

Context Dependent Models

Rita Singh and Bhiksha Raj

Recap and Lookahead

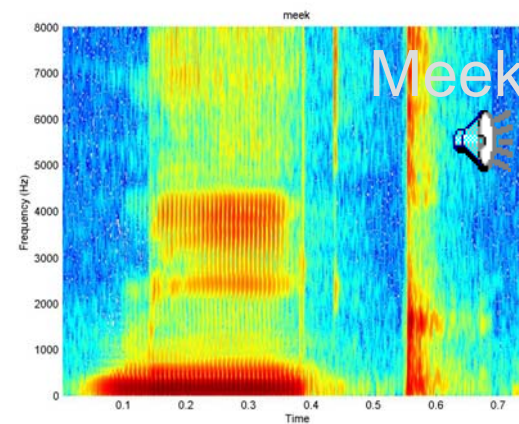
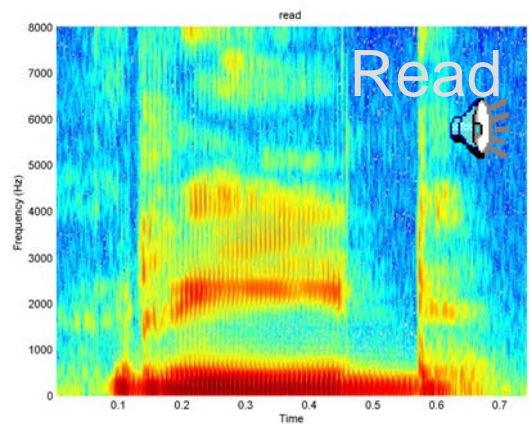
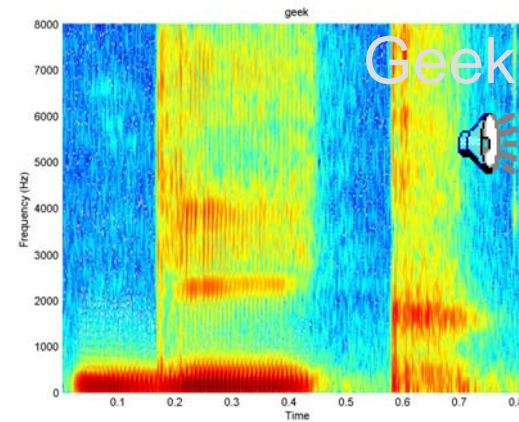
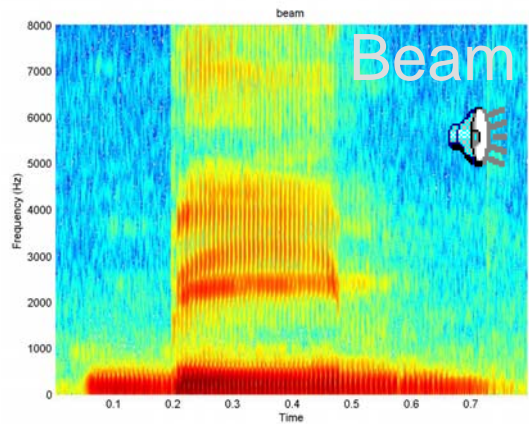
- Covered so far:
 - String Matching based Recognition
 - Introduction to HMMs
 - Recognizing Isolated Words
 - Learning word models from continuous recordings
 - Building word models from phoneme models

 - Exercise: Training phoneme models

- Phonemes in context
 - Better characterizations of fundamental sound patterns

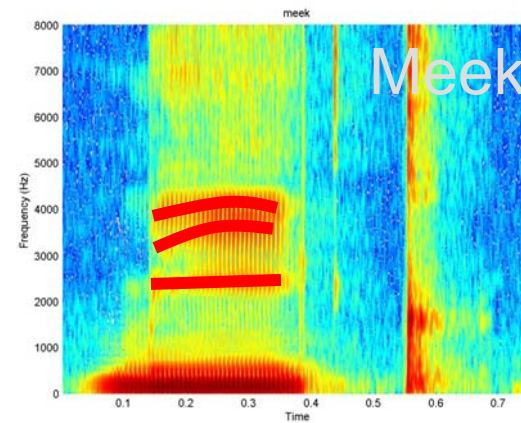
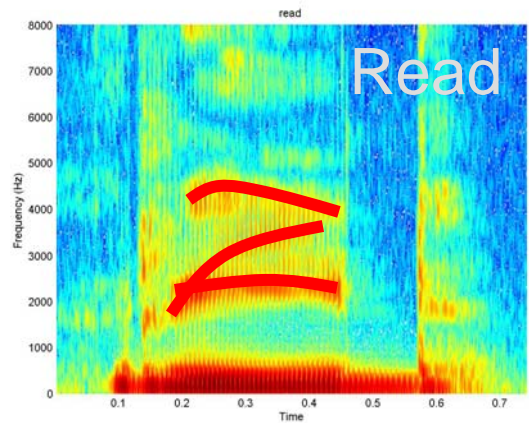
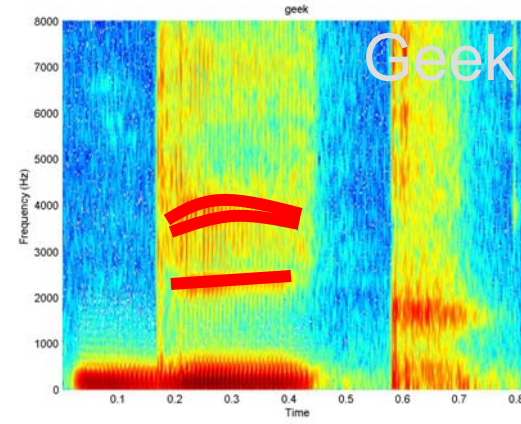
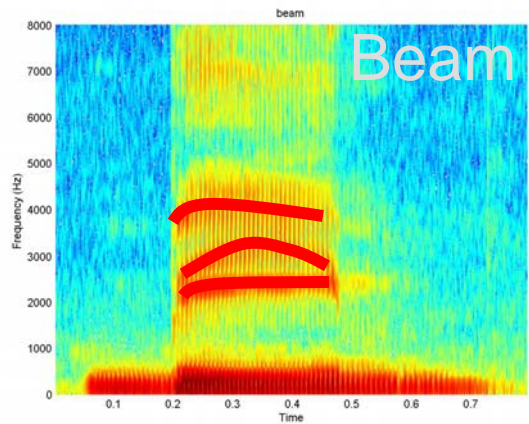
The Effect of Context

- Even phonemes are not entirely consistent
 - Different instances of a phoneme will differ according to its neighbours
 - E.g: Spectrograms of /i/ in different contexts



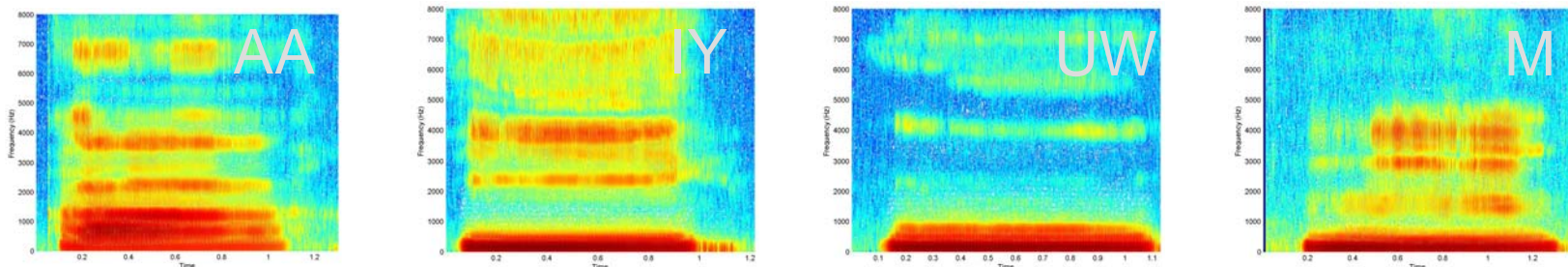
The Effect of Context

- Even phonemes are not entirely consistent
 - Different instances of a phoneme will differ according to its neighbours
 - E.g: Spectrograms of /i:/ in different contexts

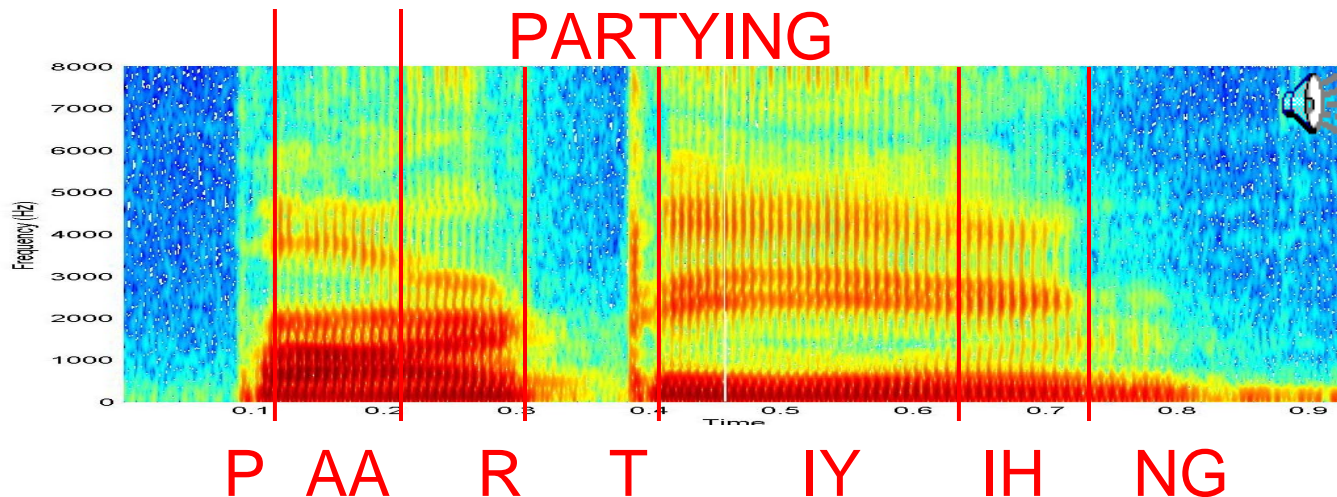


The Effect of Context

- Every phoneme has a locus
 - The spectral shape that would be observed if the phoneme were uttered in isolation, for a long time

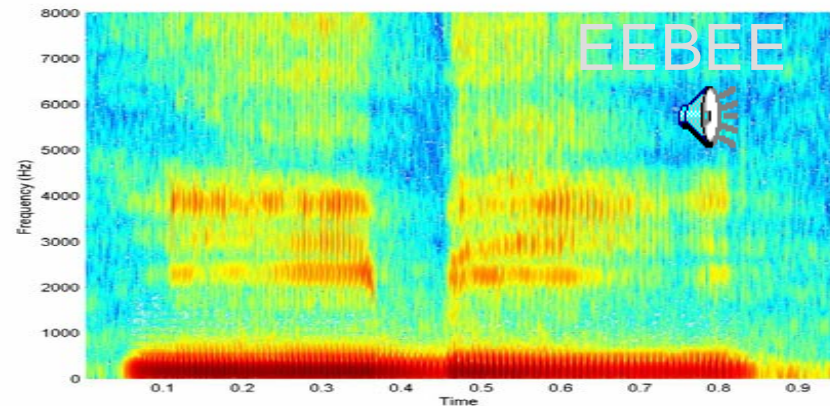
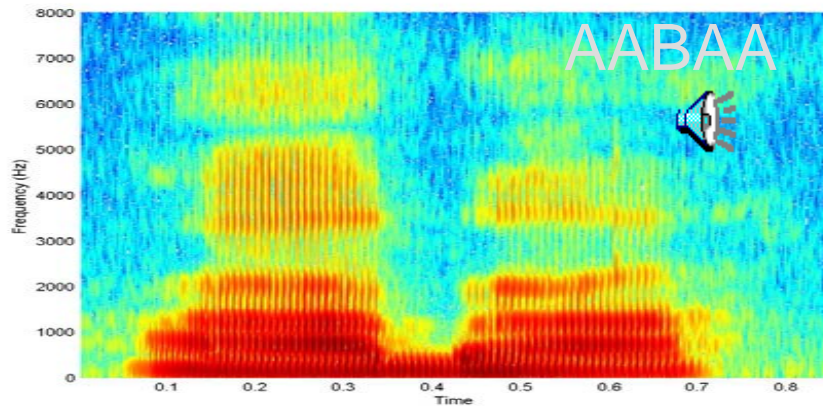


