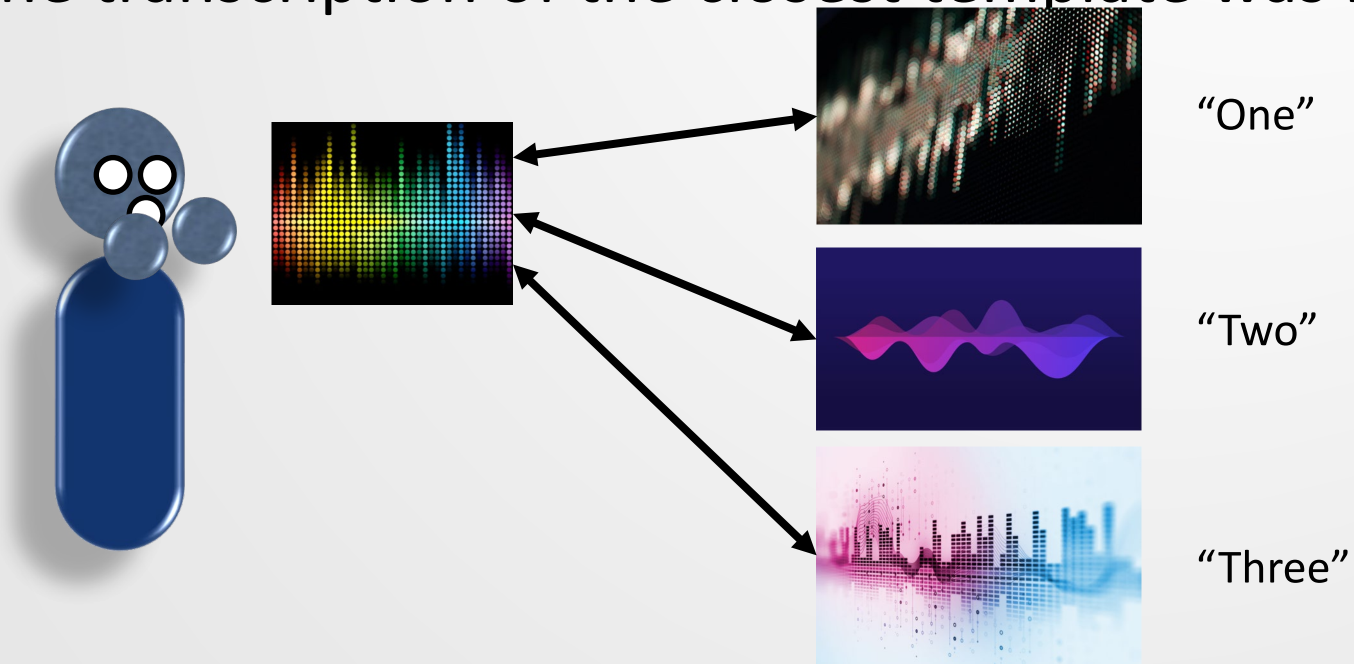
Template matching in ASR

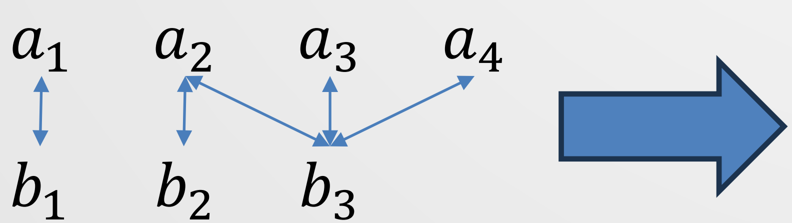
- Early ASR systems were based on template matching
- Input utterances were compared against stored templates
- The transcription of the closest template was returned



- How to deal with stretching and shrinking in time?

Aligning features

- Let $a \in \mathbb{R}^{U \times D}$ and $b \in \mathbb{R}^{T \times D}$ be seqs. of speech features
  - MFCCs, f-bank feats, *etc*.
- Choose some frame-wise (vector) distance function
  - E.g. $d(a_u, b_t) = \sum_{d=1}^{D} \left| a_{u,d} - b_{t,d} \right|$
- We can sum those frame-wise distances in a monotonic alignment to get a "distance" between utterances!



