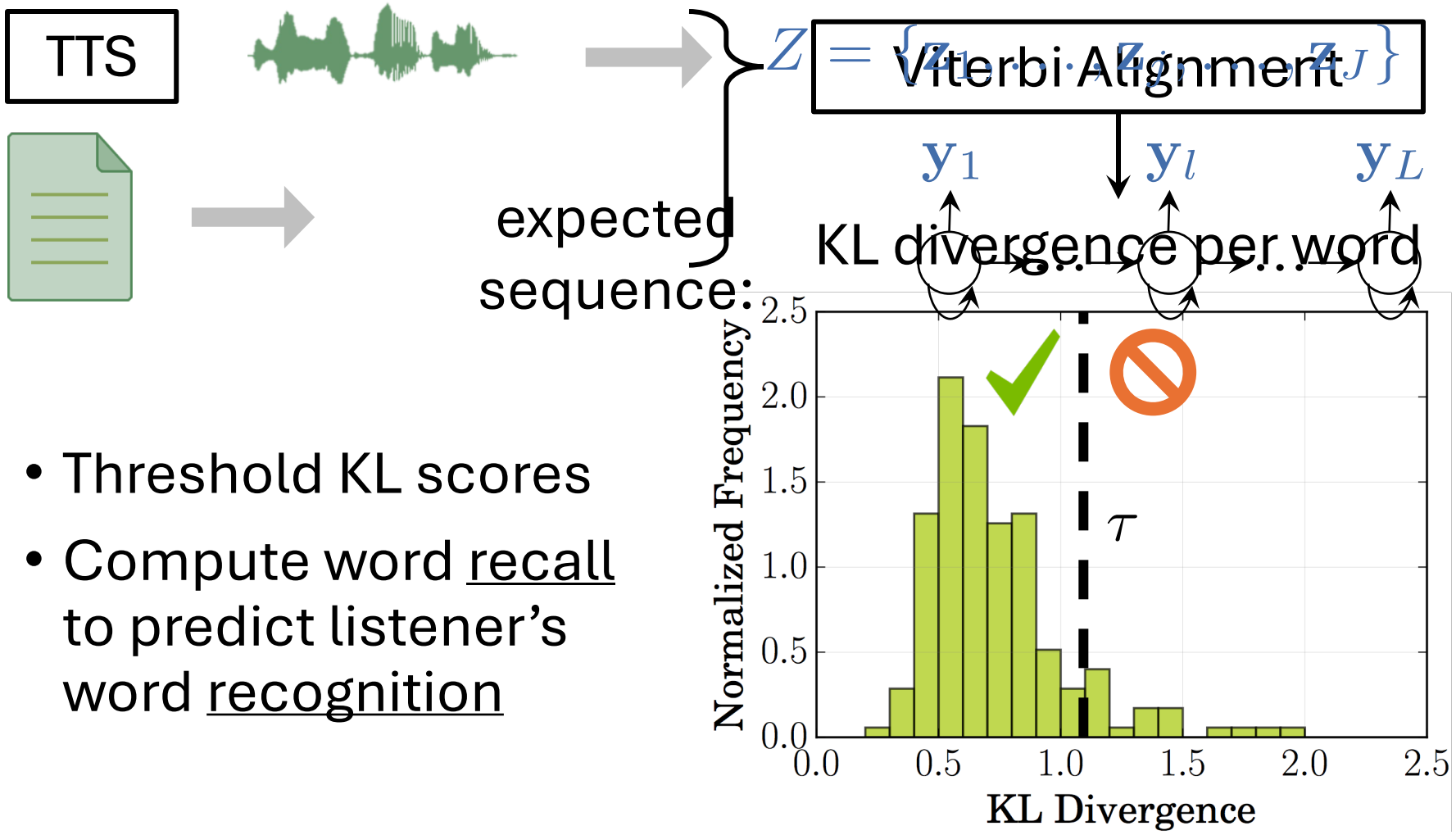


Match word hypothesis with synthetic speech



- The phonetic transcription is obtained from a lexicon
- HMM states are parameterized by distributions $\mathbf{y}_l$, with the KL divergence $\mathrm{KL}\left(\mathbf{y}_l, \mathbf{z}_j\right)$ as local score