- In continuous speech, the spectrum attempts to arrive at locus of the current sound



The Effect of Context

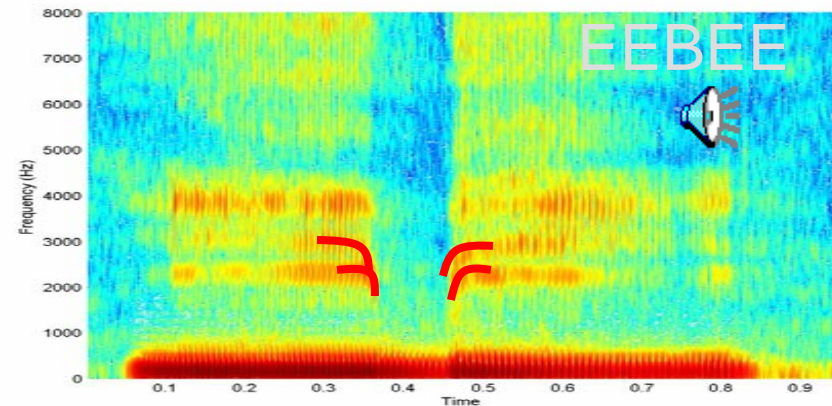
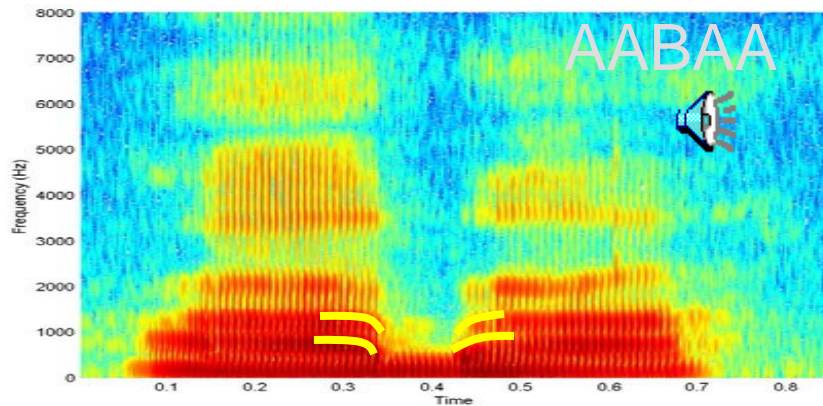
- Every phoneme has a locus
 - For many phoneme such as “B”, the locus is simply a virtual target that is never reached



- Nevertheless, during continuous speech, the spectrum for the signal tends towards this virtual locus

The Effect of Context

- Every phoneme has a locus
 - For many phoneme such as “B”, the locus is simply a virtual target that is never reached

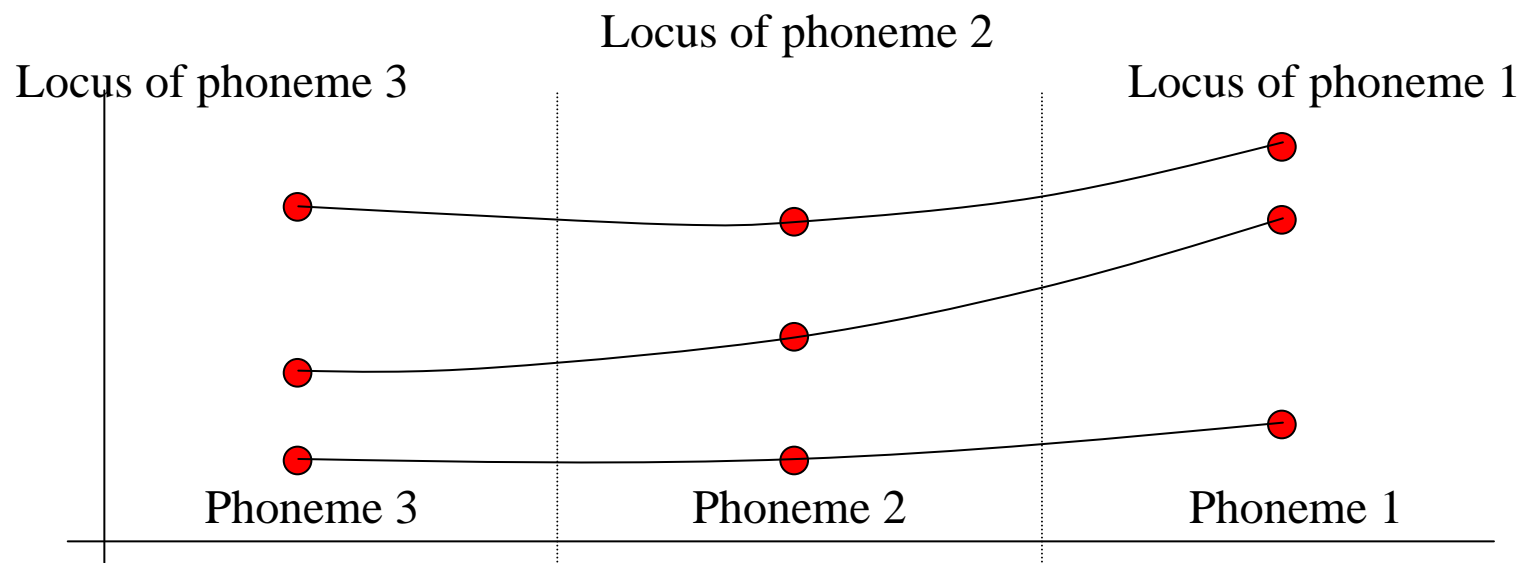
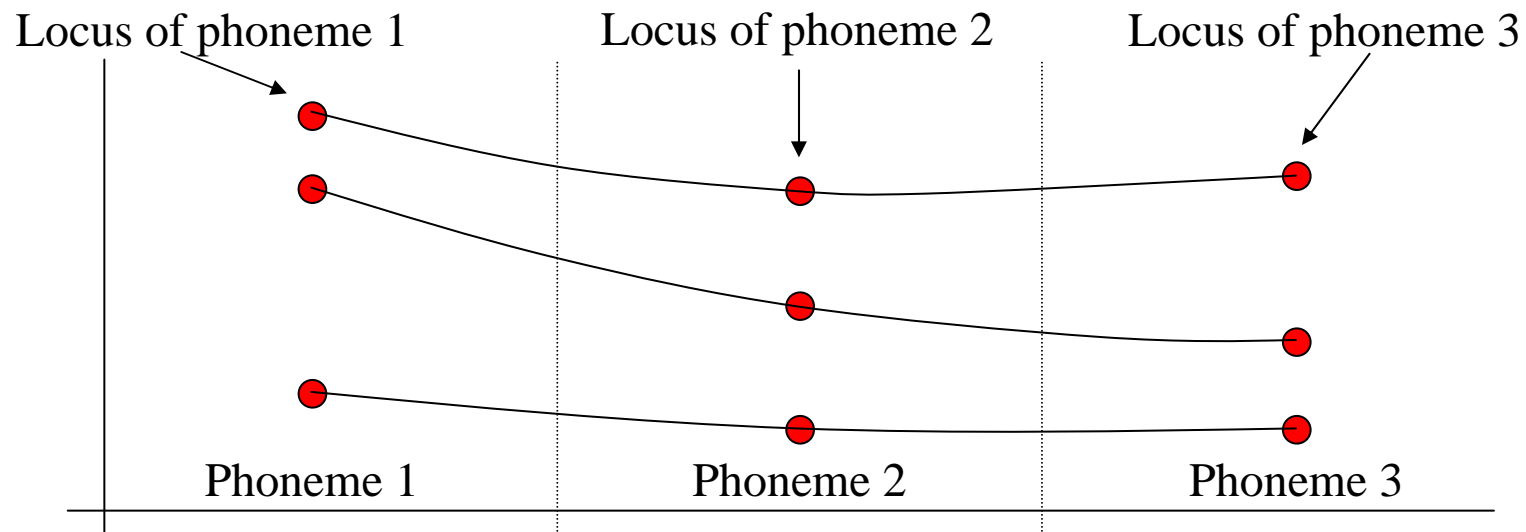


- Nevertheless, during continuous speech, the spectrum for the signal tends towards this virtual locus

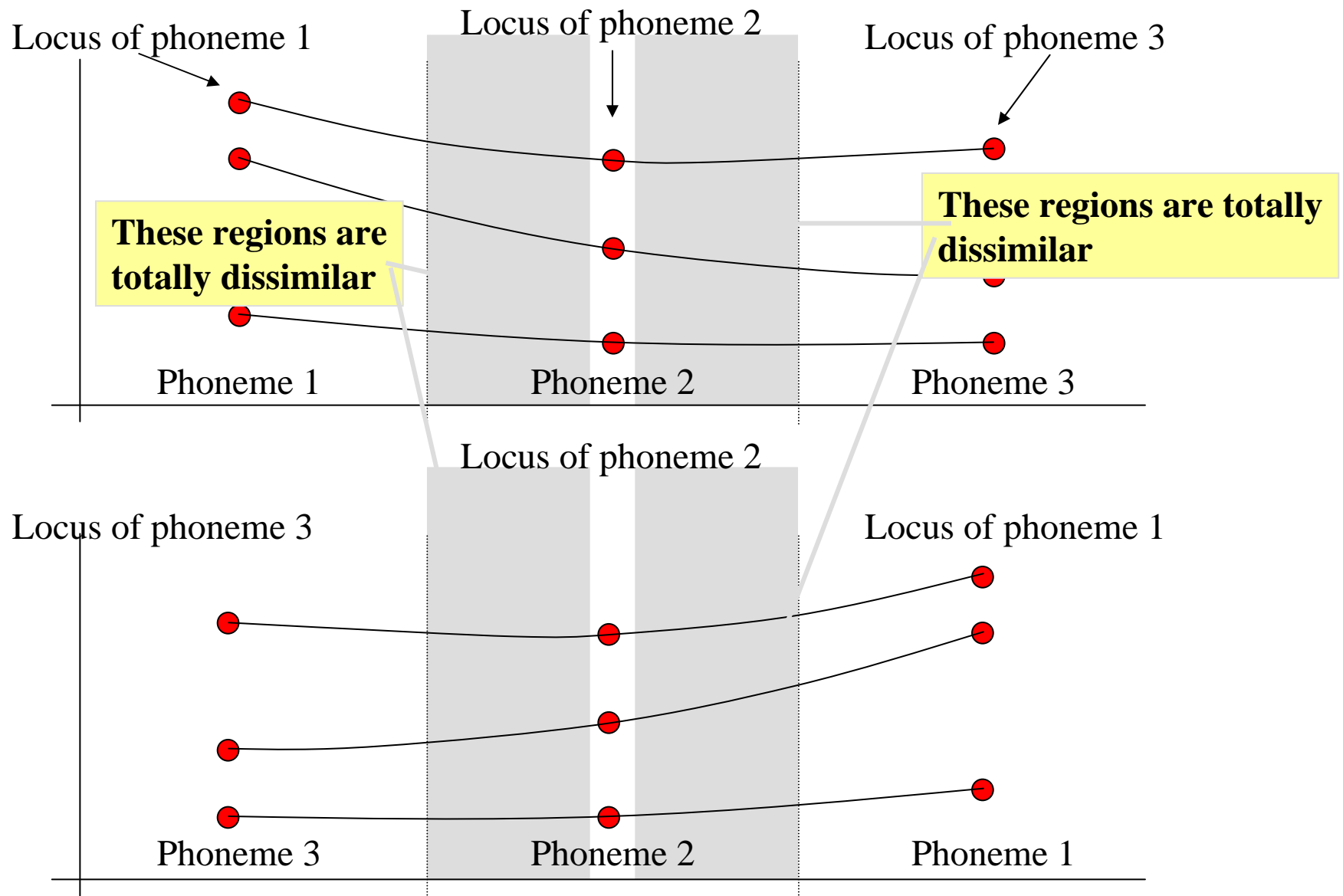
Variability among Sub-word Units

- The acoustic flow for phonemes: every phoneme is associated with a particular articulator configuration
- The spectral pattern produced when the articulators are exactly in that configuration is known as the locus for that phoneme
- As we produce sequences of sounds, the spectral patterns shift from the locus of one phoneme to the next
 - The spectral characteristics of the next phoneme affect the current phoneme
- The inertia of the articulators affects the manner in which sounds are produced
 - The manifestation lags behind the intention
 - The vocal tract and articulators are still completing the previous sound, even as we attempt to generate the next one
- As a result of articulator inertia, the spectral manifestations of any phoneme vary with the adjacent phonemes