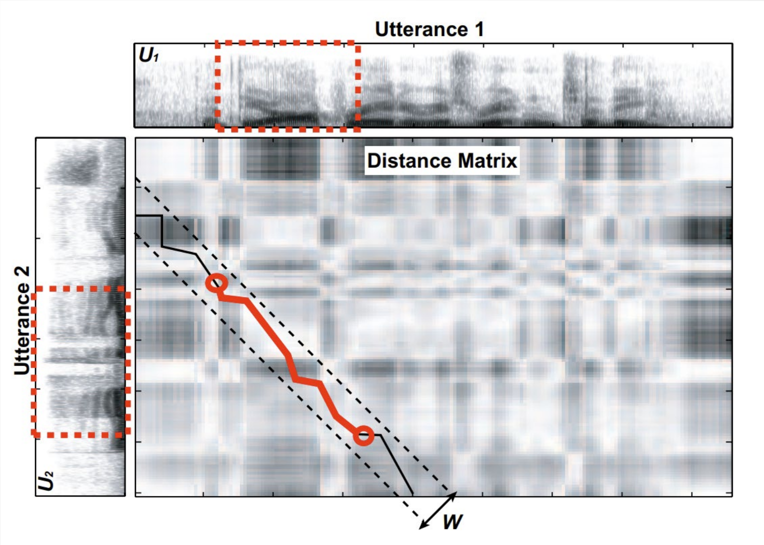
$$d(a_1, b_1) + d(a_2, b_2) + d(a_2, b_3) + d(a_3, b_3) + d(a_4, b_3)$$

Dynamic time warping

- There are many possible ways to align a and b (call them $\mathcal{A}$)
- In **Dynamic Time Warping** (DTW), *the* "distance" between a and b is that with the minimal frame-wise sum

$$\mathcal{D}_d(a,b) = \min_{\alpha \in \mathcal{A}} \sum_{\{u,t\} \in \alpha} d(a_u, b_t)$$

- $\mathcal{D}_d(a,b)$ can be computed with monotonic_forward



From Park and Glass (2006) "Towards unsupervised pattern discovery in speech"

Specifying DTW

1. What are a_u and b_t? → Feature vectors of utts a and b
2. What is a solution to a prefix?

   $\rightarrow table[u,t] = \mathcal{D}_d\left(a_{1,\ldots,u}, b_{1,\ldots,t}\right)$
3. How do we build the initial prefix(es)?

   $\rightarrow table[1,1] = d(a_1, b_1)$
4. How do we extend a prefix correctly?

$$\rightarrow table[u,t] = d(a_u, b_t) + \min \begin{cases} table[u-1,t] \\ table[u-1,t-1] \\ table[u,t-1] \end{cases}$$

5. How is the result computed from $table$?

   $\rightarrow \mathcal{D}_d(a,b) = table[U,T]$

UNIVERSITY OF TORONTO

Adapting monotonic_forward

Function monotonic_forward
Inputs $a = a_1, a_2, \ldots, a_U$ **and** $b = b_1, b_2, \ldots, b_T$
Define $table[0 \ldots U, 0 \ldots T]$
initialize$(table[0 \ldots U, 0], table[0,1 \ldots T])$
For each u **in** $1 \ldots U$:
    For each t **in** $1 \ldots T$:
        $table[u,t] =$
            $\mathrm{step}(a_u, b_t, table[u-1, t-1], table[u-1, t], table[u, t-1])$
Return $\mathrm{finalize}(table[0 \ldots U, 0 \ldots T])$

initialize: set $table[0,0] = 0$, $table[1 \ldots U, 0] = table[0,1 \ldots T] = \infty$

step: $table[u,t] = d(a_u, b_t) + \min \begin{cases} table[u-1, t] \\ table[u-1, t-1] \\ table[u, t-1] \end{cases}$

finalize: $table[U, T]$

A DTW example

		b_1	b_2	b_3
a_1		$??+0$	$??+3$	$??+1$
a_2		$??+1$	$??+2$	$??+5$
a_3		$??+1$	$??+2$	$??+4$
a_4		$??+1$	$??+0$	$??+1$

Since $d(a_u, b_t)$ is a fixed cost in $table[u, t]$, we denote it as "+ x"

A DTW example

		b_1	b_2	b_3
	0	∞	∞	∞
a_1	∞	$??+0$	$??+3$	$??+1$
a_2	∞	$??+1$	$??+2$	$??+5$
a_3	∞	$??+1$	$??+2$	$??+4$
a_4	∞	$??+1$	$??+0$	$??+1$

Initialize row 0 and column 0

A DTW example

		b_1	b_2	b_3
	0	∞	∞	∞
a_1	∞	$0+0$	$0+3$	$3+1$
a_2	∞	$??+1$	$??+2$	$??+5$
a_3	∞	$??+1$	$??+2$	$??+4$
a_4	∞	$??+1$	$??+0$	$??+1$

Fill row 1

A DTW example

		b_1	b_2	b_3
	0	∞	∞	∞
a_1	∞	$0+0$	$0+3$	$3+1$
a_2	∞	$0+1$	$0+2$	$2+5$
a_3	∞	$??+1$	$??+2$	$??+4$
a_4	∞	$??+1$	$??+0$	$??+1$

