Informative Dialect Recognition using Context-Dependent Pronunciation Modeling*

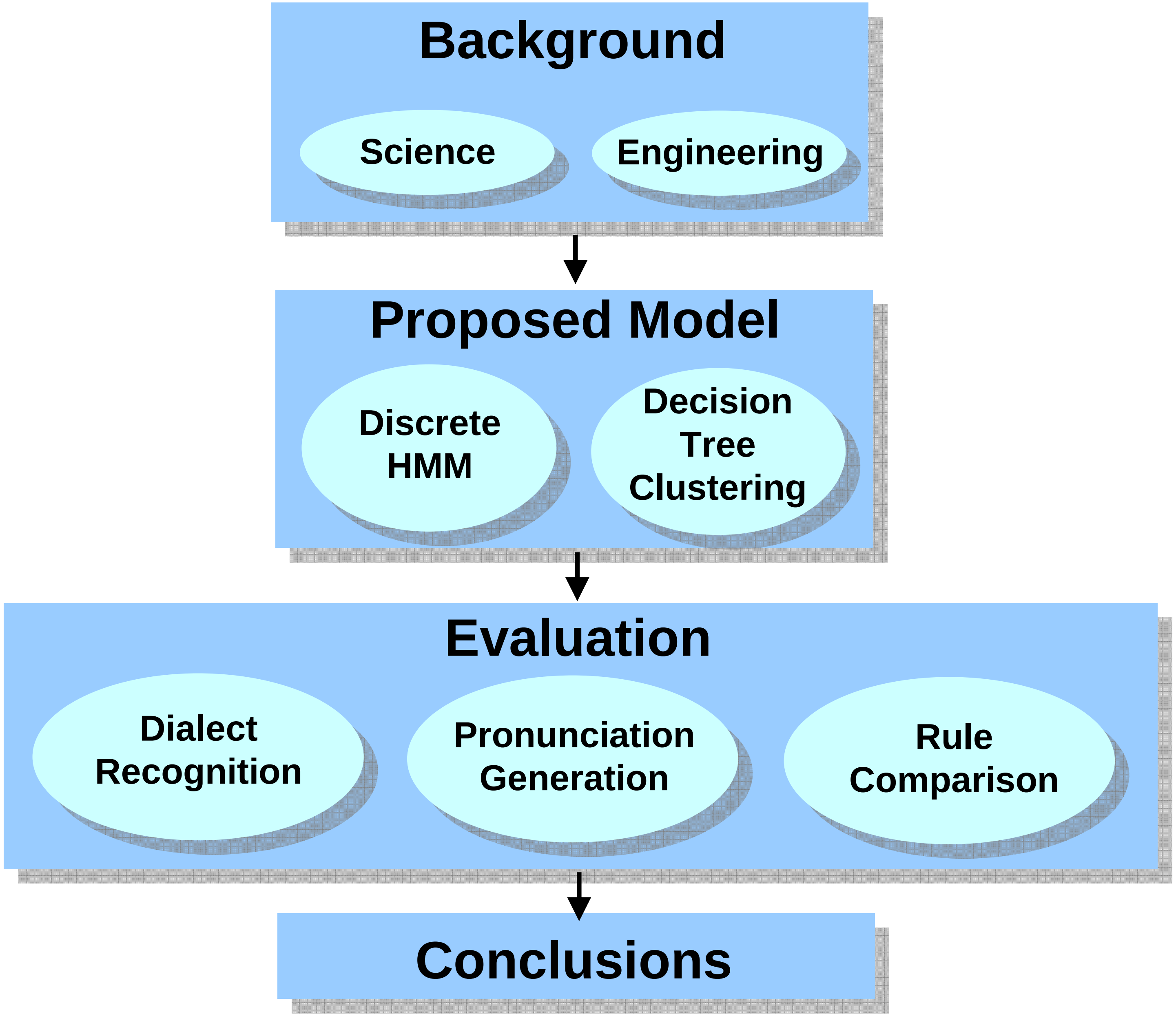
Nancy Chen, Wade Shen, Joseph Campbell, Pedro Torres-Carrasquillo

MIT/ Lincoln Laboratory

May 24, 2011

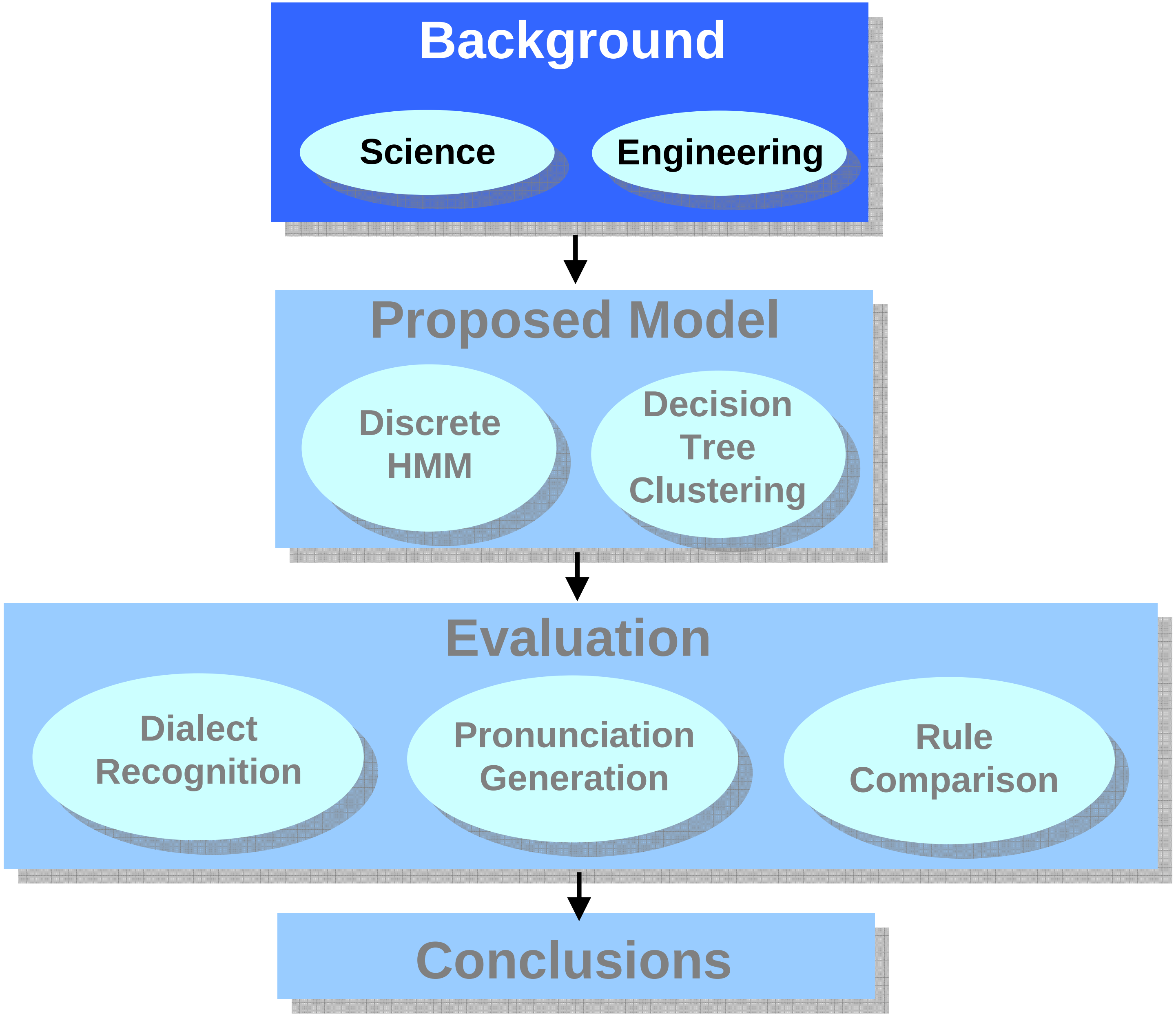
*This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Outline