Spectral trajectory of a phoneme is dependent on context



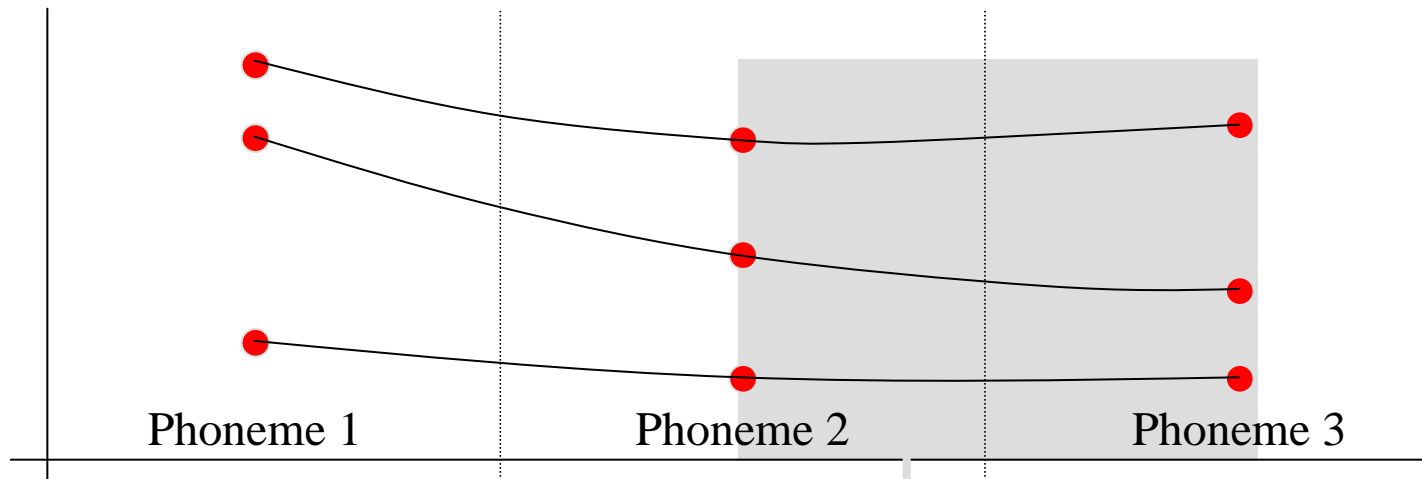
Spectral trajectory of a phoneme is dependent on context



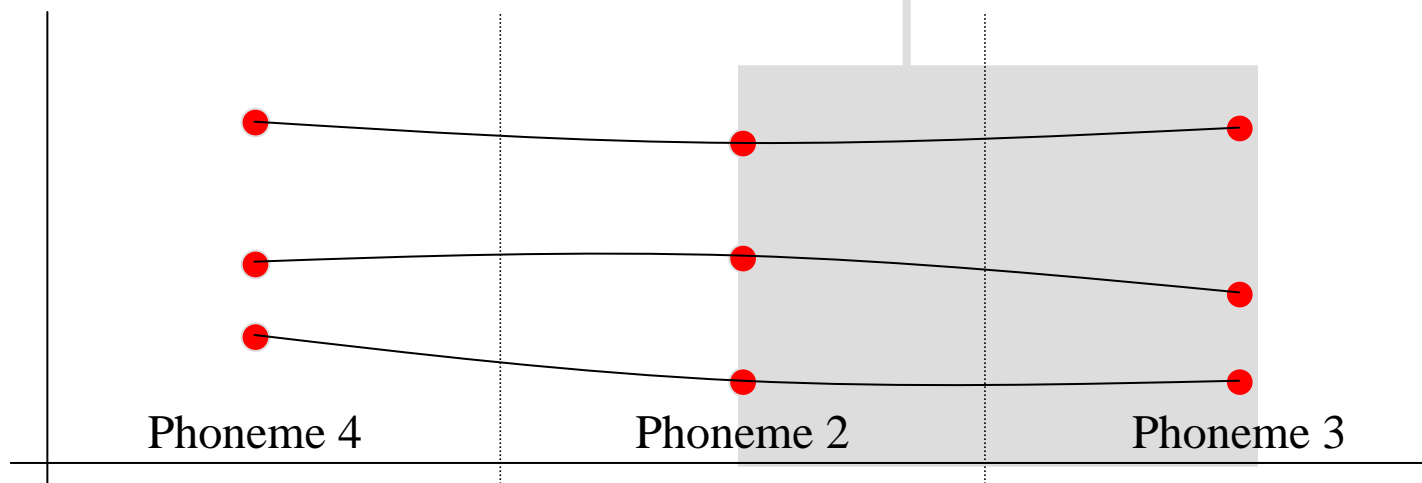
Subword units with high variability are poor building blocks for words

- Phonemes can vary greatly from instance to instance
- Due to co-articulation effects, some regions of phonemes are very similar to regions of other phonemes
 - E.g. the right boundary region of a phoneme is similar to the left boundary region of the next phoneme
- The boundary regions of phonemes are highly variable and confusable
 - They do not provide significant evidence towards the identity of the phoneme
 - Only the central regions, i.e. the loci of the phonemes, are consistent
- This makes phonemes confusable among themselves
 - In turn making these sub-word units poor building blocks for larger structures such as words and sentences
- Ideally, all regions of the sub-word units would be consistent

Diphones – a different kind of unit



The shaded regions are similar, although the phonemes to the left are different in the two cases



Diphones – a different kind of unit

- A diphone begins at the center of one phoneme and ends at the center of the next phoneme
- Diphones are much less affected by contextual, or co-articulation effects than phonemes themselves
 - All regions of the diphone are consistent, i.e. they all provide evidence for the identity of the diphone
 - Boundary regions represent loci of phonemes and are consistent
 - Central regions represent transitions between consistent loci, and are consistent
- Consequently, diphones are much better building blocks for word HMMs than phonemes
- For a language with N phonemes, there are N^2 diphones
 - These will require correspondingly larger amounts of training data
 - However, the actual number of sub-word units remains limited and enumerable
 - As opposed to words that are unlimited in number

The Diphone

- Phonetic representation of ROCK:
 - ROCK: R AO K

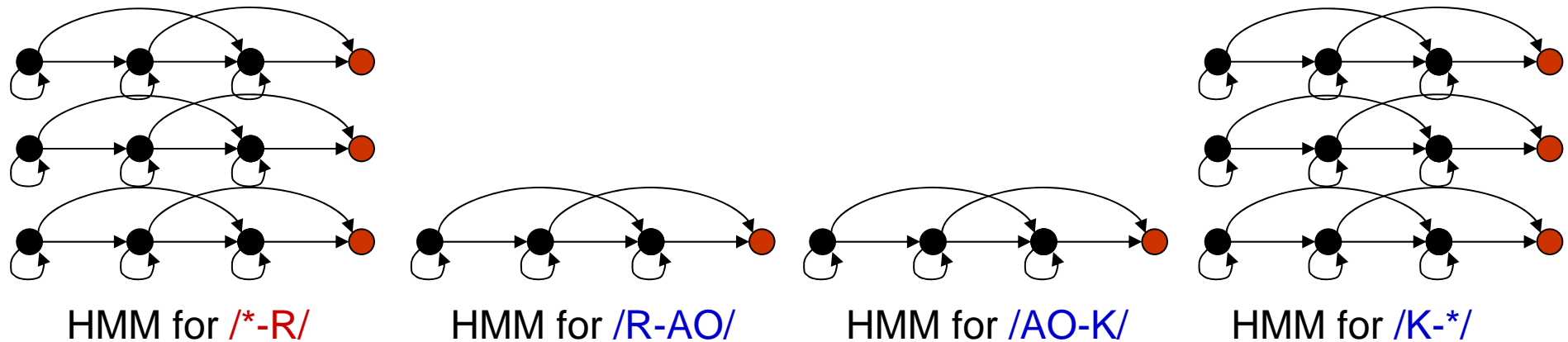
- Diphone representation:
 - ROCK: (??-R), (R-AO), (AO-K), (K-??)
 - Each unit starts from the middle of one phoneme and ends at the middle of the next one

- **Word boundaries are a problem**
 - The diphone to be used in the first position (??-R) depends on the last phoneme of the previous word!
 - ?? is the last phoneme of the previous word
 - Similarly, the diphone at the end of the word (K-??) depends on the next word

- We build a separate diphone-based word model for every combination of preceding and following phoneme observed in the grammar

- ***The boundary units are SHARED by adjacent words***

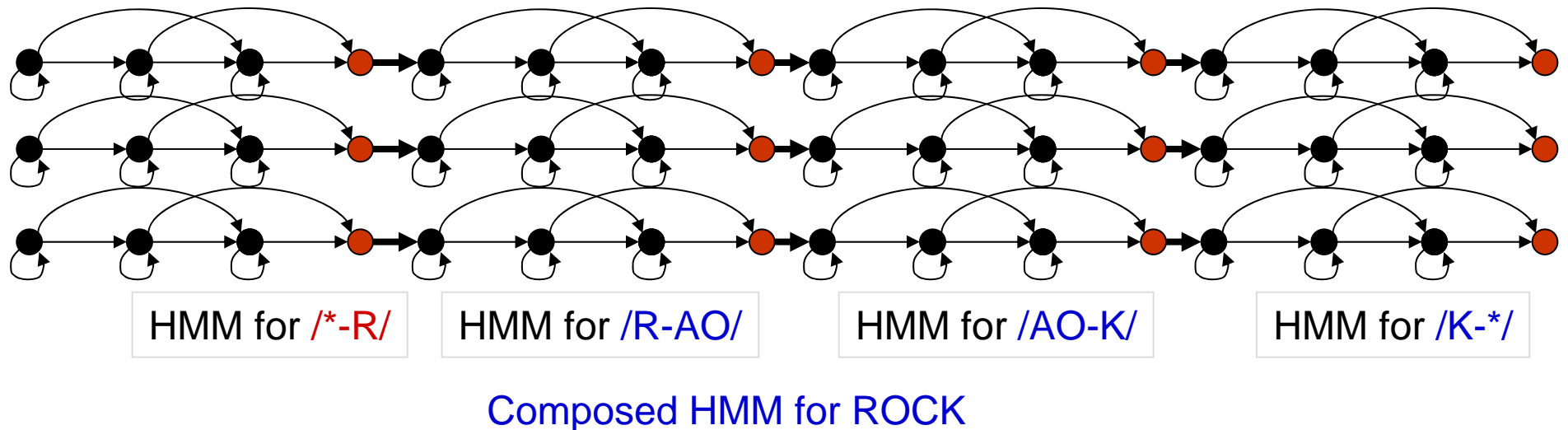
Building word HMMs with diphones



Components of HMM for ROCK

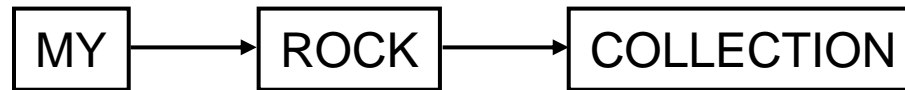
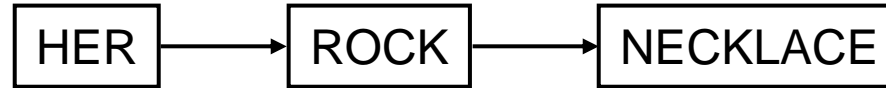
- Dictionary entry for ROCK: /R/ /AO/ /K/
- Word boundary units are not unique
- The specific diphone HMMs to be used at the ends of the word depend on the previous and succeeding word
 - E.g. The first diphone HMM for ROCK in the word series JAILHOUSE ROCK is /S-R/, whereas for PLYMOUTH ROCK, it is /TH-R/
- As a result, there are as many HMMs for "ROCK" as there are possible left-neighbor, right-neighbor phoneme combinations