Hypothesis testing and word recall

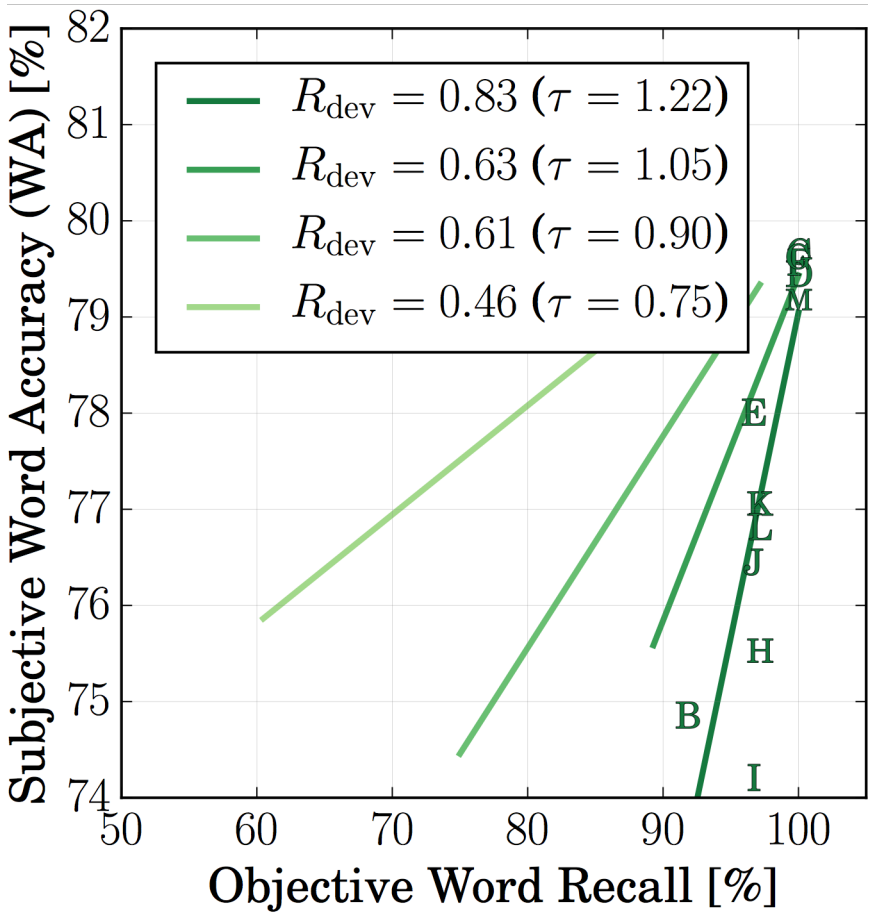


- Threshold KL scores
- Compute word recall to predict listener's word recognition