Background
Science
Engineering
Proposed Model
Discrete HMM
Decision Tree Clustering
Evaluation
Dialect Recognition
Pronunciation Generation
Rule Comparison
Conclusions

Outline



Background
Science
Engineering
Proposed Model
Discrete HMM
Decision Tree Clustering
Evaluation
Dialect Recognition
Pronunciation Generation
Rule Comparison
Conclusions

Dialect Research

Speech Science

<u>Sociolinguistics</u>

Analyze phonetic rules manually

Dialect Research

Speech Science

<u>Sociolinguistics</u>

Analyze phonetic rules manually

Speech Technology

<u>Automatic dialect recognition</u>

Not explicitly learning rules

Informative Dialect Recognition

Bridges the Gap between Speech Science and Technology

Speech Science

Sociolinguistics

Analyze phonetic rules manually

Speech Technology

Automatic dialect recognition

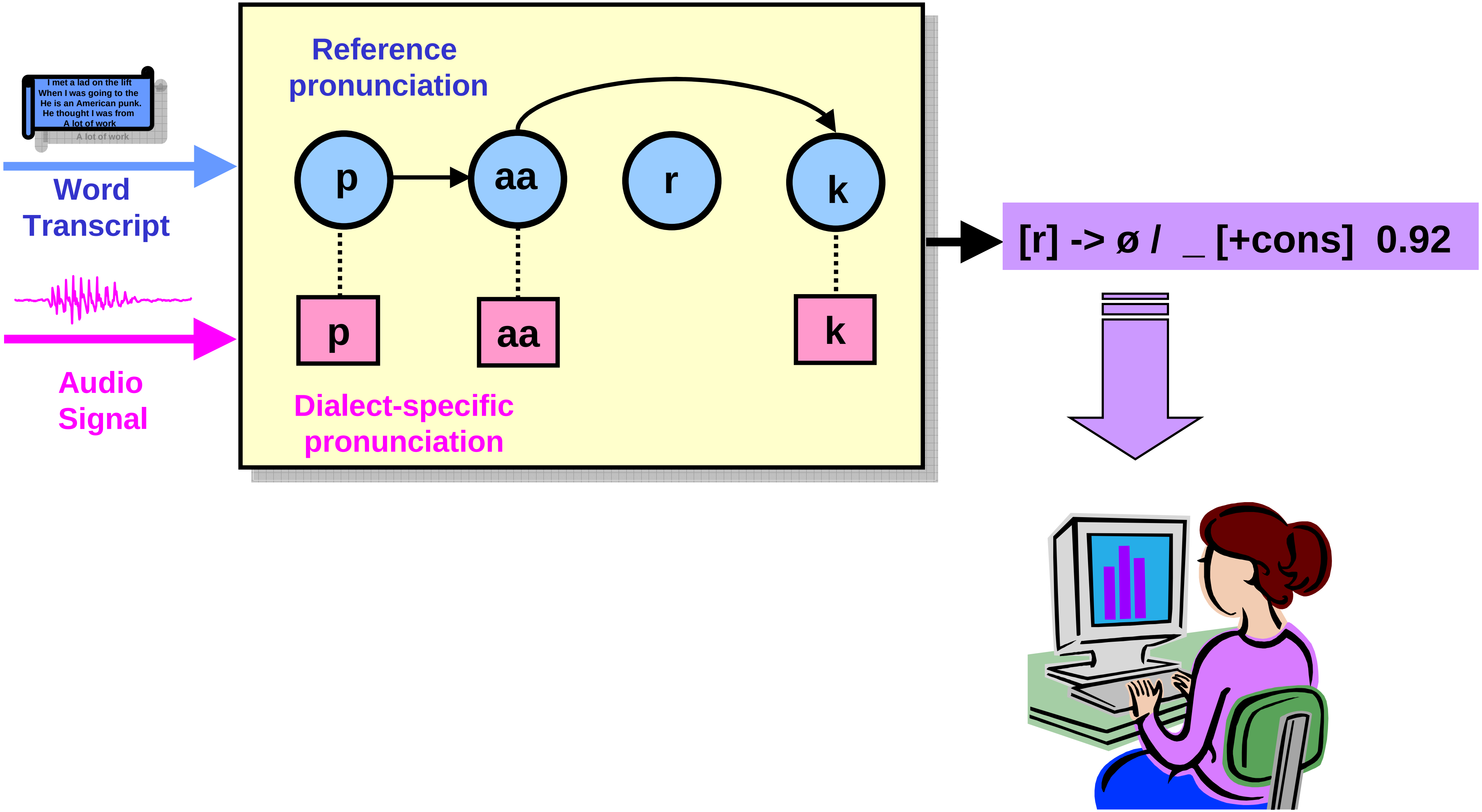
Not explicitly learning rules

Our Work

Informative Dialect Recognition

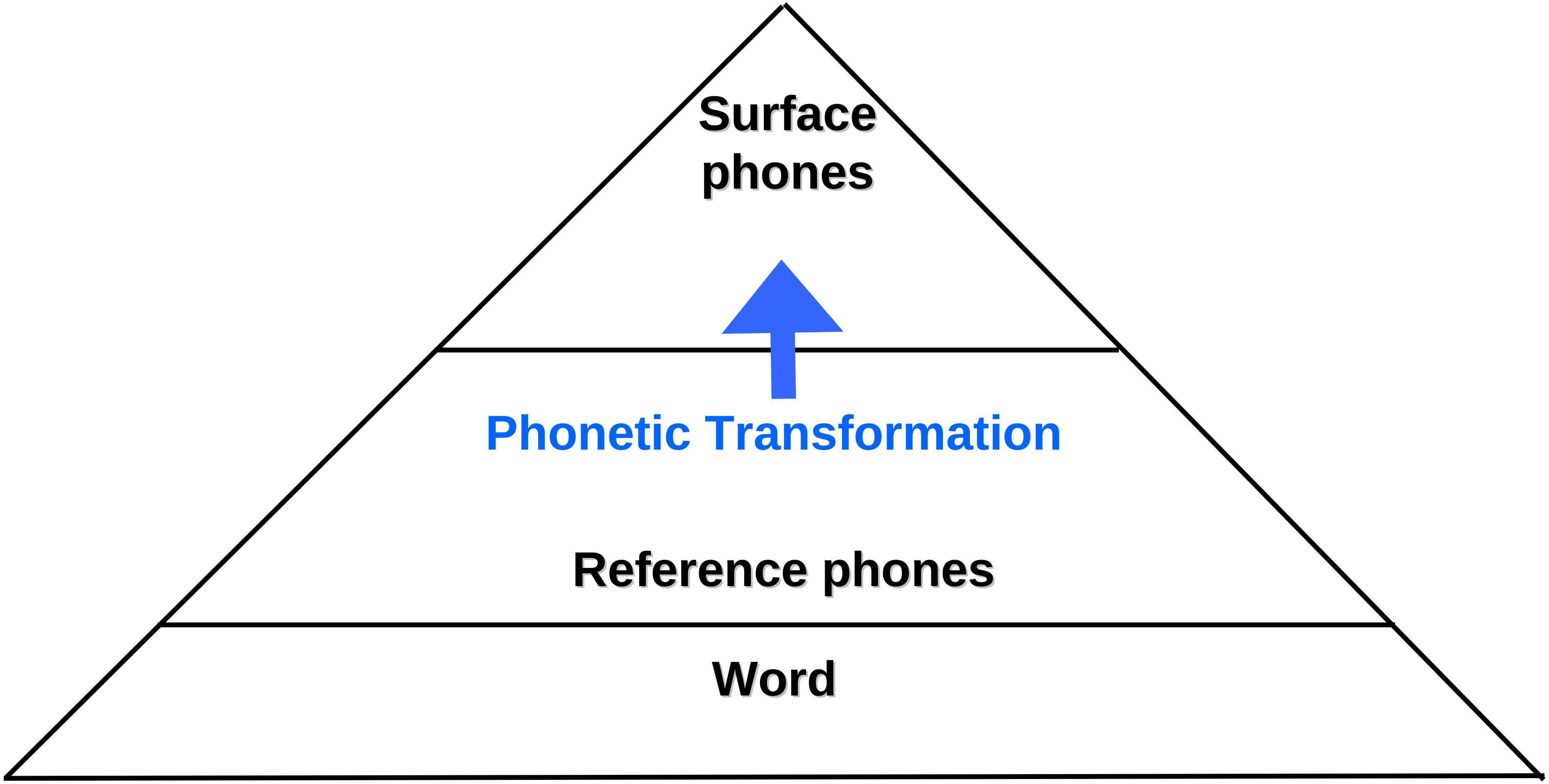
Automated system explicitly learns phonetic rules to inform human analyst