Fill row 2

A DTW example

		b_1	b_2	b_3
	0	∞	∞	∞
a_1	∞	$0+0$	$0+3$	$3+1$
a_2	∞	$0+1$	$0+2$	$2+5$
a_3	∞	$1+1$	$1+2$	$2+4$
a_4	∞	$2+1$	$2+0$	$2+1$

… and so forth

A DTW example

		b_1	b_2	b_3
	0	∞	∞	∞
a_1	∞	0	3	4
a_2	∞	1	2	7
a_3	∞	2	3	6
a_4	∞	3	2	$\boxed{3}$

Return $table[4,3] = 3$

ASR evaluation

- ASR systems often generate **hypothesis** transcripts which don't match the gold-standard **reference** transcript
- But some are closer than others:
  - Reference: `errors are common here`
  - Hypothesis 1: `his errors are commas here`
  - Hypothesis 2: `here are are`
- We may describe the errors in terms of transformations:
  - Hypothesis 1 **inserted** `his` and **substituted** `commas` for `common`
  - Hypothesis 2 **substituted** `here` for `errors`, `are` for `common`, and **deleted** `here`
    - The `here`s cannot match without re-ordering

Word-error rate (WER)

- ASR enthusiasts are often concerned with **word-error rate (WER)**, which counts the **minimum** number of 3 types of errors a hypothesis makes given a reference
  - **Substitution error**: One word being mistook for another
    e.g., hyp: `shift`, ref: `ship`
  - **Deletion error**: An input word that is 'skipped'
    e.g., hyp: `I Torgo`, ref: `I am Torgo`
  - **Insertion error**: A 'hallucinated' word not in the input.
    e.g., hyp: `steamed hams`, ref: `hams`
- Given S substitutions, D deletions, I insertions, and N reference tokens:

$$WER = \frac{S + D + I}{N} \times 100\%$$

(this is not a valid percentage)

Aligning strings

- We can count errors by building up prefixes of alignments between reference (a) and hypothesis (b)
  - Insertion: add b_t to hypothesis
  - Deletion: add a_u to reference
  - Match/substitution: add b_t, a_u
- Prepend ref + hyp with a special token <s> to allow for insertions/deletions at the beginning of the reference/hypothesis

```
<s>     errors are common here          <s> errors are common here

<s> his errors are comma  here          <s> here   are are
```

Specifying WER

1. What are a_u and b_t? → Reference and hypothesis tokens
2. What is a solution to a prefix?

   $$\rightarrow table[u,t] = \min \#errors(\text{<s>}a_{1,\ldots,u}, \text{<s>}b_{1,\ldots,t})$$
3. How do we build the initial prefix(es)?

   $$\rightarrow table[u,0] = u, table[0,t] = t$$
4. How do we extend a prefix correctly?

   $$\rightarrow table[u,t] = \min \begin{cases} table[u-1,t]+1 \\ table[u-1,t-1] + \begin{cases} 0 & a_u = b_t \\ 1 & a_u \neq b_t \end{cases} \\ table[u,t-1]+1 \end{cases}$$
5. How is the result computed from $table$?

   $$\rightarrow WER(a,b) = \frac{table[U,T]}{U} \times 100\%$$

UNIVERSITY OF TORONTO

Adapting monotonic_forward

```
Function monotonic_forward
Inputs a = a_1, a_2, ..., a_U and b = b_1, b_2, ..., b_T
1: Define table[0 ... U, 0 ... T]
2: initialize(table[0 ... U, 0], table[0,1 ... T])
3: For each u in 1 ... U :
4:     For each t in 1 ... T :
5:         table[u, t] =
               step(a_u, b_t, table[u − 1, t − 1], table[u − 1, t], table[u, t − 1])
6: Return finalize(table[0 ... U, 0 ... T])
```

initialize: set $table[u, 0] = u$, $table[0, t] = t$

step: $table[u, t] = \min \begin{cases} table[u-1, t] + 1 \\ table[u-1, t-1] + \begin{cases} 0 & a_u = b_t \\ 1 & a_u \neq b_t \end{cases} \\ table[u, t-1] + 1 \end{cases}$

finalize: $\frac{table[U,T]}{U} \times 100\%$

A WER example

	<s>	his	errors	are	comma	here
<s>						
errors						
are						
common						
here						

Rows (a_u) are reference tokens, columns (b_t) are hypothesis tokens

UNIVERSITY OF TORONTO

A WER example

	<s>	his	errors	are	comma	here
<s>	0	1	2	3	4	5
errors	1					
are	2					
common	3					
here	4					

- Initialize row 0 and column 0
  - Row 0 inserts hypothesis tokens
  - Column 0 deletes references tokens