Building word HMMs with diphones



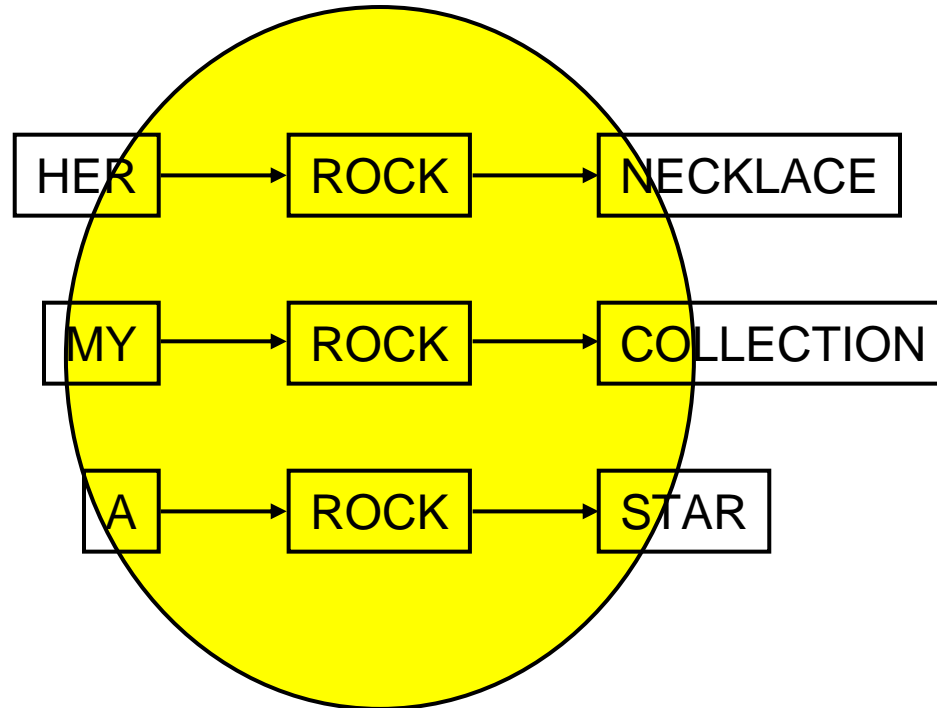
- We end up with as many word models for ROCK as the number of possible combinations of words to the right and left

Word Sequences using Diphones



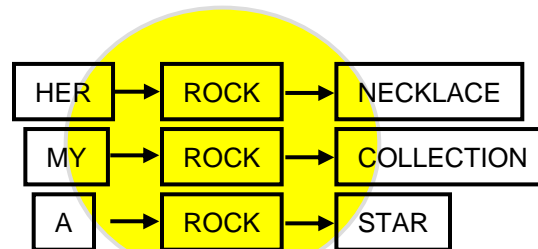
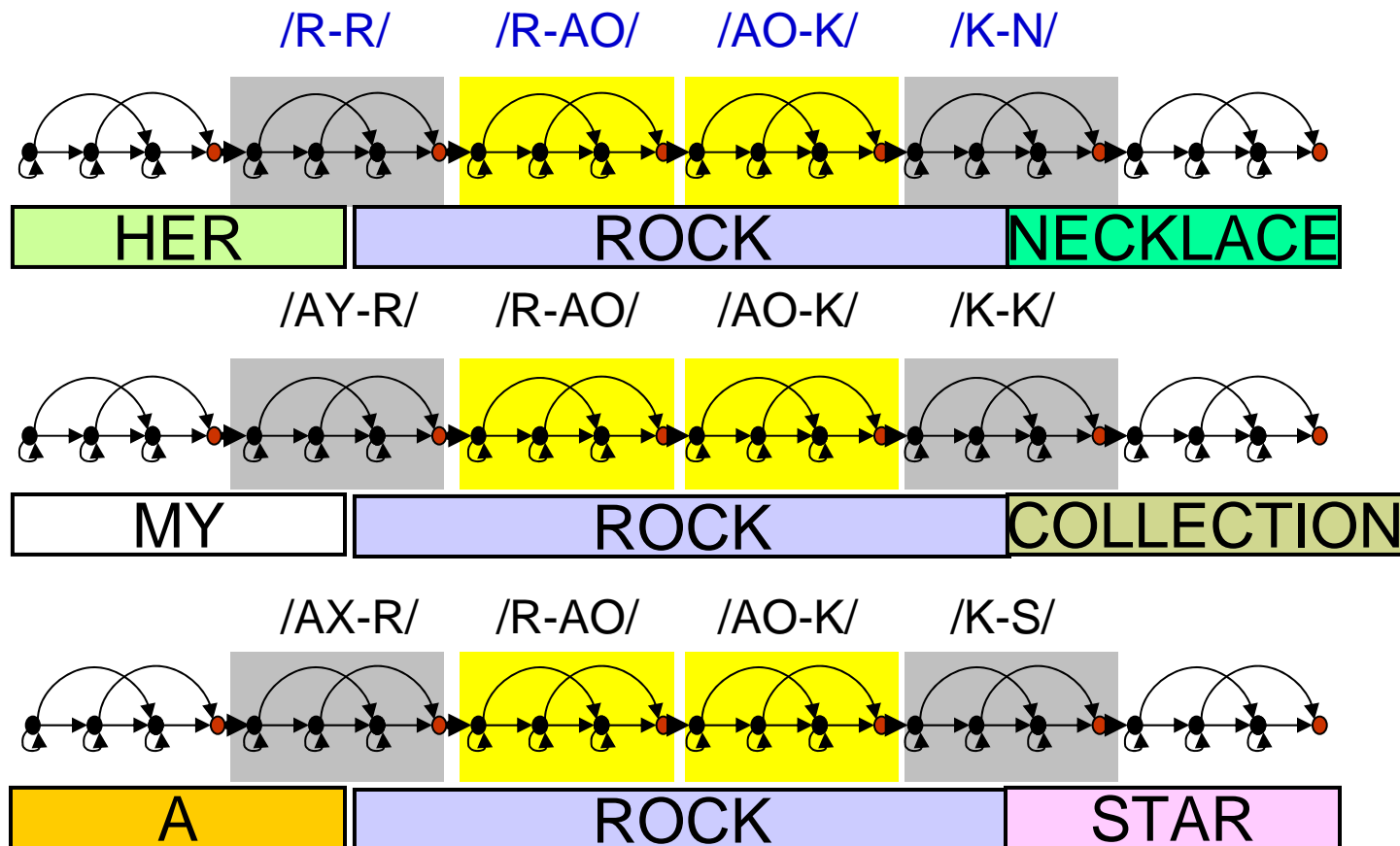
- Consider this set of sentences

Word Sequences using Diphones

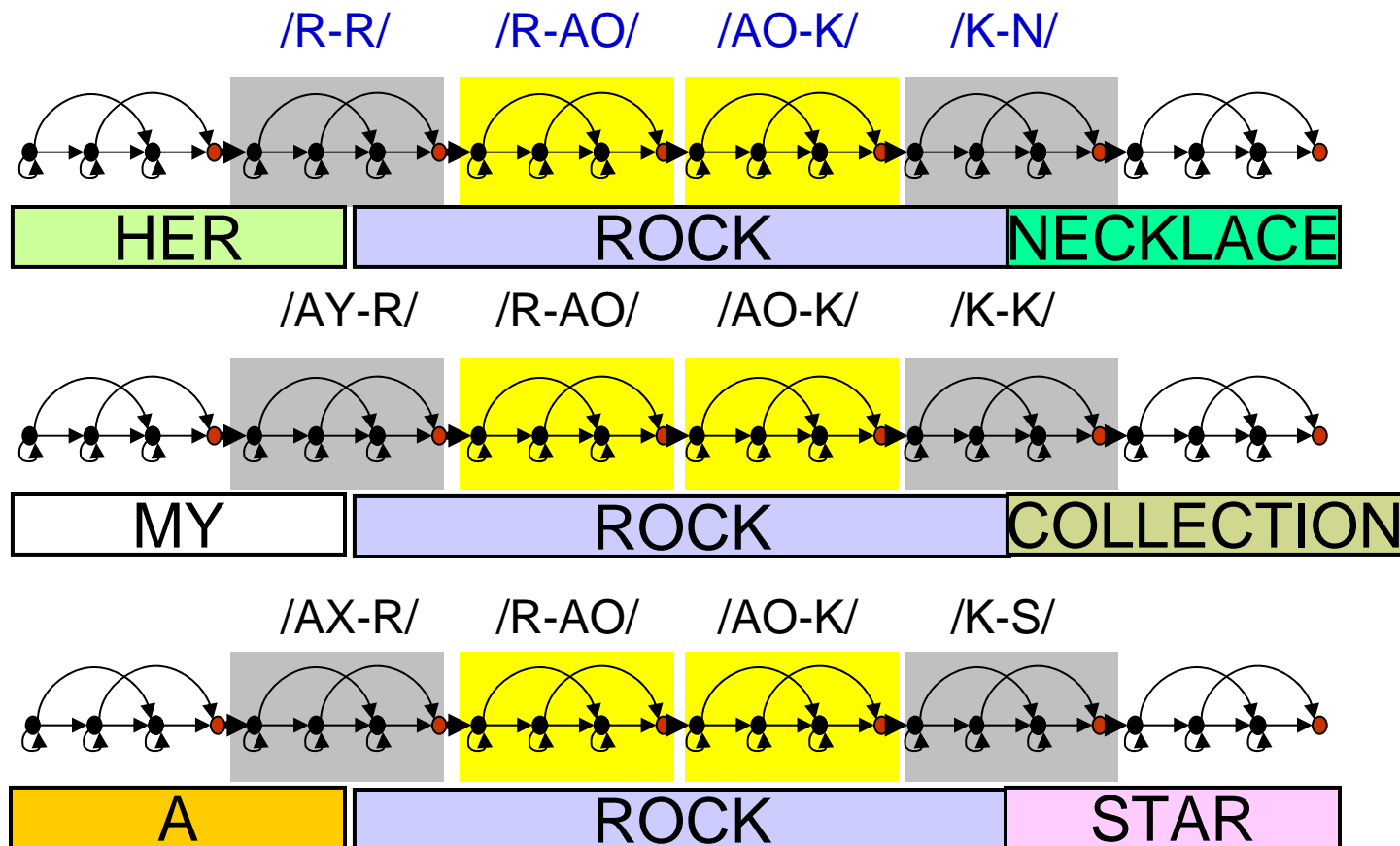


- Consider this set of sentences
- For illustration, we will concentrate on this region

Building word HMMs with diphones

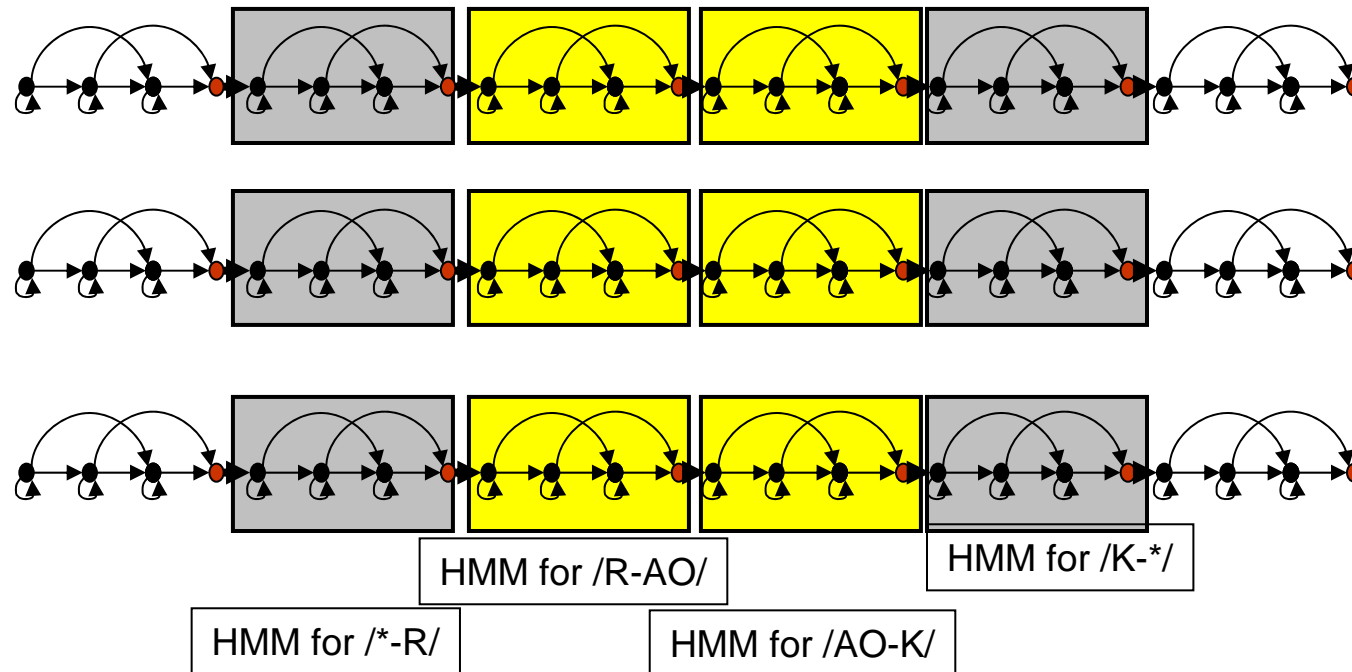


Building word HMMs with diphones



- Each instance of Rock is a different model
 - The 3 instances are not copies of one another
- The boundary units (gray) are shared with adjacent words