Informative Dialect Recognition



Phonetic Transformation

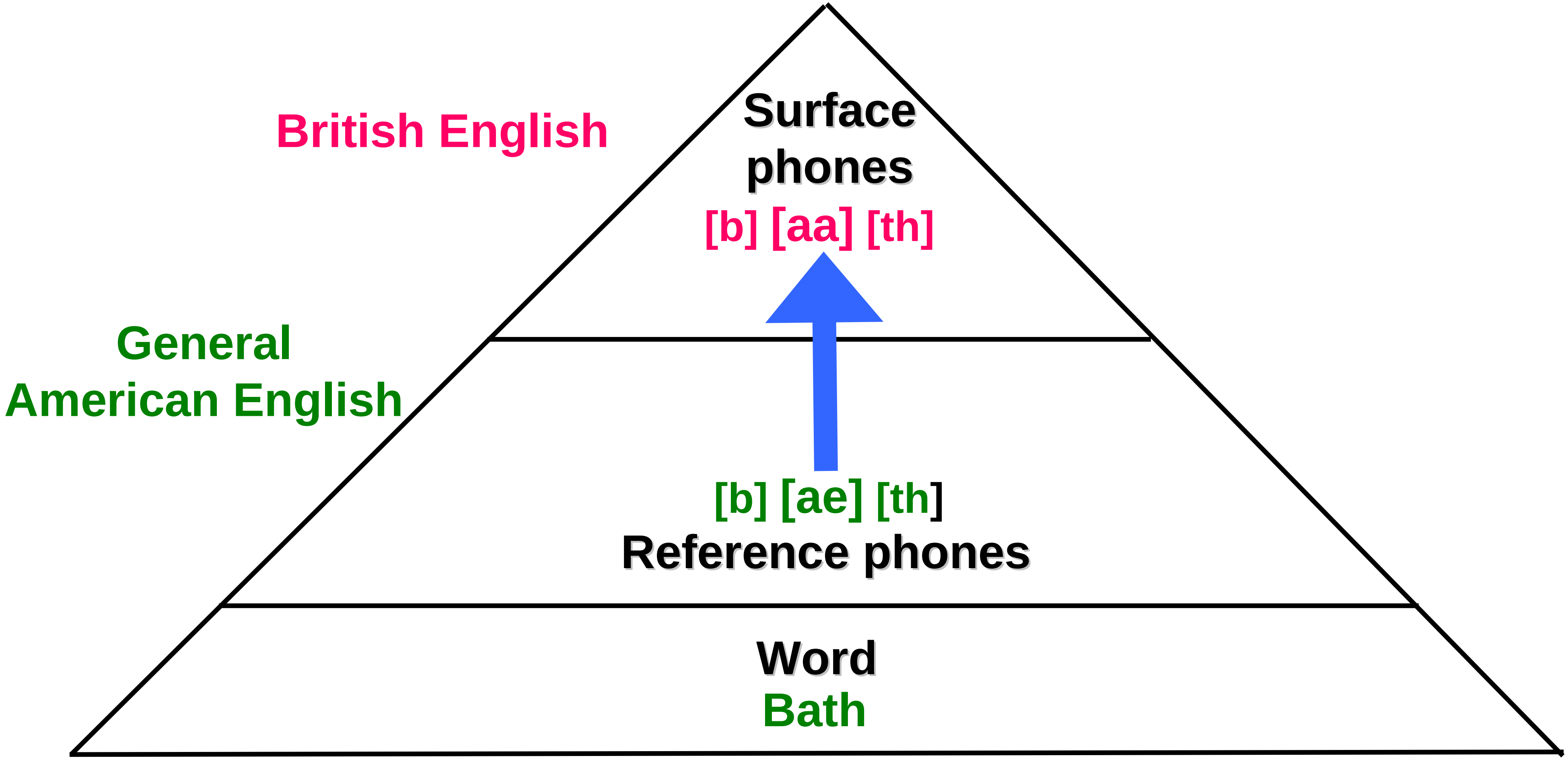
How phones in reference dialect is mapped to phones in dialect of interest