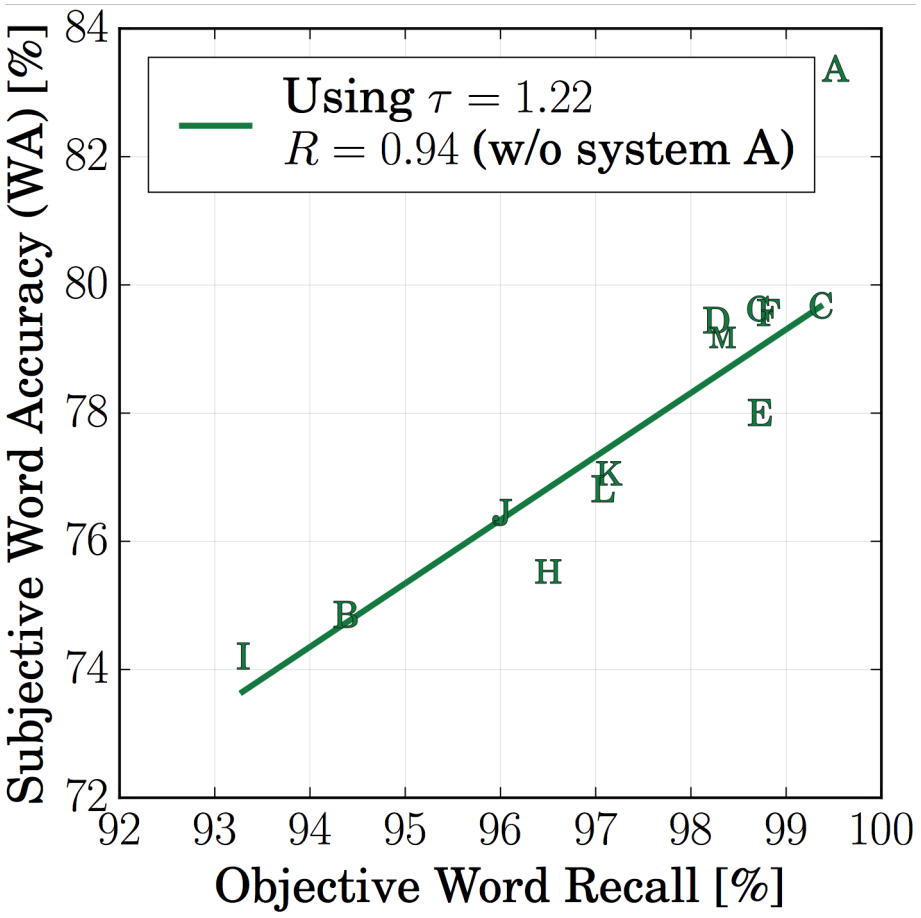
Evaluation

2011 Blizzard Challenge

- 12 TTS systems named “B” to “M”
- 26 semantically unpredictable sentences (SUS), scored by 231 listeners (26 scores per listener)
- “System A”: natural speech from a voice talent