A WER example

	<s>	his	errors	are	comma	here
<s>	0	1	2	3	4	5
errors	1	1	1	2	3	4
are	2					
common	3					
here	4					

- Fill row 1
  - Note: `errors` match

A WER example

	<s>	his	errors	are	comma	here
<s>	0	1	2	3	4	5
errors	1	1	1	2	3	4
are	2	2	2	1	2	3
common	3					
here	4					

- Fill row 2
  - Note: `are` match
  - Note: this priority goes: match > sub > ins > del

UNIVERSITY OF TORONTO

A WER example

	<s>	his	errors	are	comma	here
<s>	0	1	2	3	4	5
errors	1	1	1	2	3	4
are	2	2	2	1	2	3
common	3	3	3	2	2	3
here	4	4	4	3	3	2

- …and so on

A WER example

	<s>	his	errors	are	comma	here
<s>	0	1	2	3	4	5
errors	1	1	1	2	3	4
are	2	2	2	1	2	3
common	3	3	3	2	2	3
here	4	4	4	3	3	2

- Return $\frac{table[4,5]}{4} \times 100\% = 50\%$

Returning to ASR

- Recall: each frame gets a label by repeating transcript tokens

/ow ow ow ow ow p p p p ah ah ah ah ah n/

$\Downarrow$

/ow p ah n /

- Let $a_{1,\ldots,U}$ be the **reference transcript**, $b_{1,\ldots,T}$ be **a** frame-wise transcript with repetitions of a
- It is easy to train an RNN or Transformer to maximize the likelihood of b

$$\mathcal{L} = -\log P_\theta(b)$$

- But there are many choices of b!

Marginalization

- Solution: maximize the likelihood of all such b!
- Let $\mathcal{B}$ be a function which **removes sequential repetitions** from b: $\mathcal{B}(\mathrm{A\ B\ B\ A}) = \mathrm{A\ B\ A}$
- Let $\mathcal{B}^{-1}(a;T)$ be all b of length T which reduce to a

$$\mathcal{B}^{-1}(a;T) = \{b_{1,\ldots,T} : \mathcal{B}(b) = a\}$$

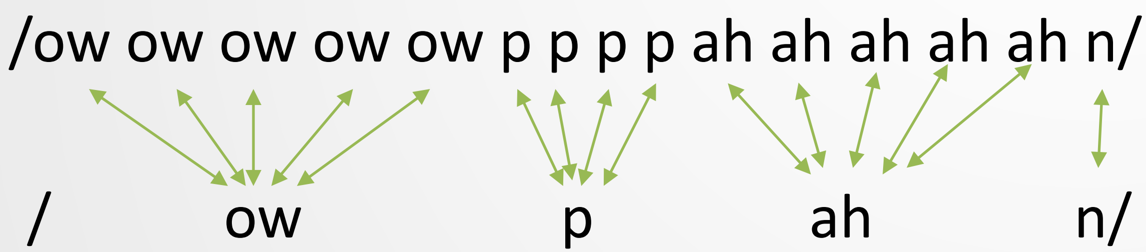
- Then we minimize

$$\mathcal{L} = -\log P_\theta(a) = -\log \sum_{b \in \mathcal{B}^{-1}(a;T)} P_\theta(b)$$

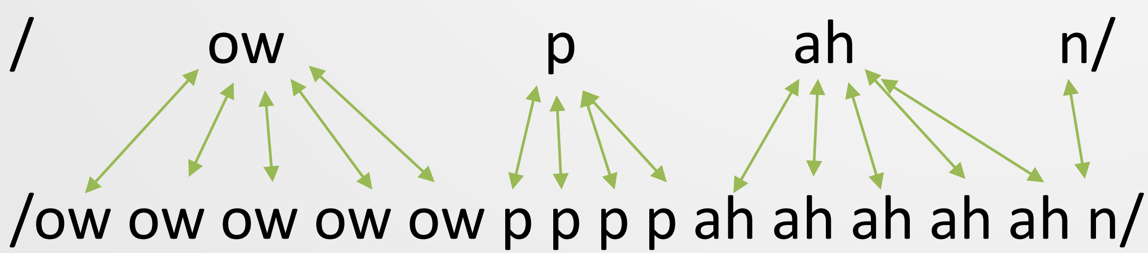
- However, there are $\binom{T-1}{U-1}$ paths in $\mathcal{B}^{-1}(a;T)$
- Can we use monotonic_forward?