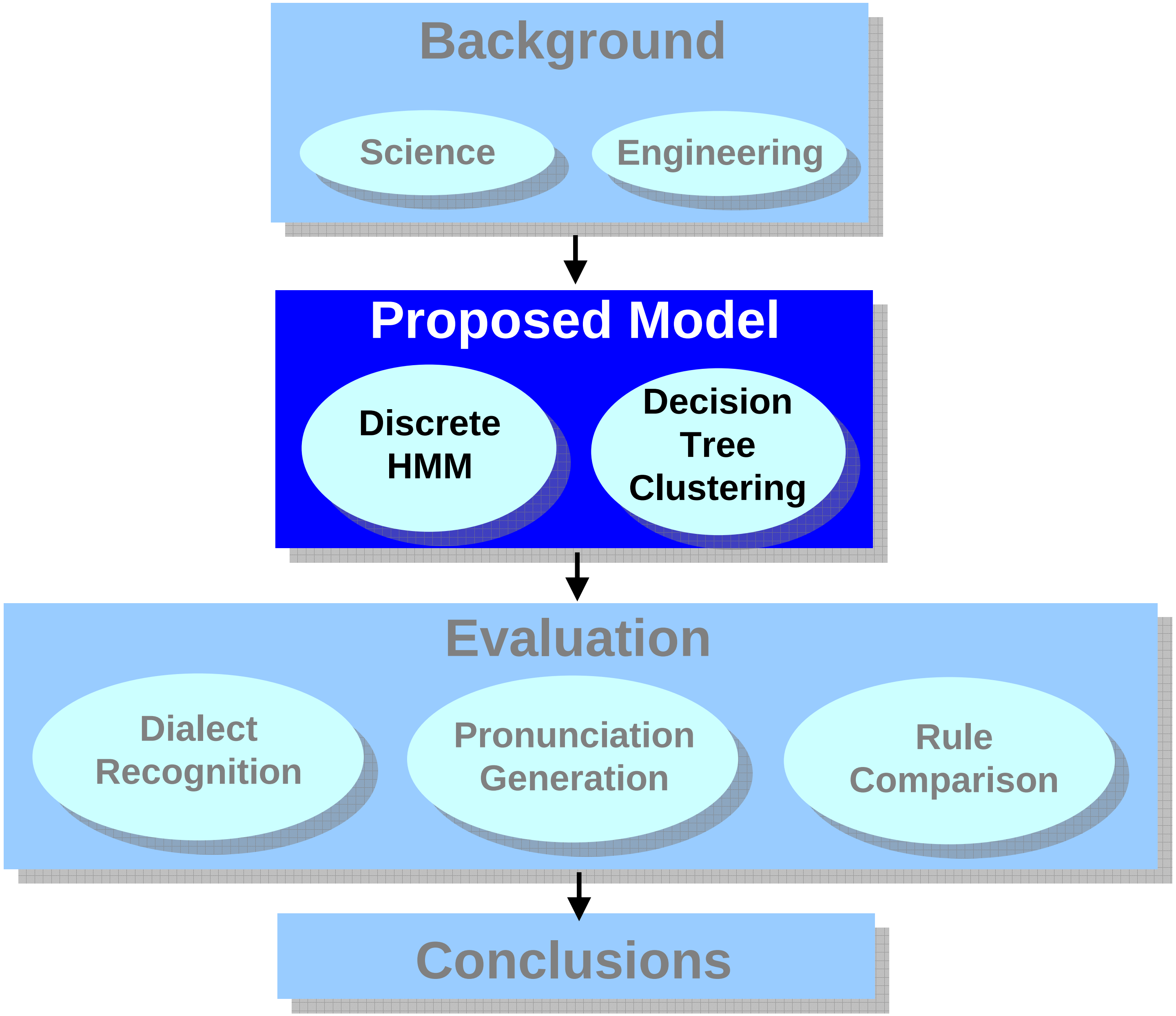
Phonetic Transformation

[ae] substitution

How phones in reference dialect is mapped to phones in dialect of interest



Outline


Background
Science
Engineering
Proposed Model
Discrete HMM
Decision Tree Clustering
Evaluation
Dialect Recognition
Pronunciation Generation
Rule Comparison
Conclusions

PPM: Phonetic Pronunciation Model

Characterizing phonetic transformations

- When a dialect is compared to a reference dialect, what kinds of substitutions, insertions, deletions occur?
- Where do they occur?
- How often do they occur?

Discrete hidden Markov model
Align reference phones
with surface phones

↓

Decision tree clustering
Generalize phonetic rules

Phonetic Transformations

Substitution: Trap/bath split

Word	bath
Reference phones *American*	b ae th
Surface phones *British*	b aa th

Deletion: Non-rhoticity

Word	park
Reference phones *American*	p aa r k
Surface phones *British*	p aa k

Insertion: Intrusive r

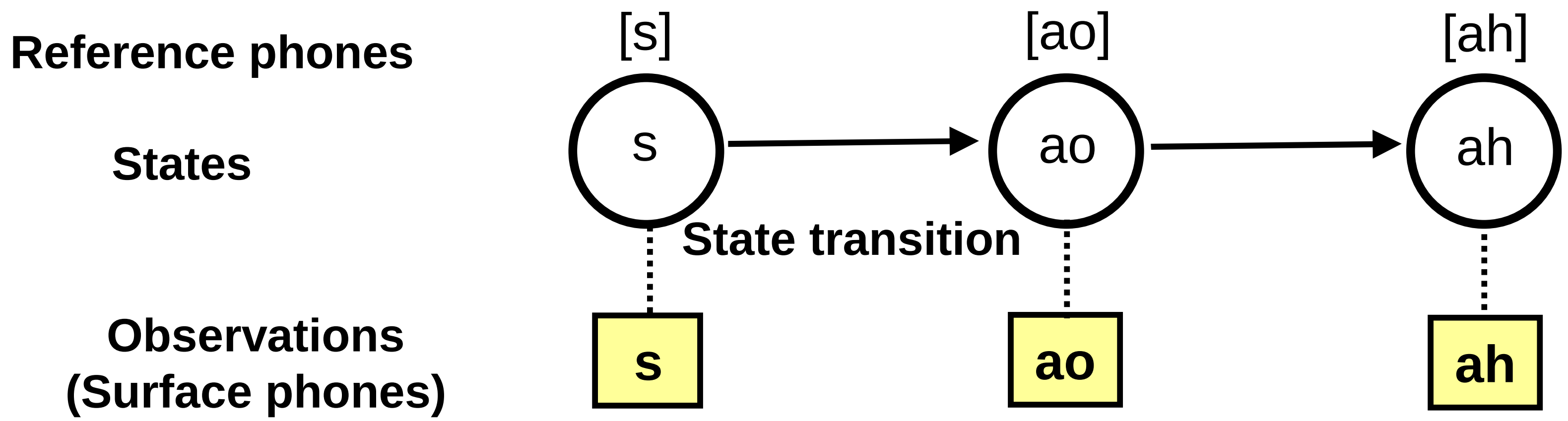
Word	saw a (film)
Reference phones *American*	s ao ah
Surface phones *British*	s ao r ah

Hidden Markov Model (HMM)

match

Traditional

Saw a (film)

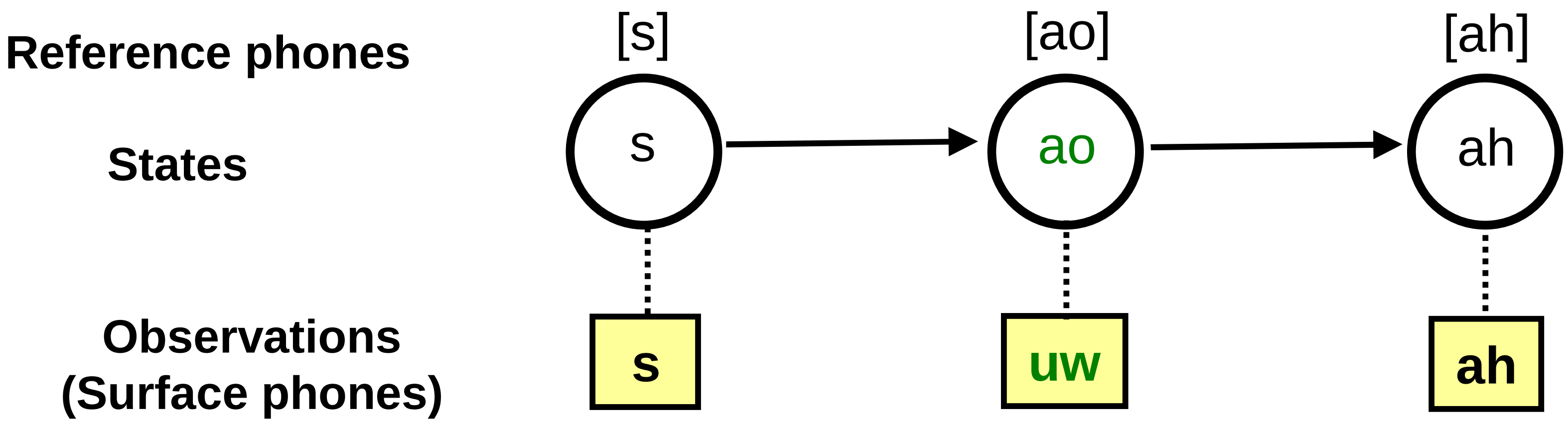


Hidden Markov Model (HMM)

substitution

Traditional

Saw a (film)



Hidden Markov Model (HMM)

insertion?