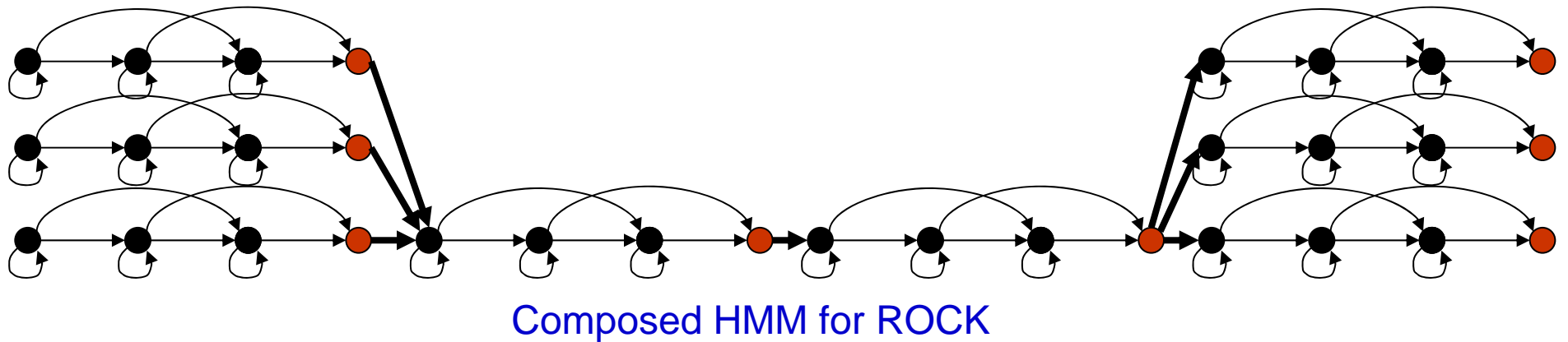
Building word HMMs with diphones



Composed HMM for ROCK

- We end up with as many word models for ROCK as the number of possible words to the right and left

Building word HMMs with diphones



- Under some conditions, portions of multiple versions of the model for a word can be collapsed
 - Depending on the grammar

Building a Diphone-based recognizer

- Train models for all Diphones in the language
 - If there are N phonemes in the language, there will be N^2 diphones
 - For 40 phonemes, there are thus 1600 diphones; still a manageable number

- Training: HMMs for word sequences must be built using Diphone models

- There will no longer be distinctly identifiable word HMMs because boundary units are shared among adjacent words

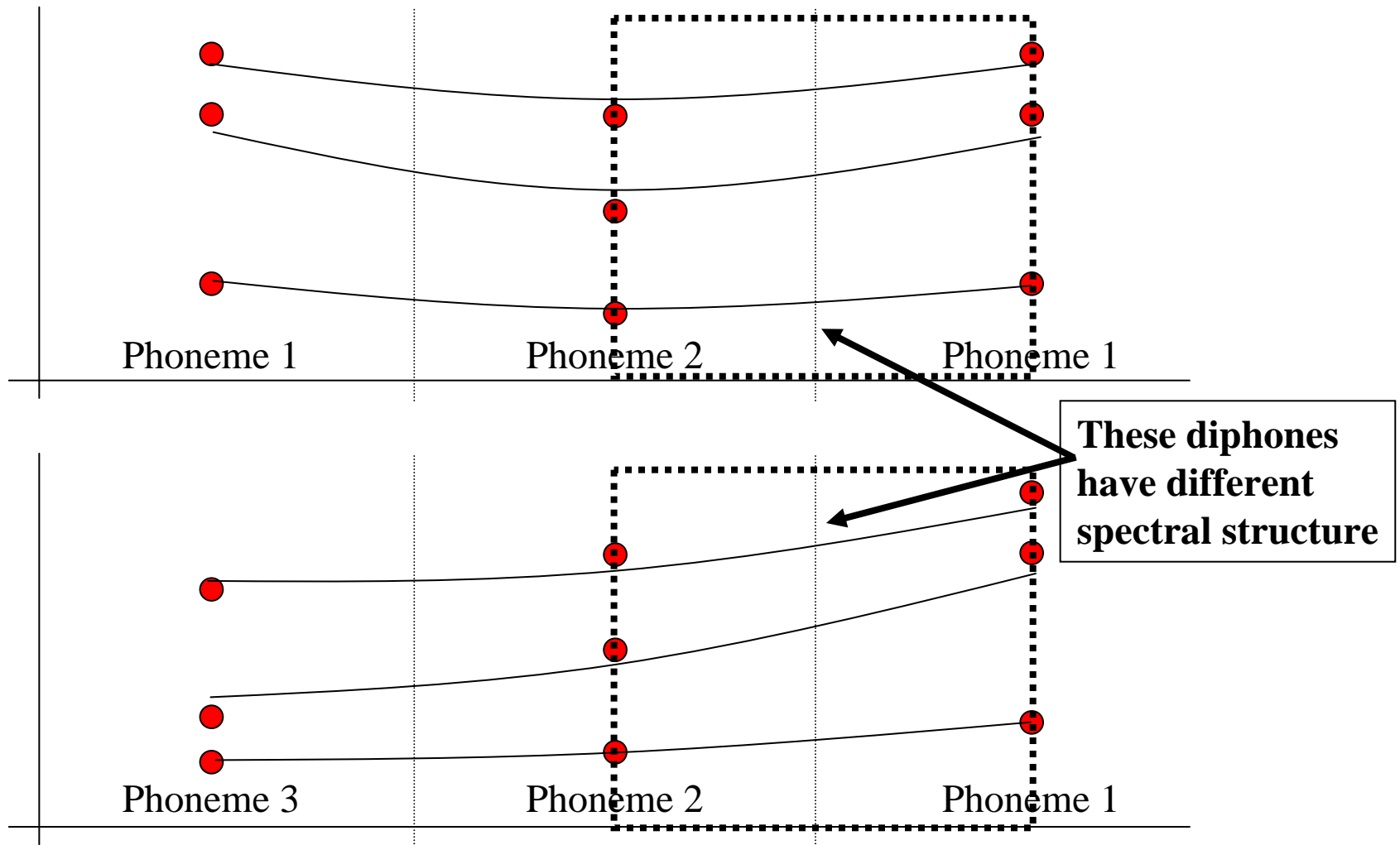
- A similar problem would arise during recognition

Word-boundary Diphones

- Cross-word diphones have somewhat different structure than within-word diphones, even when they represent the same phoneme combinations
 - E.g. the within-word diphone AA-F in the word “AFRICA” is different from the word-boundary diphone AA-F in BURKINA FASO”
 - Stresses applied to different portions of the sounds are different
 - This is true even for cross-syllabic diphones

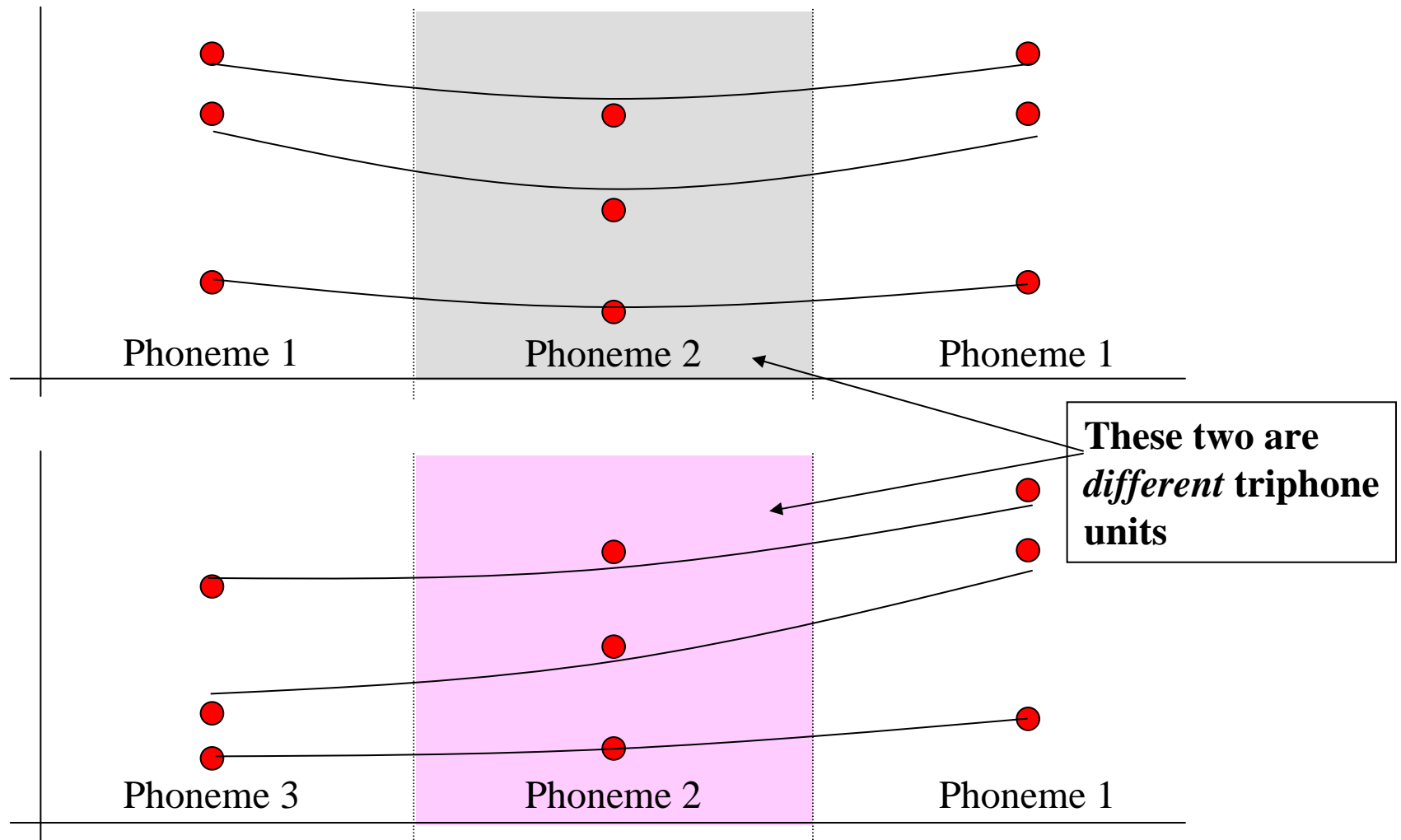
- It is advantageous to treat cross-word diphones as being distinct from within-word diphones
 - This results in a doubling of the total number of diphones in the language to $2N^2$
 - Where N is the total number of phonemes in the language
 - We will train cross-word diphones only from cross-word data and within-word diphones from only within-word data

Improving upon Diphones



- Loci may never occur in fluent speech
- Resulting in variation even in Diphones
- The actual spectral trajectories are affected by adjacent phonemes
- Diphones do not capture all effects

Triphones: PHONEMES IN CONTEXT



- Triphones represent phoneme units that are specific to a context
 - E.g. "The kind of phoneme 2 that follows phoneme 3 and precedes phoneme 3"
- **The triphone is *not* a triplet of phonemes – it still represents a single phoneme**