Monotonic alignments

- If $\mathcal{B}(b) = a$, a monotonic alignment exists between (a, b)



- But **not all** monotonic alignments (a, b) imply $\mathcal{B}(b) = a$



- We can fix this by disallowing “down” (adding a_u to a prefix)

Partial solutions

- For DTW and WER, $table[u, t]$ keeps track of **one** optimal alignment per prefix pair $a_{1,\dots,u}, b_{1,\dots,t}$

$$table[u, t] = \min \#\text{errors}\left(a_{1,\dots,u}, b_{1,\dots,t}\right)$$

- For ASR, $table[u, t]$ keeps track of **all** alignments per prefix pair $a_{1,\dots,u}, b_{1,\dots,t}$

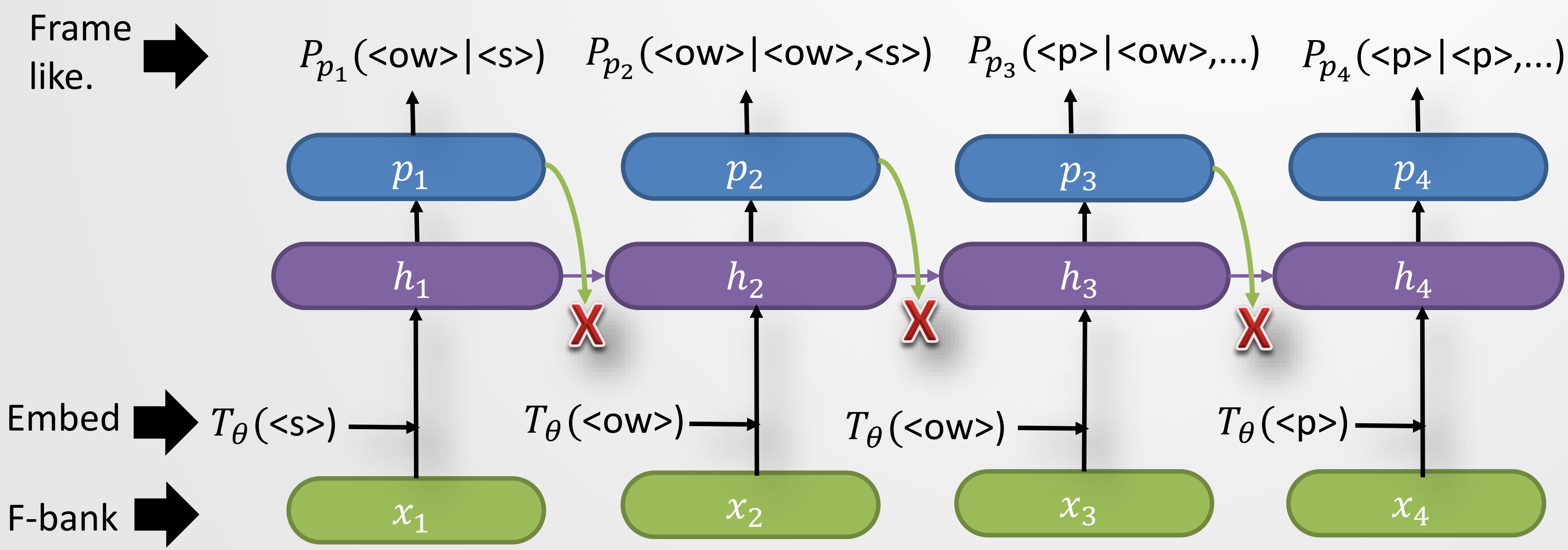
$$table[u, t] = \sum_{b_{1,\dots,t} \in \mathcal{B}^{-1}(a_{1,\dots,u};\, t)} P_\theta\left(b_{1,\dots,t} | x\right)$$

- Then $\mathcal{L} = -\log table[U, T]$

UNIVERSITY OF TORONTO

An auto-regressive approach?

- Suppose we use an auto-regressive RNN to generate frame-level predictions $P_\theta\left(b_t \middle| b_{1,\ldots,t-1}, x\right)$