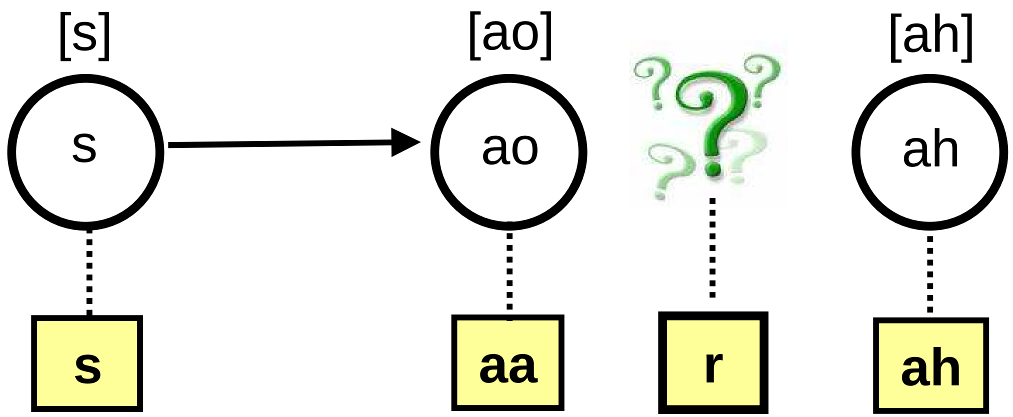
Traditional

Saw a (film)

Reference phones

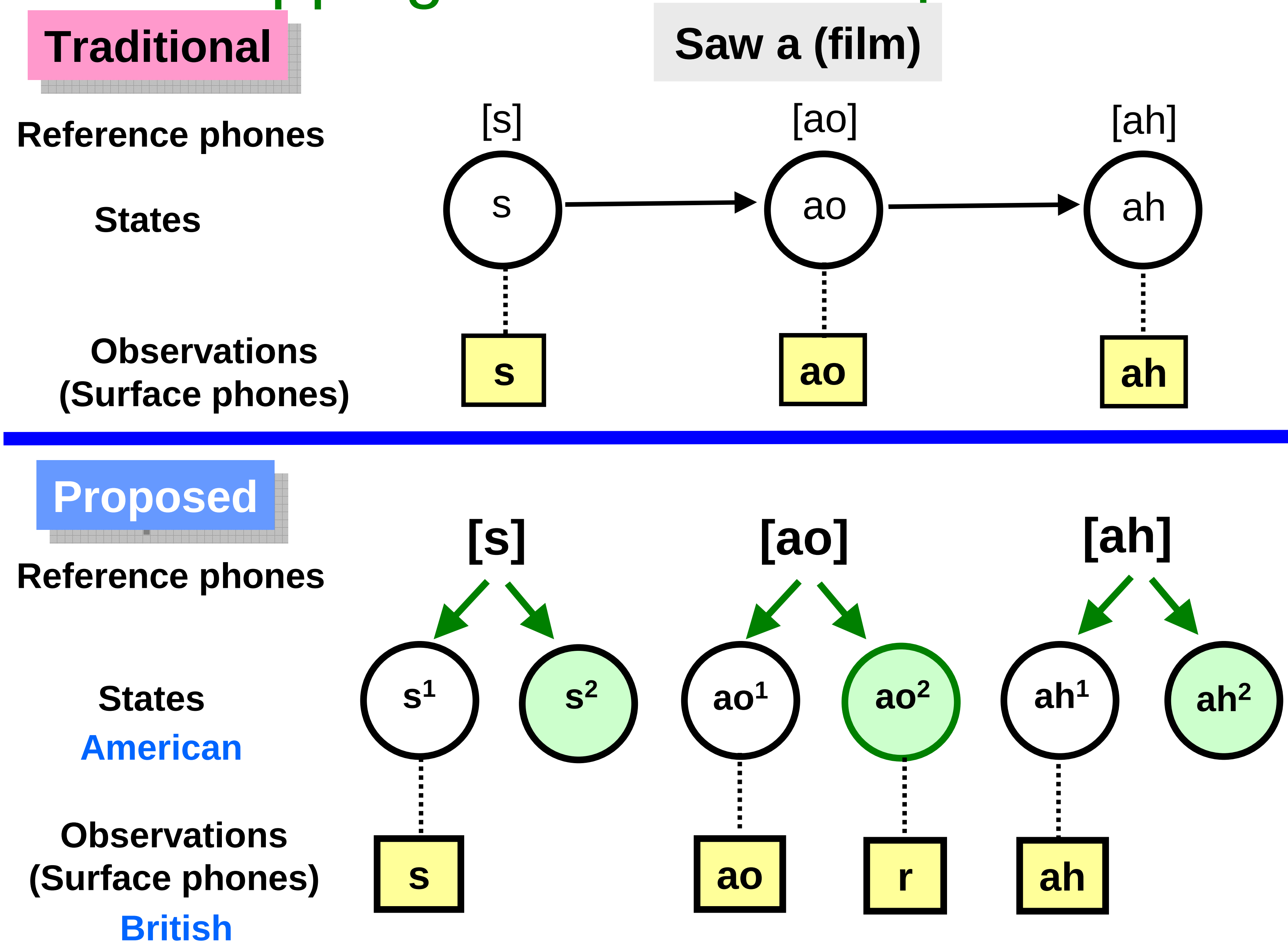
States

Observations (Surface phones)



Hidden Markov Model (HMM)

1-2 mapping btw reference phones and states



Hidden Markov Model (HMM)

insertion state & insertion arc

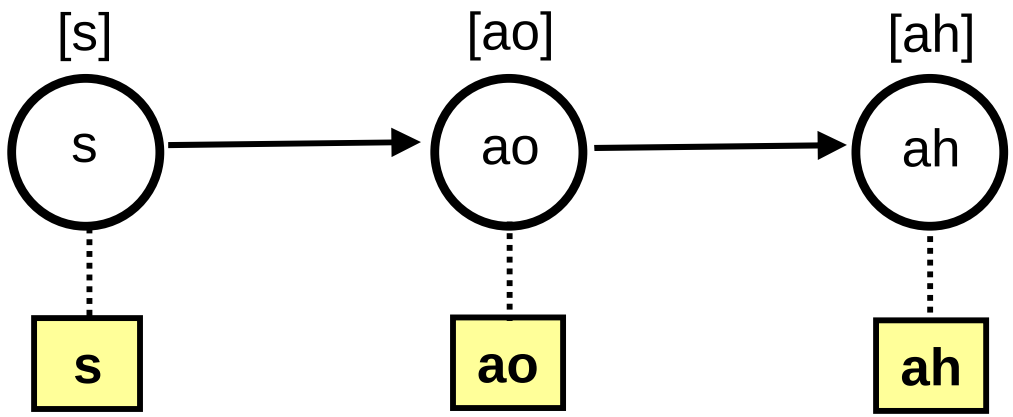
Traditional

Saw a (film)

Reference phones

States

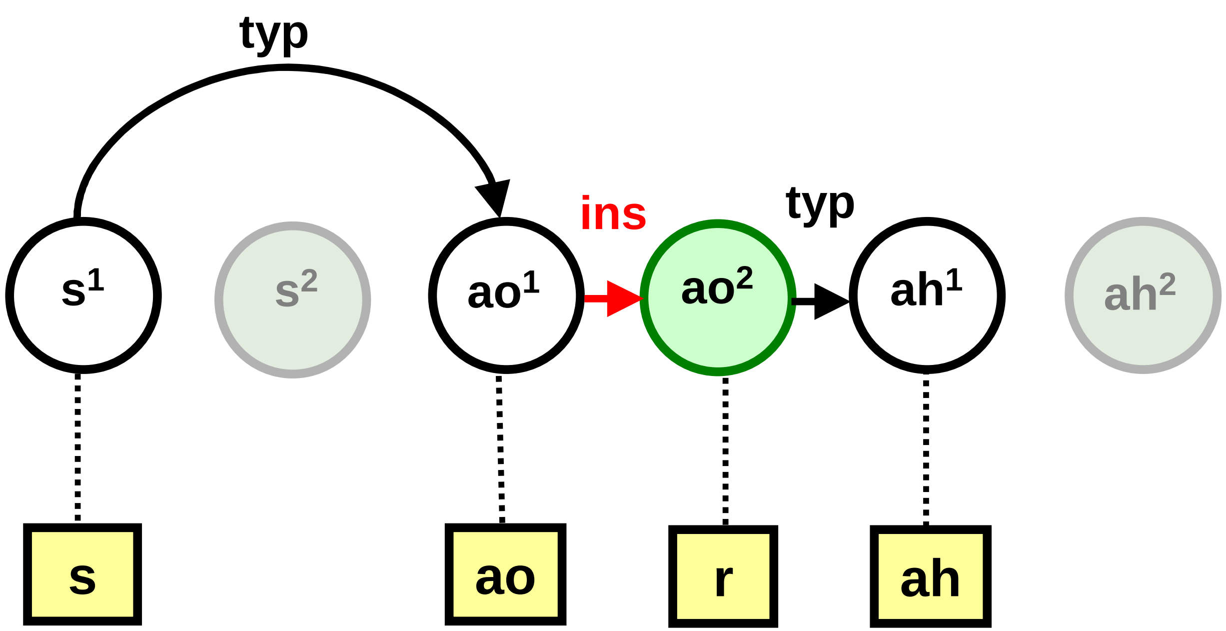
Observations (Surface phones)



