

2

Quiz

40

The start, middle and end parts of phonemes sound different. How to model a phoneme most effectively?

- ▲ Fit a sequence of states with transitions and outputs
- ◆ Fit a linear trajectory model for the outputs
- Define a set of output symbols and find their probabilities
- Fit a GMM that has a large number of mixtures

Correction answer: A

3

Quiz

86

Which assumptions are used in hidden Markov models?

- ▲ The state observation probability functions depend on time
- ◆ The state transition matrices do not depend on time
- The observations are independent from each other
- The duration of each state is normally distributed

Correction answer: B and C

4
87

Quiz

```
for ($i = 1; $i <= $iter_num; $i++)  
{  
    HERest -H hmmdefs -I rml_train.mlf  
    -S rml_train.scp -s stats  
    monophones  
}
```

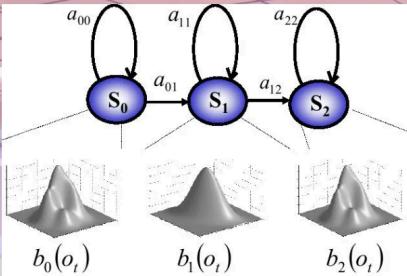
An effective way to estimate HMM parameters for the training samples is to find ...

- ▲ average values over fixed length phoneme segments
- ◆ max probability over all state sequences
- max probability of the best state sequence
- optimize frame classifications over fixed phoneme segments

Correction answer: B and C

5
87

Quiz



An effective way to find the best HMM state sequence for a speech sample is to ...

- ▲ use the Forward algorithm to compute scores of all sequences
- ◆ move to new state when its GMM fits better than the old one
- recursively find the best state for each frame using Viterbi
- compute scores of all sequences in parallel

Correction answer: C

1
87

- Let's assume words can be represented by a sequence of states, S,

$$\begin{aligned}\hat{W} &= \arg \max_W P(O | W)P(W) \\ &= \arg \max_W \sum_S P(O | S)P(S | W)P(W)\end{aligned}$$

- Words → Phonemes → States
- States represent smaller pieces of phonemes

Which claim is NOT TRUE about minimizing the classification error rate?

▲ It minimizes the posterior prob. $P(W|O)$

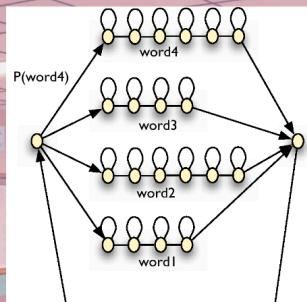
◆ It optimizes over all word strings

● Observation prob. $P(O|S)$ can be computed by a GMM

■ Language model prob. $P(W)$ can be computed by n-grams

Correct answer: A

2
88



In continuous speech ASR, which states are possible at time t=8?

▲ Only the 2nd and 4th state of each word

◆ Only the 1st state of each word

● Only the states of word₁

■ Any state is possible

Correct answer:

D

3 Quiz

Exercise1: Token passing

What is NOT TRUE about the last empty box?

Answers:

- ▲ It has 1 token coming from state S_0 and 2 from state S_1
- ◆ The state at $t=0$ has no more effect in red 3-gram "state-LM"
- All token probabilities are lower than 0.001 (= 1 E-3)
- All tokens can now be merged and only the best goes to $t=4$

Correct answer:

D

4 Quiz

**1. Pass tokens from all states to all successor states
2. Update the tokens for each state
3. Prune unlikely tokens
4. Merge tokens with the same N last words**

What is NOT TRUE about the Token Passing Decoder (steps of the algorithm below)?

Answers:

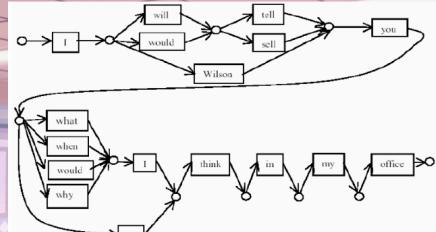
- ▲ Without pruning & merging the #tokens grows linearly in time
- ◆ The update needs transition, observation and LM probability
- Pruning means dropping the lowest probability tokens
- Tokens with identical n-gram history can be merged

Correct answer:

A

87

Quiz



-What is not true about rescoring a lattice or n-best list?

▲ The rescoring pass is slower than the first decoding pass

◆ You can use LMs that take the whole sentence as input

● The words pruned out in first pass can not be rescored in 2nd pass

■ Lattice is a compact way to represent a large n-best list

Correct answer:

A

Xuan Binh



2559 points

On your way to the podium!

1

Quiz

57

Which statement about end-to-end ASR is True?

- ▲ It can only be trained with a small amount of data
- ◆ It does not suffer from overfitting
- It eliminates the need for neural networks
- It directly maps the input audio signals to transcriptions

Correct answer:

D

2

Quiz

58

Which statement about Transformer is True?

- ▲ It does not require pre-training to achieve high accuracy
- ◆ It was developed primarily for computer vision tasks
- It is only effective for small-scale datasets
- It relies on multi-head attention to accumulate past information

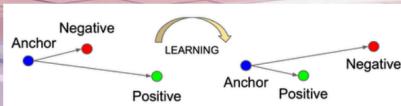
Correct answer:

D

3

58

Quiz



Which statement about Self-supervised learning (SSL) is **False**?

- ▲ SSL models learn from the data itself without explicit supervision
- ◆ Contrastive learning aims to separate positive and negative samples
- SSL has been successfully applied in NLP and speech recognition
- SSL always requires a large amount of labelled data for pre-training

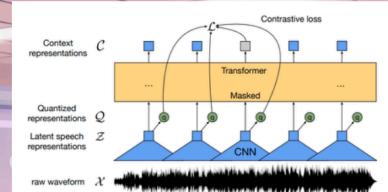
Correct answer:

D

4

57

Quiz



Which statement about the Wav2Vec ASR model is **False**?

- ▲ It uses the contrastive learning objective during pretraining
- ◆ It is designed for supervised learning and does not use pre-training
- It is generally finetuned using the CTC loss
- It takes MFCC features as input

Correct answer:

B, D

5

Quiz

56

Which statement about the CTC algorithm is **True**?

It is a supervised learning algorithm that requires aligned training data

It introduces a blank symbol in its output space

It is primarily used for computer vision tasks

It aims to align audio with transcriptions without explicit alignment

Correct answer:

B, D