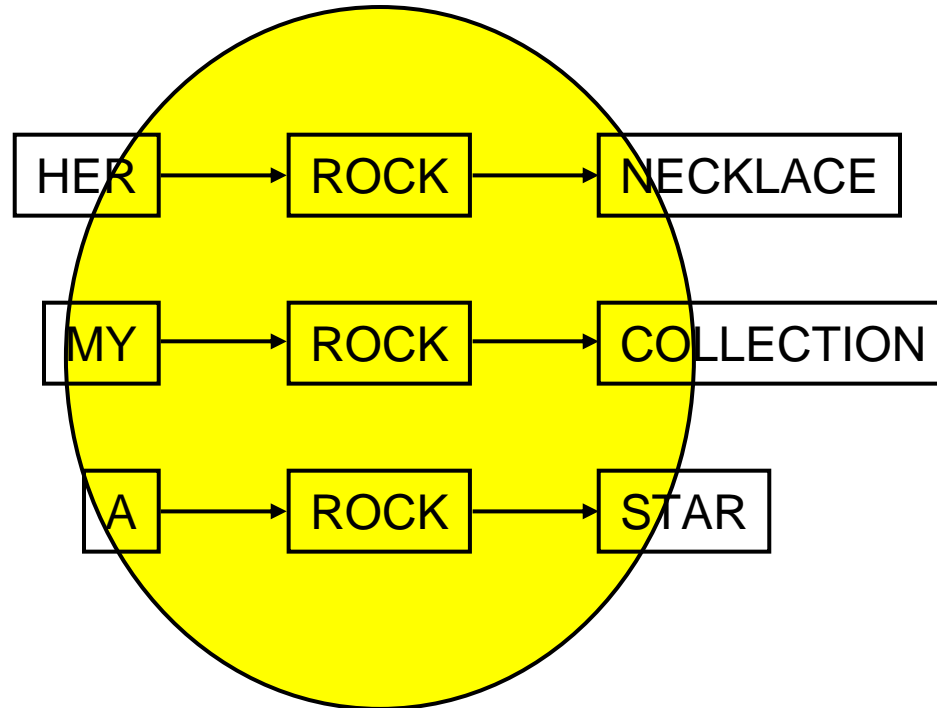
Triphones

- Triphones are actually *specialized phonemes*
 - The triphones AX (B, T), AX(P, D), AX(M,G) all represent variations of AX
- To build the HMM for a word, we simply concatenate the HMMs for individual triphones in it
 - E.g. R AO K : R(??, AO) AO (R,K) K(AO,??)
 - We link the HMMs for the triphones for R(??,AO), AO(R,K) and K(AO,??)
- Unlike diphones, no units are shared across words
- Like diphones, however, the boundary units R(??,AO), K(AO,??) depend on the boundary phonemes of adjacent words
 - In fact it becomes more complex than for diphones

Building HMMs for word sequences

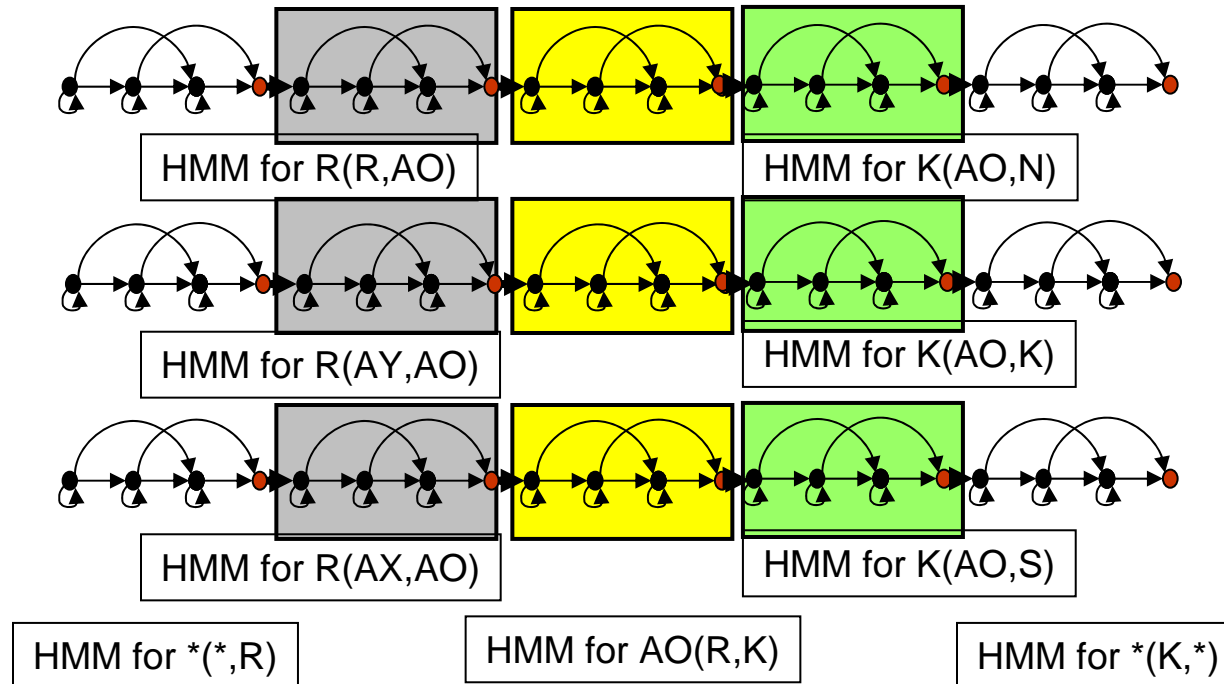
- As in the case of diphones, we cannot compose a unique model for a word
- The model that is composed will depend on the adjacent words

Word Sequences using Triphones



- Consider this set of sentences
- For illustration, we will concentrate on this region

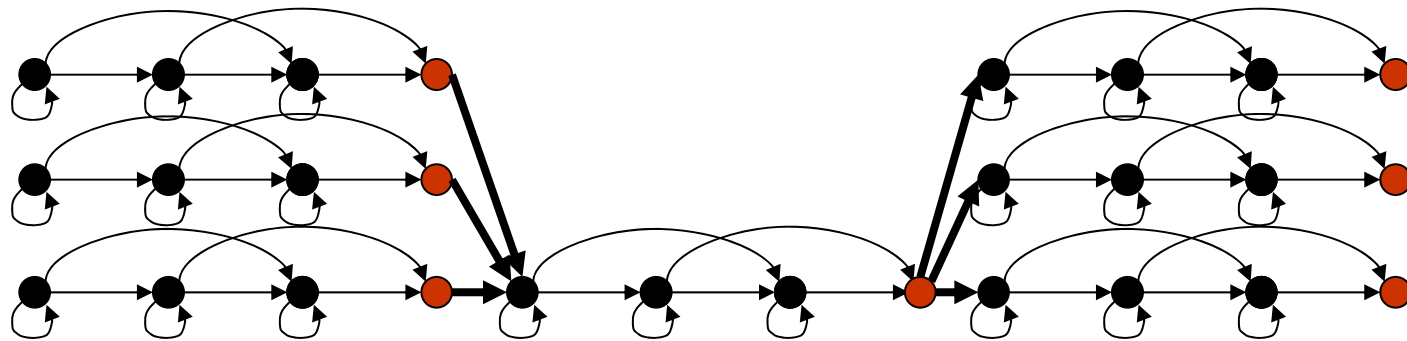
Building word HMMs with triphones



Composed HMM for ROCK

- As for diphones, we end up with as many word models for ROCK as the number of possible words to the right and left

Building word HMMs with triphones



Composed HMM for ROCK

- As with diphones, portions of multiple versions of the model for a word can be collapsed
 - Depending on the grammar












It can get more complex

- Triphones at word boundaries are dependent on neighbouring words.
- This results in significant complication of the HMM for the language (through which we find the best path, for recognition)
 - Resulting in larger HMMs and slower search








Dictionary

Five: **F AY V**
Four: **F OW R**
Nine: **N AY N**
<sil>: **SIL**
++breath++: **+breath+**

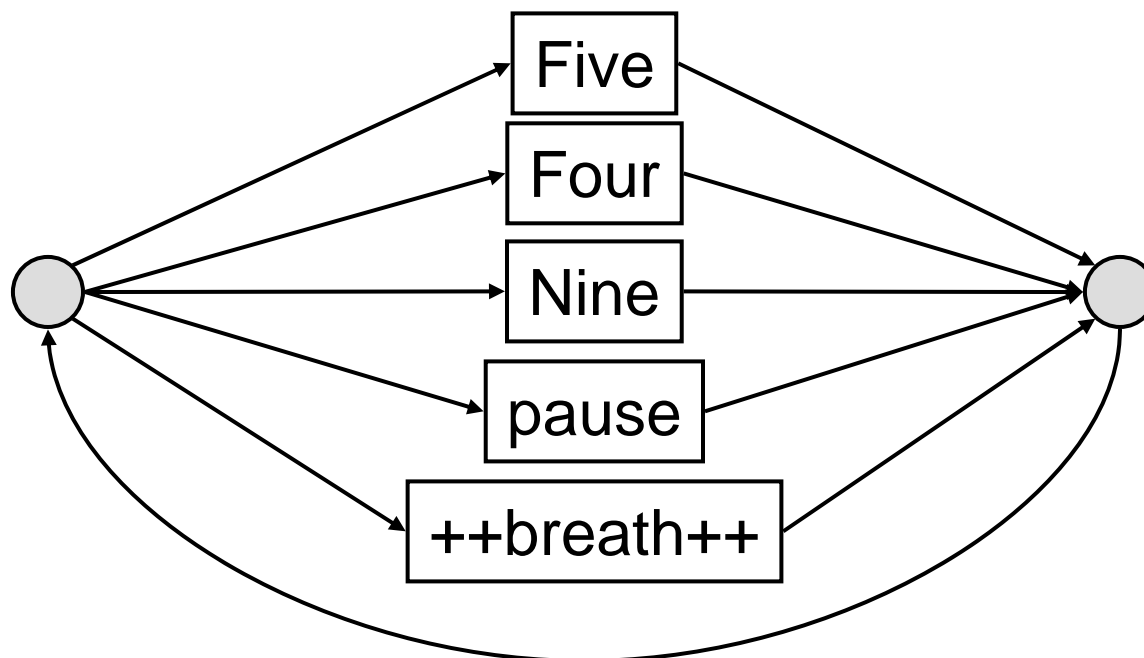
Lexicon

Five   
Four   
Nine   
<sil> 
++Breath++ 

Listed here are five “words” and their pronunciations in terms of “phones”. Let us assume that these are the only words in the current speech to be recognized. The recognition vocabulary thus consists of five words. The system uses a **dictionary** as a reference for these mappings.