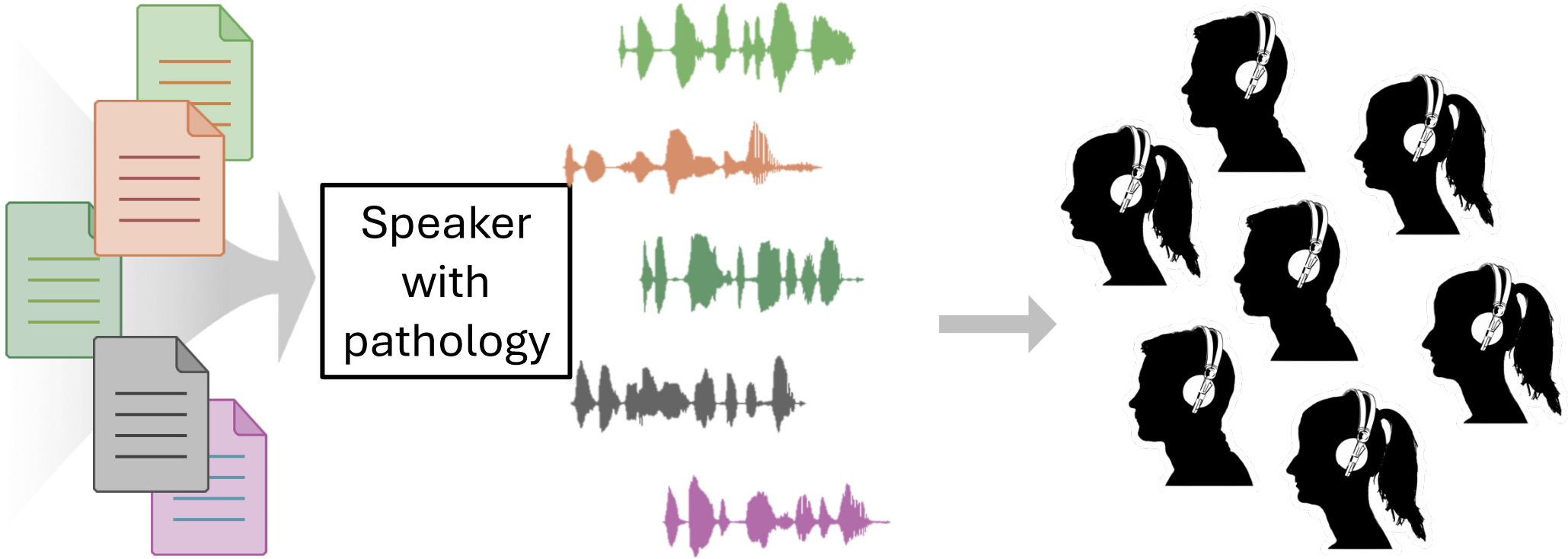
Determine threshold on development data



Performance on evaluation data

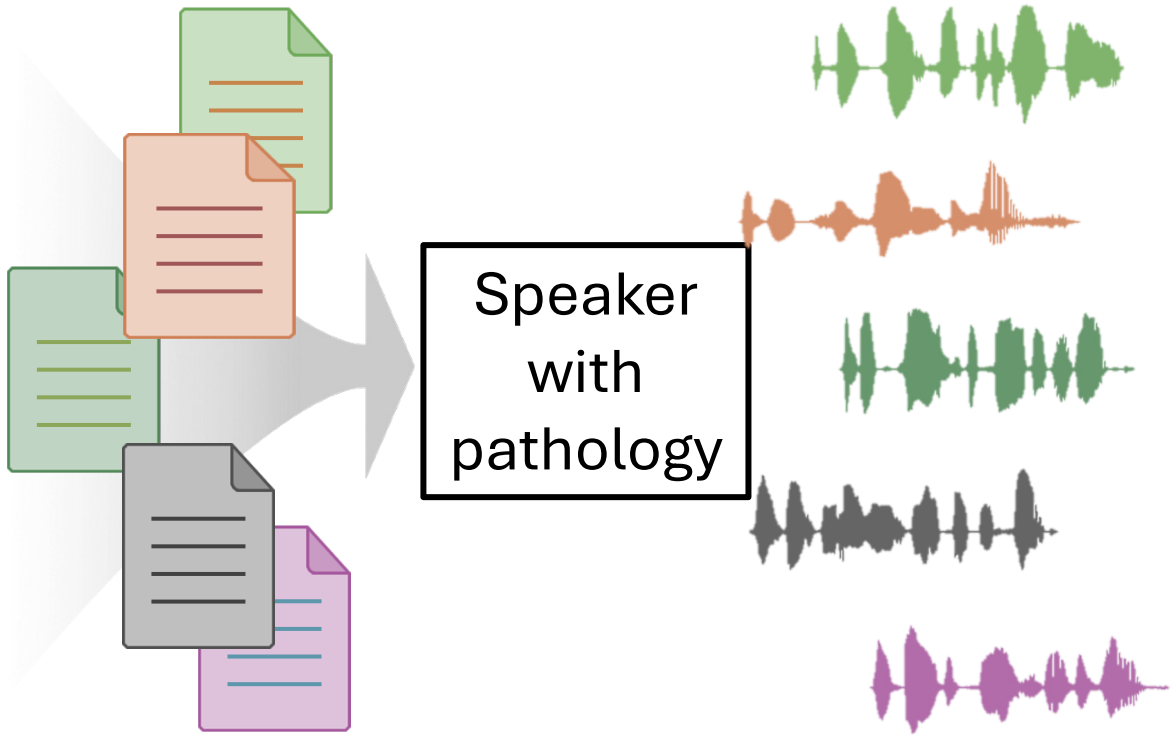
Objective Intelligibility Assessment of Text-to-Speech Systems Through Utterance Verification, Raphael Ullmann, Ramya Rasipuram, Mathew Magimai-Doss and Hervé Bourlard, in: Proceedings of Interspeech, Dresden, Germany, pages 3501-3505, 2015

Subjective intelligibility assessment (2)



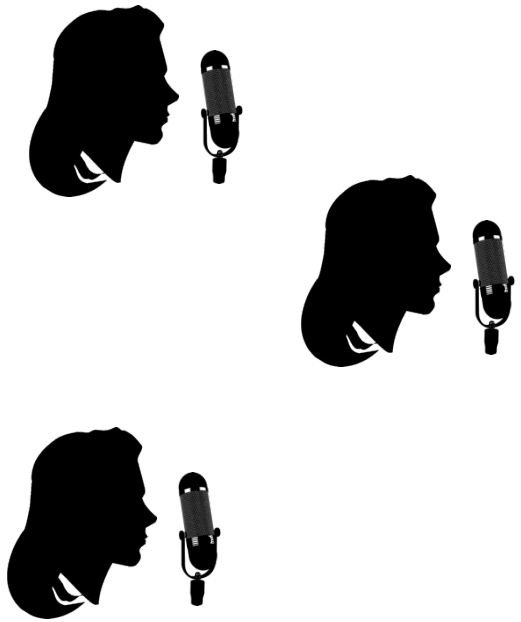
e.g., speaker with dysarthria

Objective intelligibility assessment (1)