Proposed

States
American

Observations (Surface phones)
British

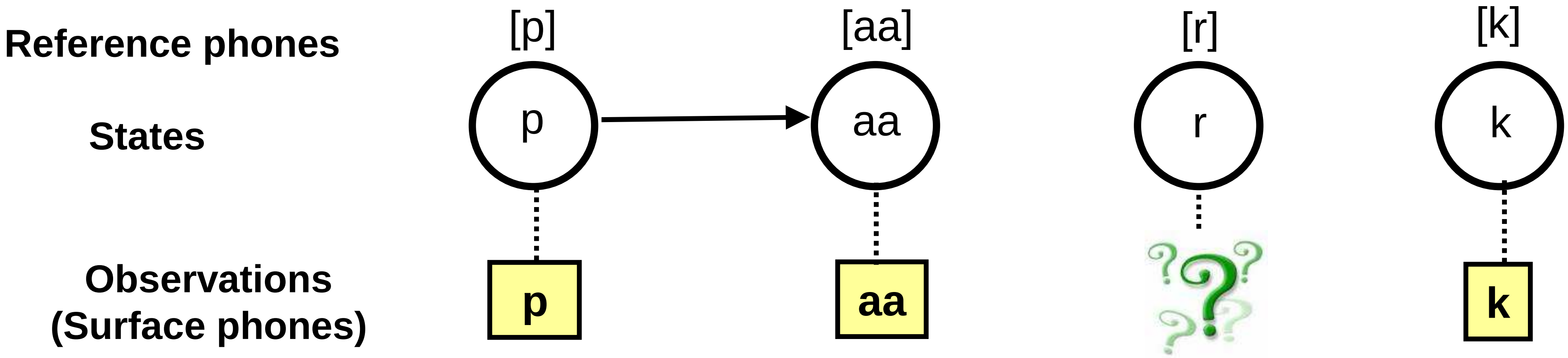


Hidden Markov Model (HMM)

deletion?

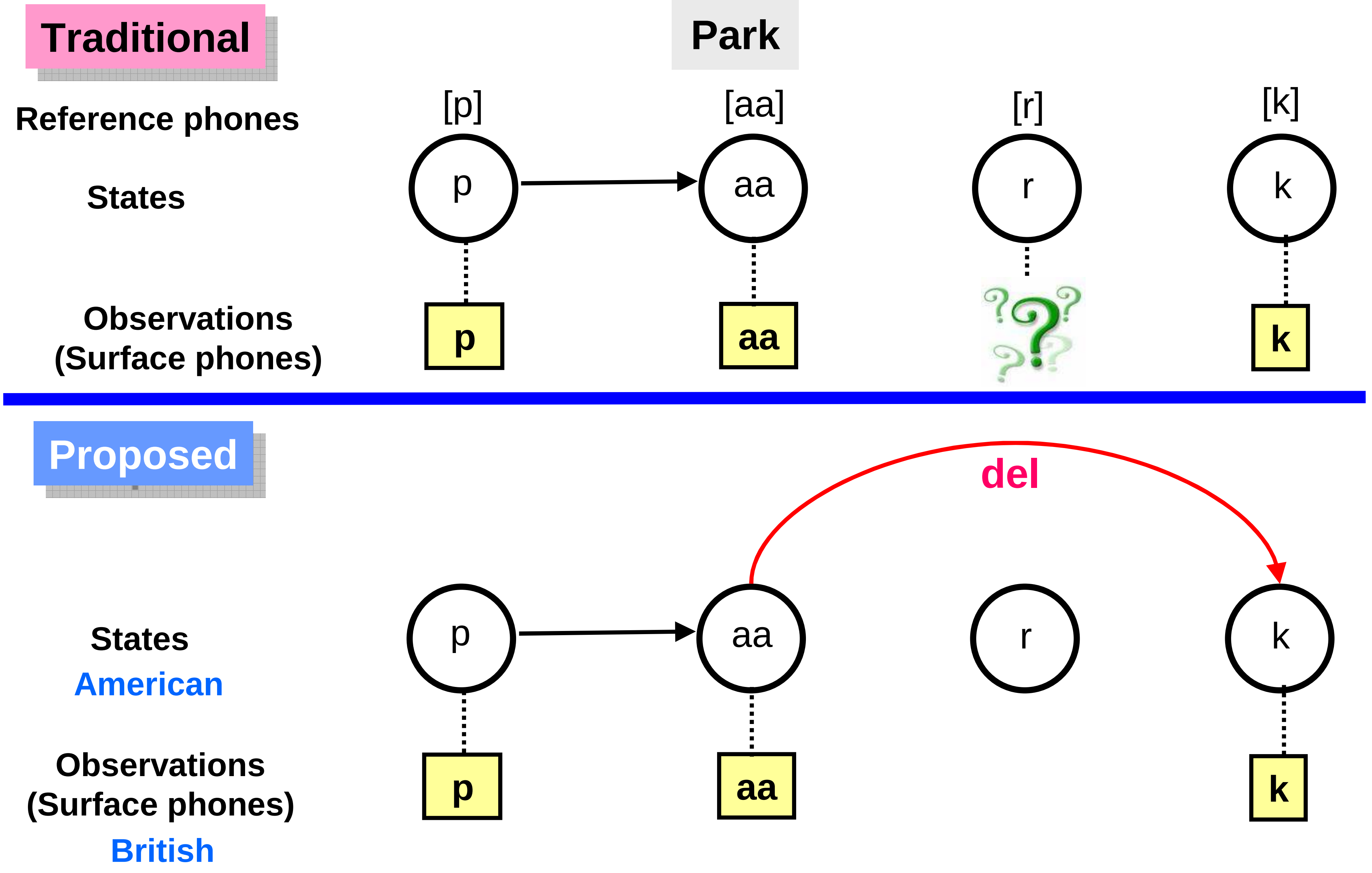
Traditional

Park



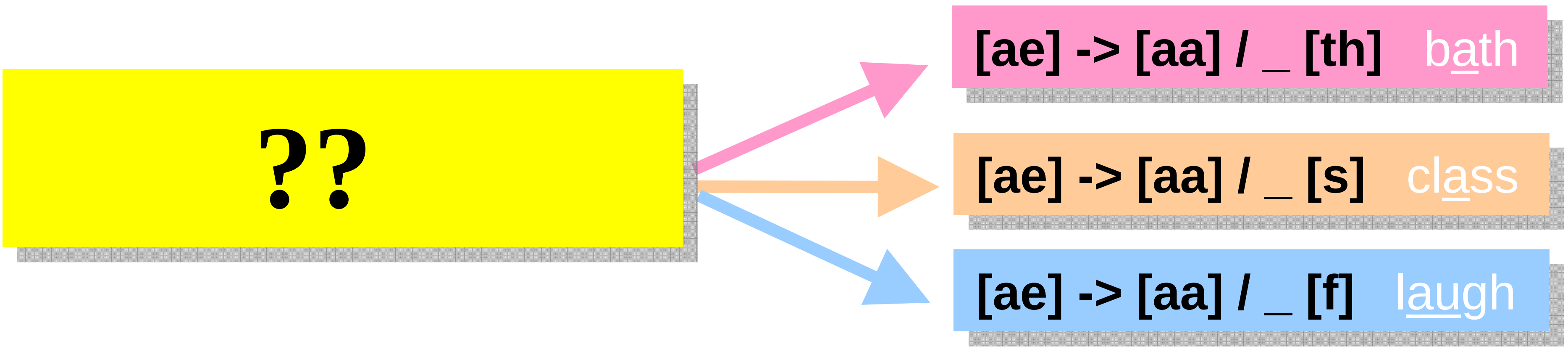
Hidden Markov Model (HMM)

deletion arc



Generalizing Rules

What is the underlying rule of [ae] transforming to [aa]??