 = F
 = AY
 = V
 = OW
 = R
 = N
 = SIL
 = +breath+

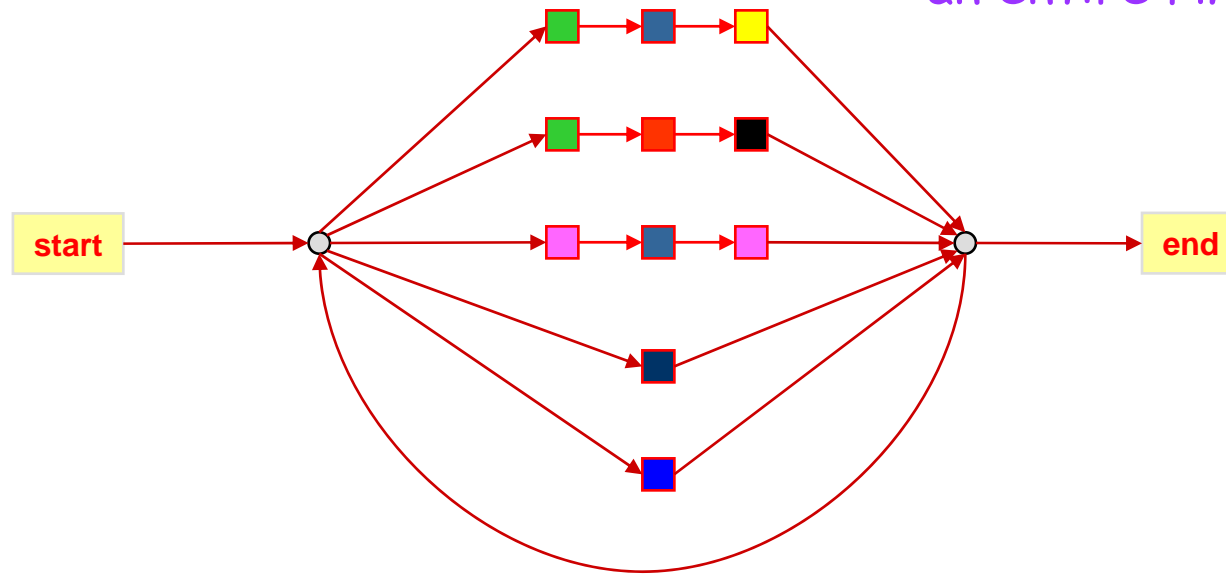
A Simple Looping Graph For Words



- We will compose an HMM for the simple grammar above
 - A person may say any combination of the words five, nine and four with silences and breath inbetween

Using (CI) Phonemes

Each coloured square represents an entire HMM



Lexicon				
Five	<table border="1"><tr><td>Green</td><td>Blue</td><td>Yellow</td></tr></table>	Green	Blue	Yellow
Green	Blue	Yellow		
Four	<table border="1"><tr><td>Green</td><td>Red</td><td>Black</td></tr></table>	Green	Red	Black
Green	Red	Black		
Nine	<table border="1"><tr><td>Pink</td><td>Blue</td><td>Pink</td></tr></table>	Pink	Blue	Pink
Pink	Blue	Pink		
<sil>	<table border="1"><tr><td>Dark Blue</td></tr></table>	Dark Blue		
Dark Blue				
++Breath++	<table border="1"><tr><td>Blue</td></tr></table>	Blue		
Blue				

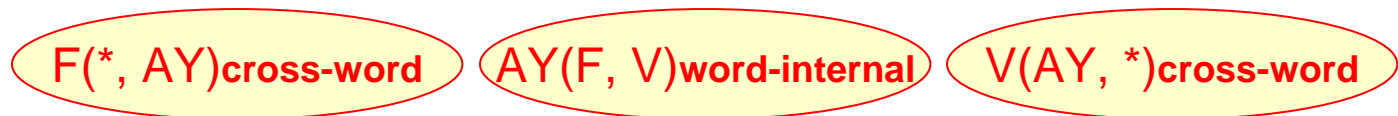
Word boundary units are not context specific.
All words can be connected to (and from) null nodes

Green	= F
Blue	= AY
Yellow	= V
Red	= OW
Black	= R
Pink	= N
Dark Blue	= SIL
Blue	= +breath+

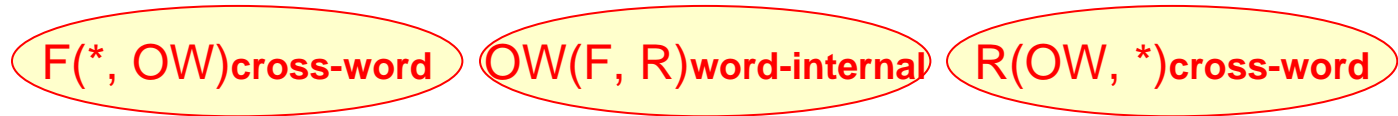
Using Triphones

Using the dictionary as reference, the system first maps each word into triphone-based pronunciations. Each triphone further has a characteristic **label or type**, according to where it occurs in the word. *Context is not initially known for cross-word triphones.*

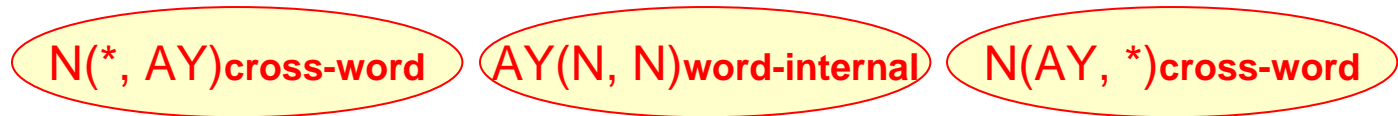
Five



Four



Nine



<sil>



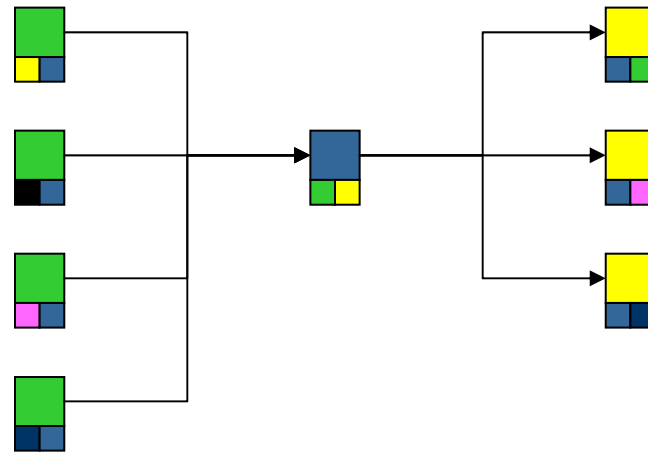
++Breath++



Each triphone is modelled by an HMM
Silence is modelled by an HMM
breath is modelled by an HMM

Using Triphones

HMM for “Five”.
This is composed of 8 HMMs.














Each triple-box represents a triphone. Each triphone model is actually a left-to-right HMM.









A triphone is a *single* context-specific phoneme.
It is *not* a sequence of 3 phones.

Expand the word Five

- All last phones (except +breath+) become left contexts for first phone of Five.
- All first phones (except +breath+) become right contexts for last phone of Five
- Silence can form contexts, but itself does not have any context dependency.
- **Filler** phones (e.g. +breath+) are treated as silence when building contexts. Like silence, they themselves do not have any context dependency.

Lexicon

Five			
Four			
Nine			
<sil>			
++Breath++			

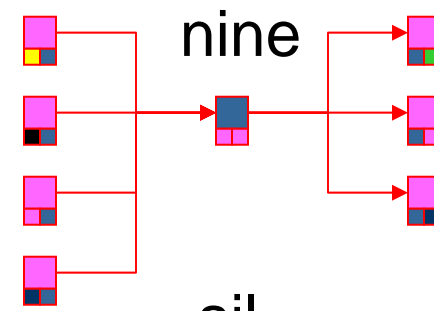
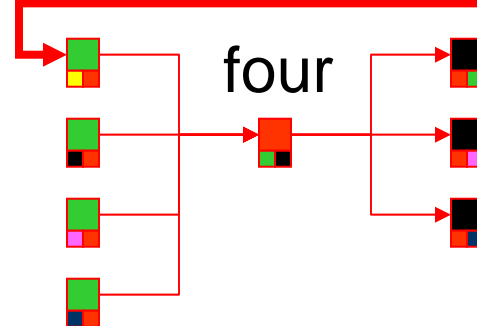
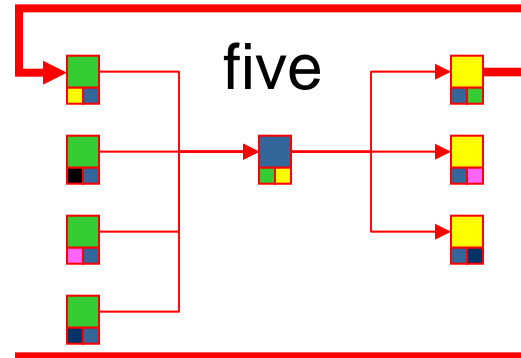
	= F
	= AY
	= V
	= OW
	= R
	= N
	= SIL
	= +breath+

The triphone based Grammar HMM

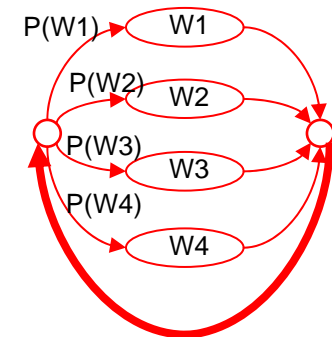
- Even a simple looping grammar becomes very complicated because of cross-word triphones

- The HMM becomes *even more* complex when some of the words have only one phoneme
 - There will be as many instances of the HMM for the word as there are word contexts
 - No portion of these HMMs will be shared!

The triphone based Grammar HMM

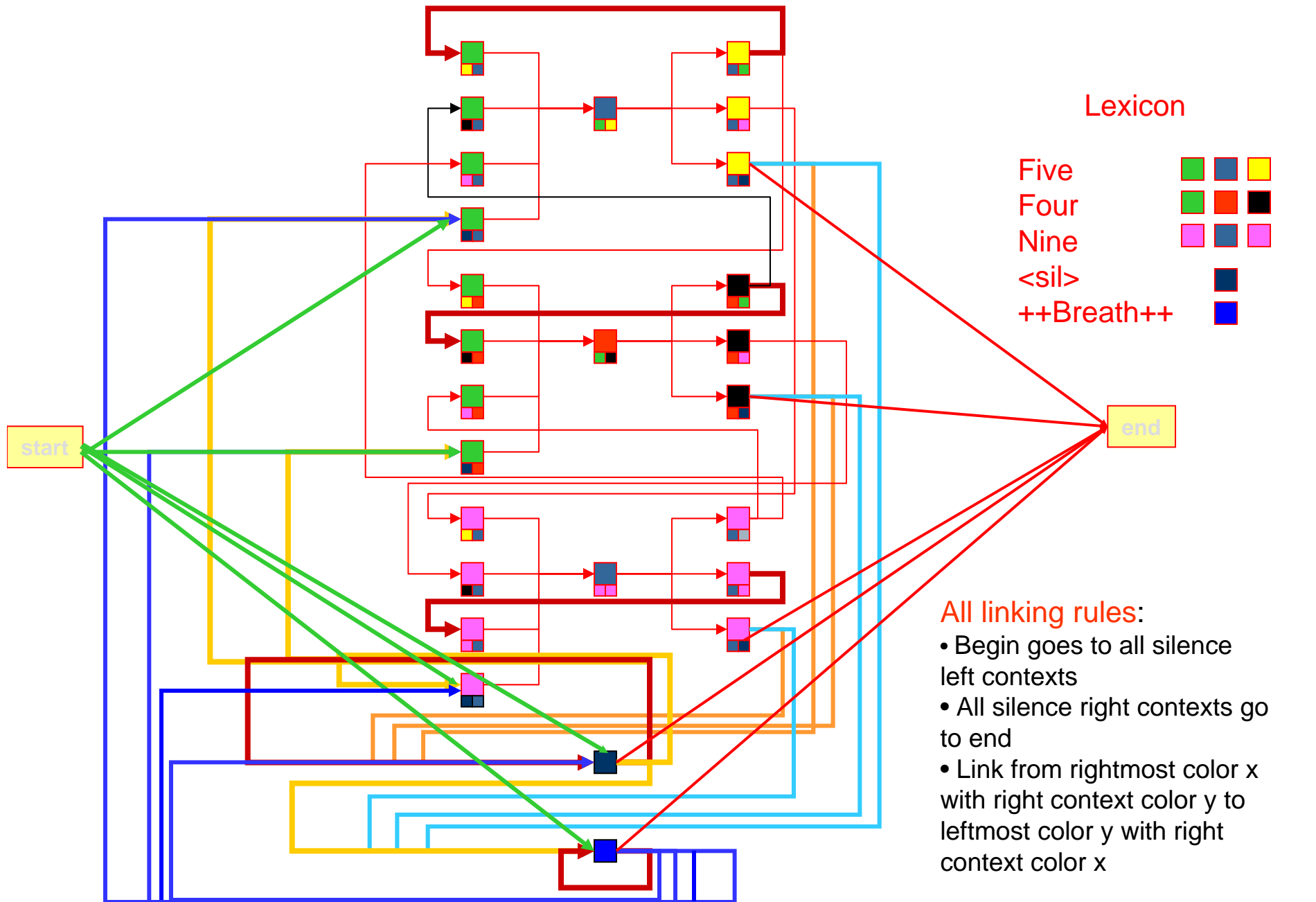


Lexicon	
Five	
Four	
Nine	
<sil>	
++Breath++	



Linking rule:
 Link from rightmost color x
 with right context color y
 to leftmost color y with
 right context color x

The triphone based Grammar HMM



Types of triphones

- A triphone in the middle of a word sounds different from the same triphone at word boundaries
 - e.g the word-internal triphone AX(G,T) from GUT: G AX T
 - Vs. cross-word triphone AX(G,T) in BIG ATTEMPT