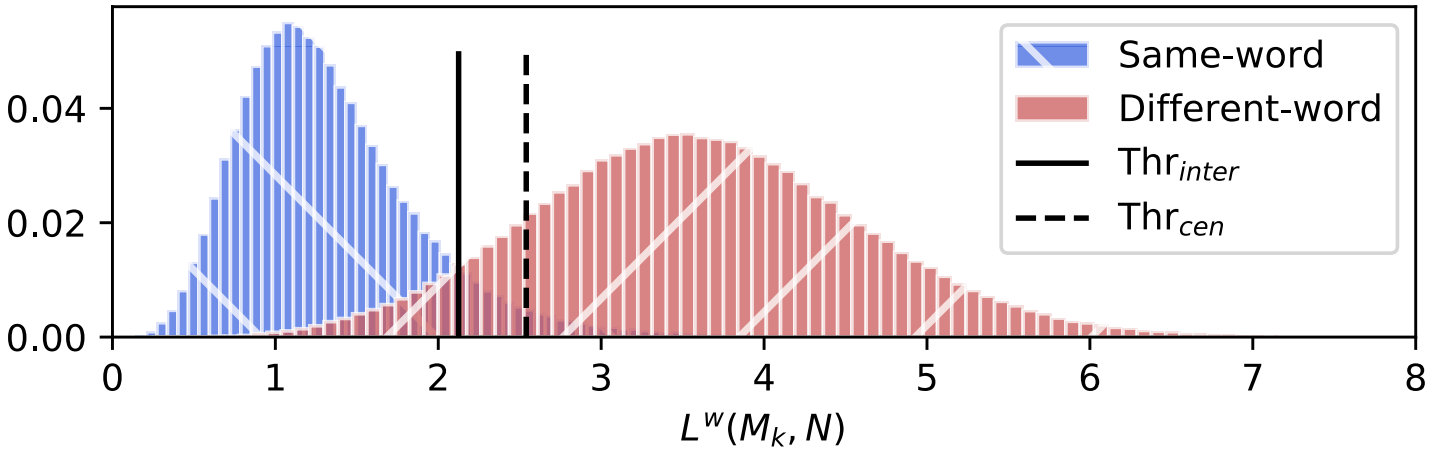


e.g., speaker with dysarthria

Replace listeners by healthy control speakers' speech utterances of each word or even synthetic speech.



Objective intelligibility assessment (2)