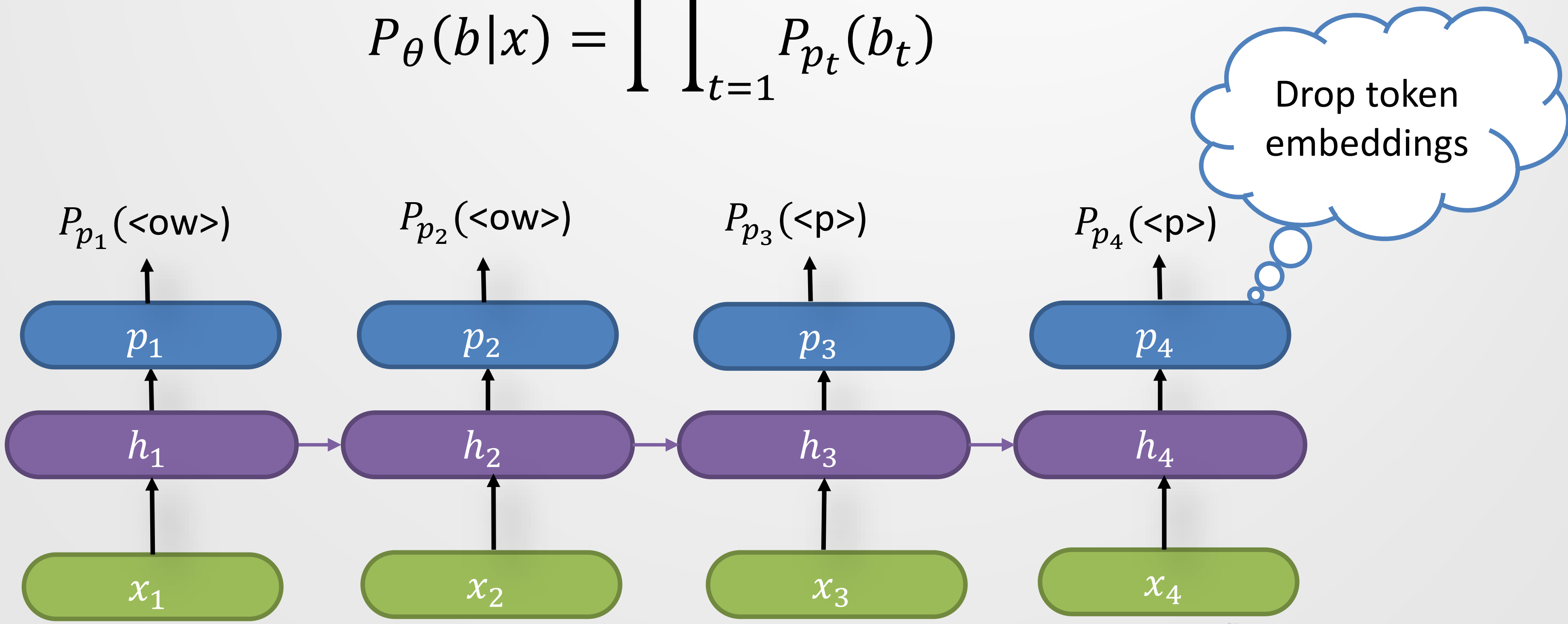
- Can we use this with monotonic_forward?

UNIVERSITY OF TORONTO

Conditional independence

- **No.** The frame-wise likelihoods depend on a specific prefix
  - $P_\theta(\text{<ow>}|\text{<ow>},\text{<s>}) \neq P_\theta(\text{<ow>}|\text{<p>},\text{<s>})$
  - We cannot share computations over prefixes
- We require elements of b to be **conditionally independent** given x:

$$P_\theta(b|x) = \prod_{t=1}^{T} P_{p_t}(b_t)$$



Extending prefixes

- For $b_{1,\dots,t} \in \mathcal{B}^{-1}\left(a_{1,\dots,u}; t\right)$, let
  - $b_{1,\dots,t}^{u-1}$ denote a prefix where a_u is first aligned to b_t
    - a_u aligns to b_t **and** a_{u-1} aligns to b_{t-1}
  - $b_{1,\dots,t}^{u}$ denote a prefix where a_u has already been aligned
    - a_u aligns to b_t **and** a_u aligns to b_{t-1}
- Using conditional independence, we have

$$
\begin{aligned}
table[u,t] &= \textstyle\sum_{b_{1,\dots,t} \in \mathcal{B}^{-1}(a_{1,\dots,u};\, t)} P_\theta\left(b_{1,\dots,t} | x\right) \\
&= P_{p_t}(b_t = a_u) \textstyle\sum_{b_{1,\dots,t} \in \mathcal{B}^{-1}(a_{1,\dots,u};\, t)} P_\theta\left(b_{1,\dots,t-1} | x\right) \\
&= P_{p_t}(b_t = a_u) \left(\textstyle\sum_{b_{1,\dots,t}^{u-1}} P_\theta\Big(\underbrace{b_{1,\dots,t-1}^{u-1}}_{\in \mathcal{B}^{-1}\left(a_{1,\dots,u-1};\, t-1\right)} | x\Big) + \textstyle\sum_{b_{1,\dots,t}^{u}} P_\theta\Big(\underbrace{b_{1,\dots,t-1}^{u}}_{\in \mathcal{B}^{-1}(a_{1,\dots,u};\, t-1)} | x\Big) \right) \\
&= P_{p_t}(b_t = a_u)(table[u-1, t-1] + table[u, t-1])
\end{aligned}
$$

Specifying the ASR loss

1. What are a_u and b_t? → Reference and frame-level tokens
2. What is a solution to a prefix?

   $\rightarrow table[u,t] = \sum_{b_{1,\ldots,t} \in \mathcal{B}^{-1}(a_{1,\ldots,u};\, t)} P_\theta(b_{1,\ldots,t} | x)$
3. How do we build the initial prefix(es)?