Generalizing Rules

Underlying rule can be found using *Decision-Tree Clustering*

Decision Tree Clustering

[ae] -> [aa] / _ [voiceless fricative]

Underlying Rule

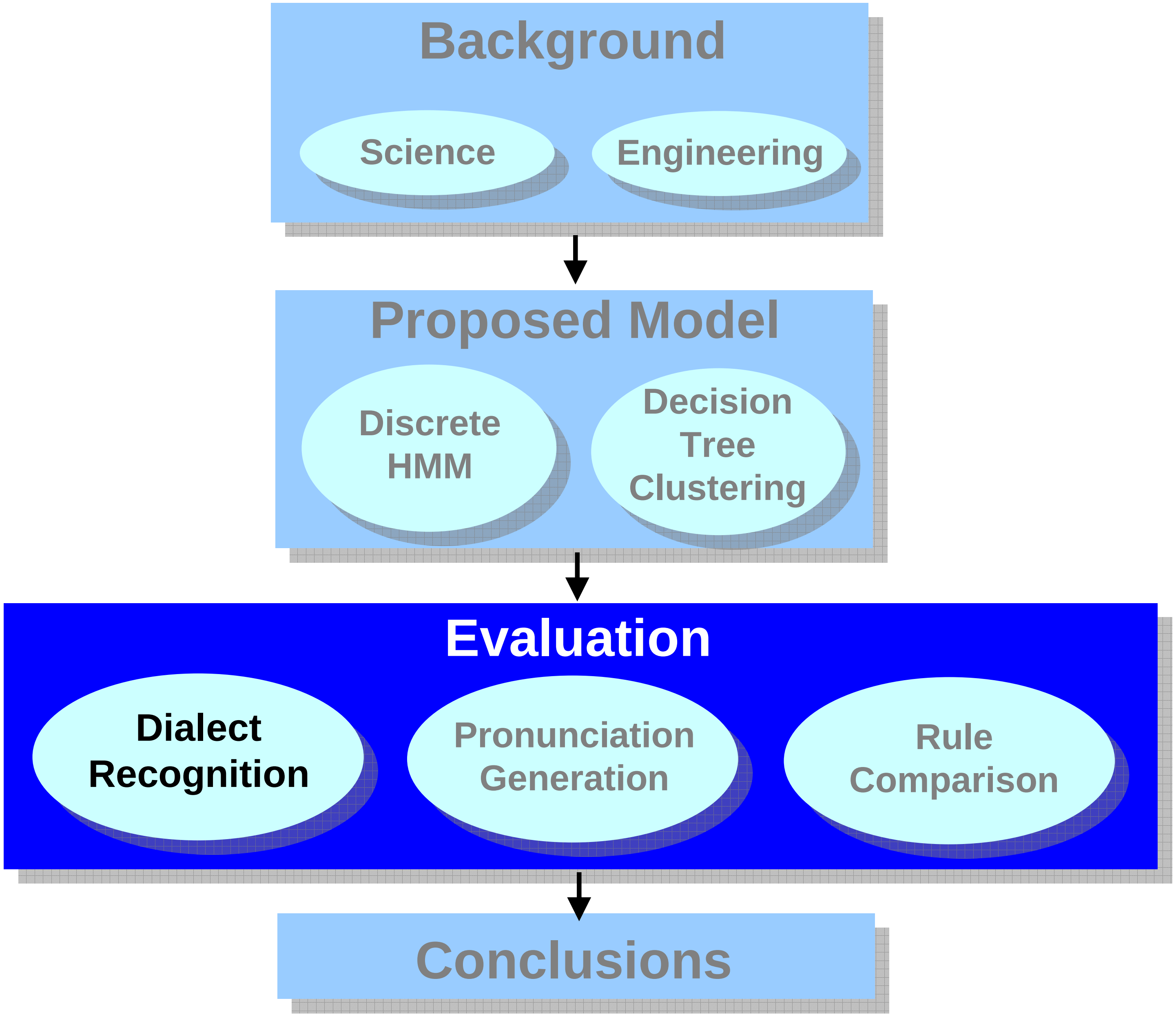
[ae] -> [aa] / _ [th] bath

[ae] -> [aa] / _ [s] class

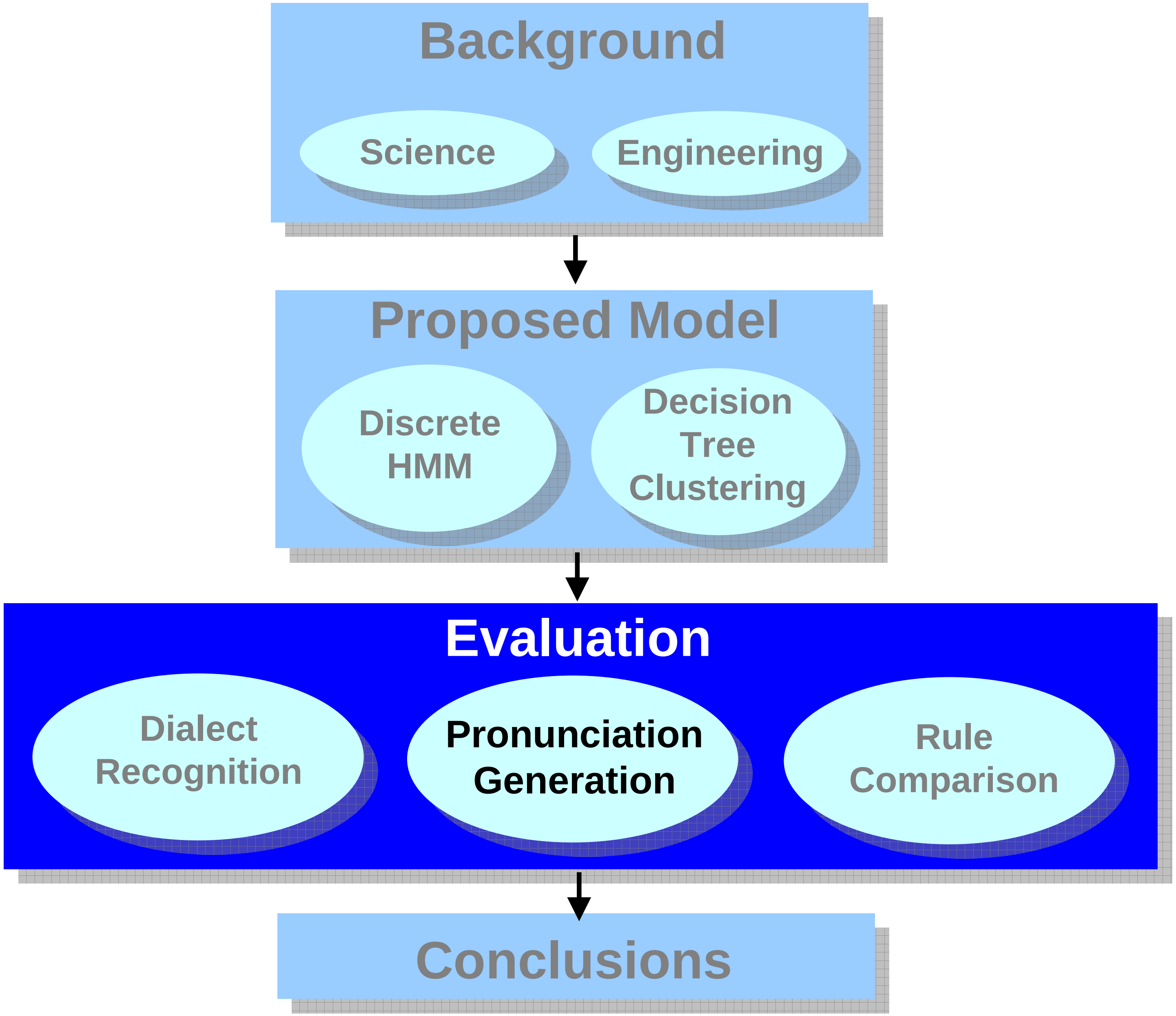
[ae] -> [aa] / _ [f] laugh

Details

Outline


Background
Science
Engineering
Proposed Model
Discrete HMM
Decision Tree Clustering
Evaluation
Dialect Recognition
Pronunciation Generation
Rule Comparison
Conclusions