- The same triphone in a single-word (e.g when the central phoneme is a complete word) sounds different
 - E.g. AX(G,T) in WAG A TAIL
 - The word A: AX is a single-phone word and the triphone is a single-word triphone

- We distinguish four types of triphones:
 - Word-internal
 - Cross-word at the beginning of a word
 - Cross-word at the end of a word
 - Single-word triphones

- Separate models are learned for the four cases and appropriately used

Context Issues

- Phonemes are affected by adjacent phonemes.
- If there is no adjacent phoneme, i.e. if a phoneme follows or precedes a silence or a pause, that will also have a unique structure
- We will treat “silence” as a context also
- “Filler” sounds like background noises etc. are typically treated as equivalent to silence for context
- “Voiced” filler sounds, like “UM”, “UH” etc. do affect the structure of adjacent phonemes and must be treated as valid contexts

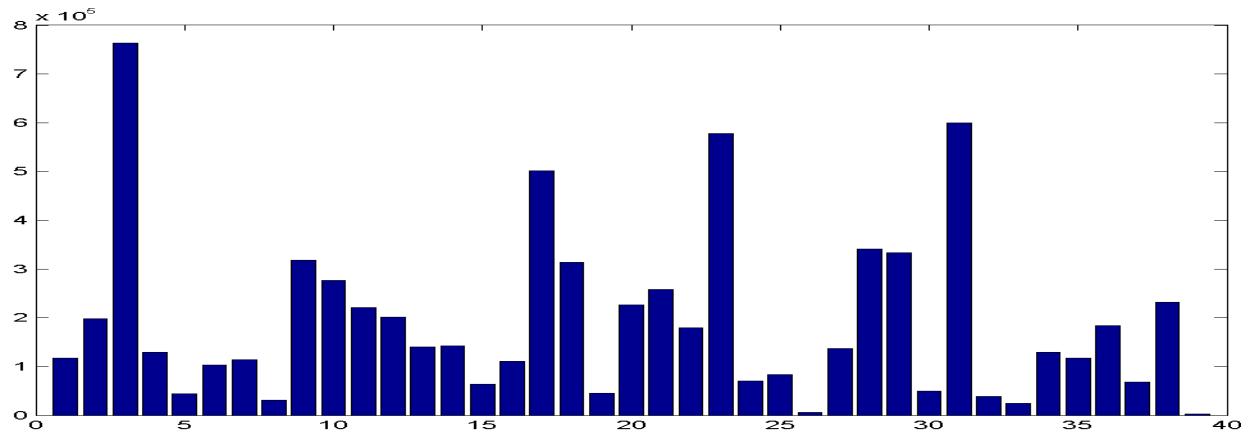
Training Data Considerations for Nphone Units

- 1.53 million words of training data (~70 hours)
- All 39 phonemes are seen (100%)

- 1387 of 1521 possible diphones are seen (91%)
 - Not considering cross-word diphones as distinct units

- 24979 of 59319 possible triphones are seen (42%)
 - not maintaining the distinction between different kinds of triphones

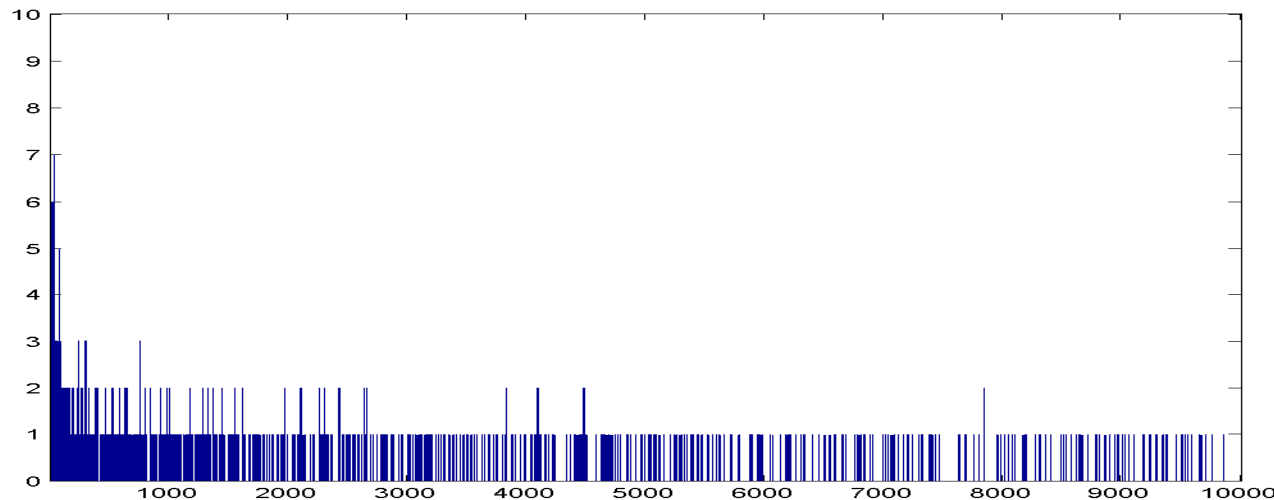
Counts of CI Phones



Histogram of the number of occurrences of the 39 phonemes in 1.5 million words of Broadcast News

- All context-independent phonemes occur in sufficient numbers to estimate HMM parameters well
- Some phonemes such as “ZH” may be rare in some corpora. In such cases, they can be merged with other similar phonemes, e.g. ZH can be merged with SH, to avoid a data insufficiency problem

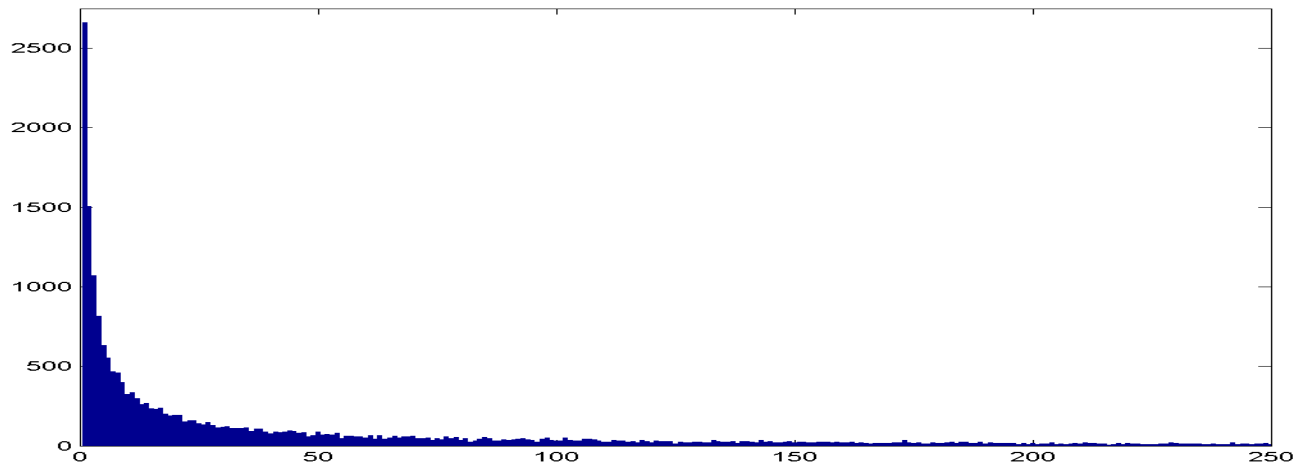
Counts of Diphones



“Count of counts” histogram for the 1387 diphones in 1.5 million words of Broadcast News

- Counts (x axis) > 10000 not shown, Count of counts (Y axis) clipped from above at 10
- The figure is much flatter than the typical trend given by Zipf’s law
- The mean number of occurrences of diphones is 4900
- Most diphones can be well trained
 - Diphones with low (or 0) count, e.g. count < 100, account for less than 0.1% of the total probability mass
 - Again, contrary to Zipf’s law
 - Diphones with low count, or unseen diphones, can be clustered with other similar diphones with minimal loss of recognition performance

Counts of Triphones



"Count of counts" histogram for the 24979 triphones in 1.5 million words of Broadcast News

- Counts (x axis) > 250 not shown
- Follows the trends of Zipf's law very closely
 - Because of the large number of triphones
- The mean number of occurrences of *observed* triphones is 300
- The majority of the triphones are not seen, or are rarely seen
 - 58% of all triphones are never seen
 - 86% of all triphones are seen less than 10 times
 - The majority of the triphones in the language will be poorly trained

Higher order Nphones

- Spectral trajectories are also affected by farther contexts
 - E.g. by phonemes that are two phonemes distant from the current one
- The effect of longer contexts is much smaller than that of immediate context, in most speech
 - The use of longer context models only results in relatively minor improvements in recognition accuracy
- There are far too many possibilities for longer context units
 - E.g, there are 40^5 possible quinphone units (that consider a 2-phoneme context on either side), even if we ignore cross-word effects
- Cross-word effects get far more complicated with longer-context units
 - Cross-word contexts may now span multiple words. The HMM for any word thus becomes dependent on the the previous two (or more) words, for instance.

Training Tradeoffs

- Word models
 - Very effective, if they can be well trained
 - Difficult to train well, when vocabularies get large
 - Cannot recognize words that are not seen in training data

- Context-independent phoneme models
 - Simple to train; data insufficiency rarely a problem
 - All phonemes are usually seen in the training data
 - If some phonemes are not seen, they can be mapped onto other, relatively common phonemes

- Diphones
 - Better characterization of sub-word acoustic phenomena than simple context-independent phonemes
 - Data insufficiency a relatively minor problem. Some diphones are usually not seen in training data.
 - Their HMMs must be inferred from the HMMs for seen diphones

Training Tradeoffs

- Triphones
 - Characterizations of phonemes that account for contexts on both sides
 - Result in better recognition than diphones
 - Data insufficiency is a problem: the majority of triphones are not seen in training data.
 - Parameter sharing techniques must be used to train uncommon triphones derive HMMs for unseen triphones
 - Advantage: can back-off to CI phonemes when HMMs are not available for the triphones themselves

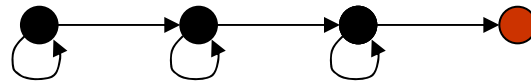
- Higher order Nphone models
 - Better characterizations than triphones, however the benefit to be obtained from going to Nphones of higher contexts is small
 - Data insufficiency a huge problem – most Nphones are not seen in the training data
 - Complex techniques must be employed for inferring the models for unseen Nphones
 - Alternatively, one can backoff to triphones or CI-phonemes

Context-Dependent Phonemes for Recognition

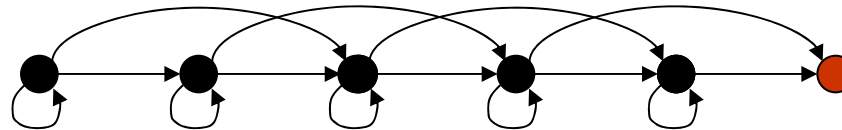
- Context-independent phonemes are rarely used by themselves
 - Recognition performance is not good enough
- Diphone models are typically used in embedded systems
 - Compact models
 - Nearly all diphones will be seen in the training data
 - Recognition accuracy still not as good as triphones
- Triphones are the most commonly used units
 - Best recognition performance
 - Model size controlled by tricks such as parameter sharing
 - Most triphones not see in training data
 - Nevertheless models for them can be constructed as we will see later
- Quinphone models are used by some systems
 - Much increase in system complexity for small gains

HMM Topology for N-phones

- Context-dependent phonemes are modeled with the same topology as context independent phonemes
 - Most systems use a 3-state topology
 - All C-D phonemes have the same topology



- Some older systems use a 5-state topology
 - Which permits states to be skipped entirely
 - This is not demonstrably superior to the 3-state topology



Training Triphone Models

- Like training context-independent (CI) phoneme models
- First step: Train CI models
 - Important
 - Needed for initialization
- Initialize all triphone models using corresponding CI model
 - E.g. the model for IY (B,T) [the phoneme IY that occurs after B and before T, such as the IY in BEAT] is initialized by the model for the context independent phoneme IY
- Create HMMs for the word sequences using triphones
- Train HMMs for all triphones in the training data using Baum Welch

Training Triphone Models with SphinxTrain

- A simple exercise:
 - Train triphone models using a small corpus
 - Recognize a small test set using these models