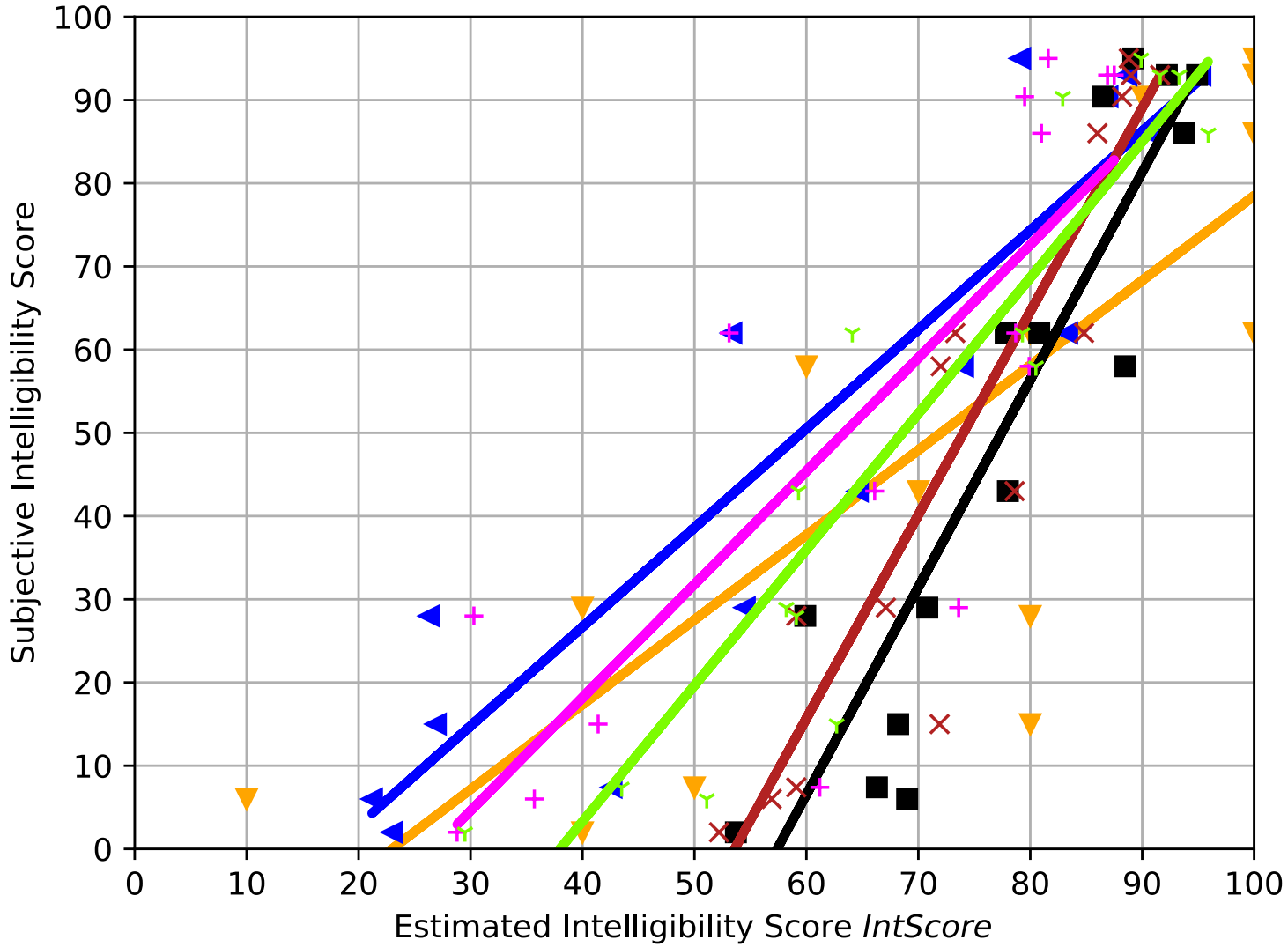


Determing threshold from healthy control speakers' data by forming (a) same-word pair and (b) different word pair.

For each pathological speech utterance

1. Performance utterance by matching to the respective word utterances from healthy control speakers
2. Count number of same word decisions
3. Count more than or equal to half of number of healthy control speakers then decide the pathological speech utterance is recognizable else non-recognizable
4. Count the number of recognizable words

Objective intelligibility assessment (3)



- Phone Thr_{inter} (R=.950, RMSE=21.1)
- AF_{binary} Thr_{inter} (R=.922, RMSE=16.9)
- $AF_{multi-manner}$ Thr_{inter} (R=.917, RMSE=35.2)
- Synth. Phone Thr_{inter} (R=.931, RMSE=32.3)
- Synth. AF_{binary} Thr_{inter} (R=.822, RMSE=24.4)
- Synth. $AF_{multi-manner}$ Thr_{inter} (R=.930, RMSE=25.1)

Experiments conducted on UA-Speech dataset

Utterance Verification-based Dysarthric Speech Intelligibility Assessment using Phonetic Posterior Features, Julian Fritsch and Mathew Magimai-Doss, in: IEEE Signal Processing Letters, 28:224 - 228, 2021

Posterior feature based approach

A modular and flexible framework that enables abstraction and integrated modeling of communication phenomenon

- **Q1** – Flexibility to choose and/or explore symbol set
- **Q2** – Flexibility to choose symbol probability estimation method
- **Q3** – Flexibility to text-based and acoustic instance-based representation of word hypothesis
- **Q4** – Intepretable and explainable information-theoretic approach akin to string matching

Systematically assimilate developments in different fields such as, linguistics, signal processing, machine learning, information theory, coding theory, communication theory, and probability and statistics

W_k

Listener side / Observer side / synthesis

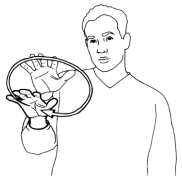
Top-down

$$Y_k = \mathbf{y}_{k,1} \cdots \mathbf{y}_{k,n} \cdots \mathbf{y}_{k,N}$$

$$Z = \mathbf{z}_1 \cdots \mathbf{z}_m \cdots \mathbf{z}_M$$

Speaker side / Signer side / analysis

Bottom-up



S