   $\rightarrow table[0,0] = 1, table[0,1 \ldots T] = table[1 \ldots U, 0] = 0$
4. How do we extend a prefix correctly?

   $\rightarrow table[u,t] = P_{p_t}(b_t = a_u)(table[u-1, t-1] + table[u, t-1])$
5. How is the result computed from $table$?

   $\rightarrow \mathcal{L} = -\log table[U, T]$

UNIVERSITY OF TORONTO

Adapting monotonic_forward

Function monotonic_forward
Inputs $a = a_1, a_2, \dots, a_U$ **and** $b = b_1, b_2, \dots, b_T$
Define $table[0 \dots U, 0 \dots T]$
initialize($table[0 \dots U, 0], table[0,1 \dots T]$)
For each u **in** $1 \dots U$:
    For each t **in** $1 \dots T$:
        $table[u,t] =$
            $\text{step}(a_u, b_t, table[u-1, t-1], table[u-1, t], table[u, t-1])$
Return finalize($table[0 \dots U, 0 \dots T]$)

initialize: set $table[0,0] = 1$, $table[0,1 \dots T] = table[1 \dots U, 0] = 0$

step: $P_{p_t}(b_t = a_u)(table[u-1, t-1] + table[u, t-1])$

finalize: $-\log table[U,T]$

UNIVERSITY OF TORONTO

An ASR example

		b_1	b_2	b_3	b_4
a_1		$??\times 0.1$	$??\times 0.3$	$??\times 0.1$	$??\times 0.1$
a_2		$??\times 0.1$	$??\times 0.2$	$??\times 0.5$	$??\times 0.1$
a_3		$??\times 0.1$	$??\times 0.2$	$??\times 0.4$	$??\times 0.1$

Since $P_{p_t}(b_t = a_u)$is a fixed cost in $table[u, t]$, we denote it as “x x”

An ASR example

		b_1	b_2	b_3	b_4
	1	0	0	0	0
a_1	0	$??\times 0.1$	$??\times 0.3$	$??\times 0.1$	$??\times 0.1$
a_2	0	$??\times 0.1$	$??\times 0.2$	$??\times 0.5$	$??\times 0.1$
a_3	0	$??\times 0.1$	$??\times 0.2$	$??\times 0.4$	$??\times 0.1$

Initialize the first row and column

An ASR example

		b_1	b_2	b_3	b_4
	1	0	0	0	0
a_1	0	1×0.1	0.1×0.3	0.03×0.1	0.003×0.1
a_2	0	$?? \times 0.1$	$?? \times 0.2$	$?? \times 0.5$	$?? \times 0.1$
a_3	0	$?? \times 0.1$	$?? \times 0.2$	$?? \times 0.4$	$?? \times 0.1$

Fill row 1

An ASR example

		b_1	b_2	b_3	b_4
	1	0	0	0	0
a_1	0	1×0.1	0.1×0.3	0.03×0.1	0.003×0.1
a_2	0	0×0.1	0.1×0.2	0.05×0.5	0.028×0.1
a_3	0	0×0.1	0×0.2	0.02×0.4	0.033×0.1

… and so on

An ASR example

		b_1	b_2	b_3	b_4
	1	0	0	0	0
a_1	0	0.1	0.03	0.003	0.0003
a_2	0	0	0.02	0.025	0.0028
a_3	0	0	0	0.008	$\boxed{0.0033}$

Return $-\log table[U, T] = -\log 0.0033 \approx 5.7$

Inference

- We can compute an error signal for $\mathcal{L} = -\log P_\theta(a|x)$

- At test time, how do we generate transcriptions a?
- Computing $b^* = \text{argmax}_b P_\theta(b|x)$ is **easy**
  - $b^* = \text{argmax}_b \prod_{t=1}^{T} P_{p_t}(b_t) \Rightarrow b_t^* = \text{argmax}_{b_t} P_{p_t}(b_t)$
- Computing $a^* = \text{argmax}_a P_\theta(a|x)$ is **hard**
  - We use DP to compute $P_\theta(a|x)$ **once**
  - For vocab size V, there are $\frac{V(V^T-1)}{V-1}$ possible a
  - Vanilla beam search on b will just return b^*

Prefix search (sketch)

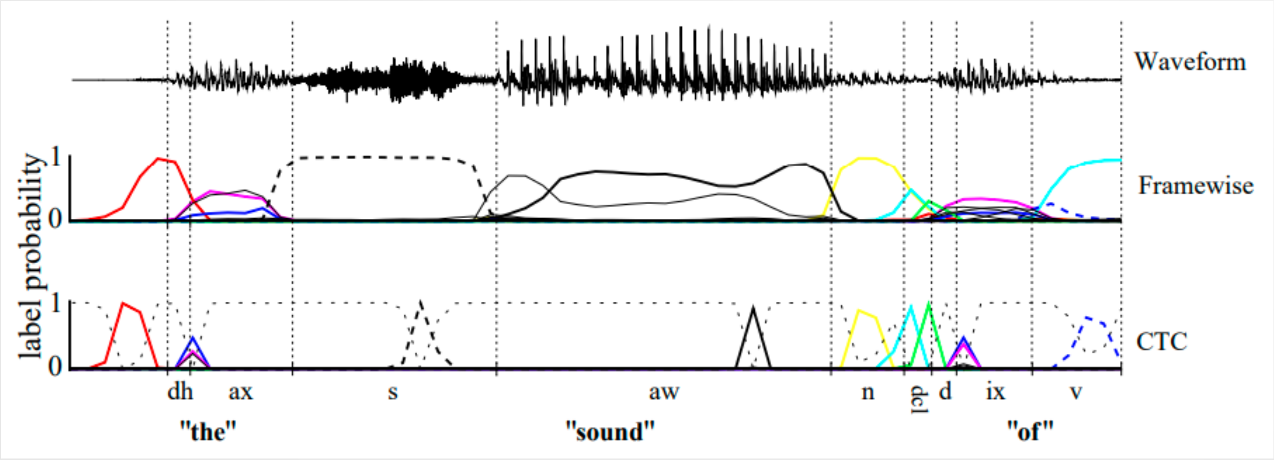
- We use a modified beam search called **prefix search**
- Keeps track of K prefixes of a, **not** b
- For each frame t
  - Extend each prefix $a^{(k)}$ with each element of the vocabulary v and frame-level likelihoods $P_{p_t}(v)$: $a^{(k,v)}$

NEW

  - **Collapse** any repeats in extensions using $\mathcal{B}$, summing the likelihoods of matching prefixes: $\mathcal{B}\left(a^{(k,v)}\right) = \mathcal{B}\left(a^{(k',v')}\right)$
  - Pick the top K most likely extensions as new $a^{(k)}$

Dealing with doubles

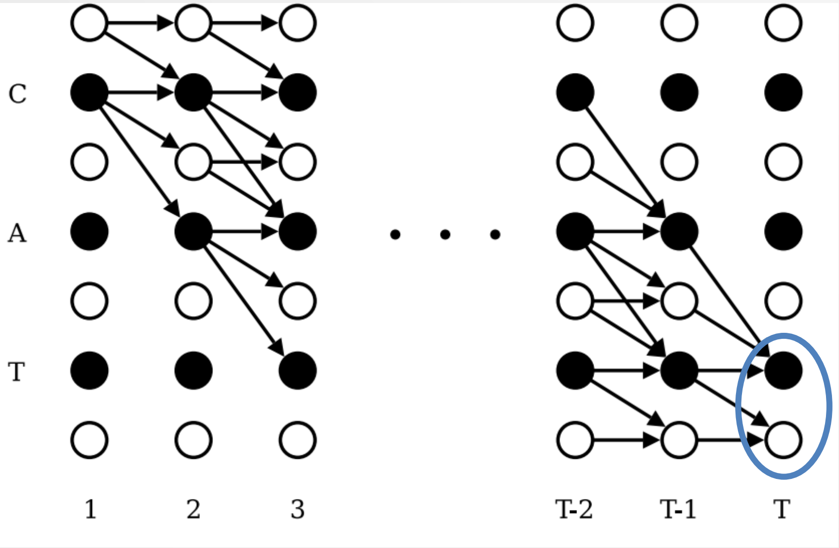
- What happens when a has natural repetitions?
  - E.g. hesitation <I- I am...> = [ey ey ae m...]
  - $\mathcal{B}(a)$ gives us <I am> = [ey ae m]!
- **Connectionist Temporal Classification** (CTC) fixes this by introducing a **blank token** ε to the vocabulary
- $\mathcal{B}$ removes ε **after** removing duplicates
  - $\mathcal{B}$([ey ey ε ε ey ε ae ε ε ε m]) = [ey ey ae m]



From Graves *et al.* (2006) "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks"

CTC and monotonic_forward (sketch)

- monotonic_forward cannot handle blanks **as-is**
  - There are optional ε between each reference token
- Three step solution:
  1. Add blanks between reference tokens
     - $a =$ <C A T> $\mapsto a' =$ <ε C ε A ε T ε>
  2. Add extra diagonal dependency to skip blanks
  3. $\mathcal{L} = -\log\left(table[U-1, T] + table[U, T]\right)$



From Graves *et al.* (2006) “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks”

UNIVERSITY OF TORONTO