Outline



Background
Science
Engineering
Proposed Model
Discrete HMM
Decision Tree Clustering
Evaluation
Dialect Recognition
Pronunciation Generation
Rule Comparison
Conclusions

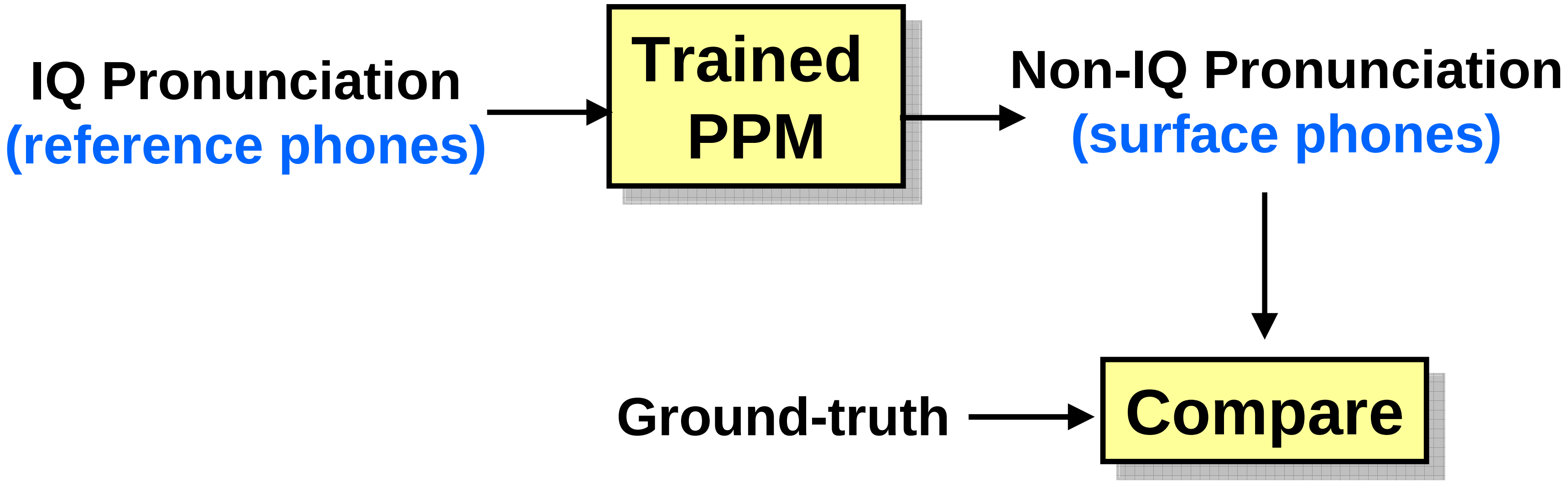
Corpus

- 5 Arabic dialects regions
  - UAE (AE), Egypt (EG), Iraq (IQ), Palestine (PS), Syria (SY)
- Conversational telephone speech
- IQ: reference dialect

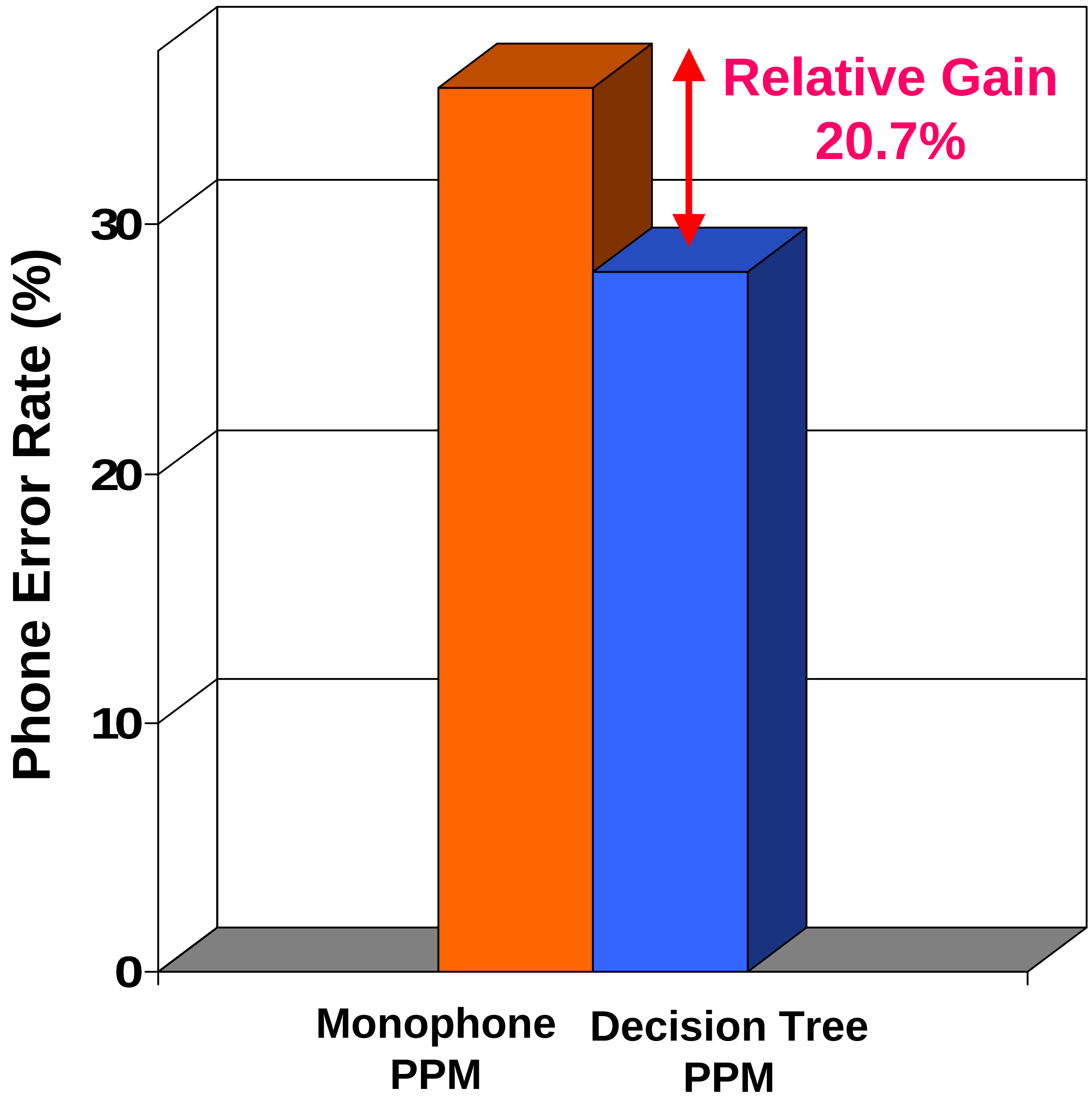
Data set	Speaker number	Duration
Train	276	46.25 hr
Dev	83	13.9 hr
Test	88	14.75

Generating Dialect-Specific Pronunciation

- Assumption
  - If trained model has learned rules correctly, then the model is able to convert IQ pronunciation to non-IQ pronunciation

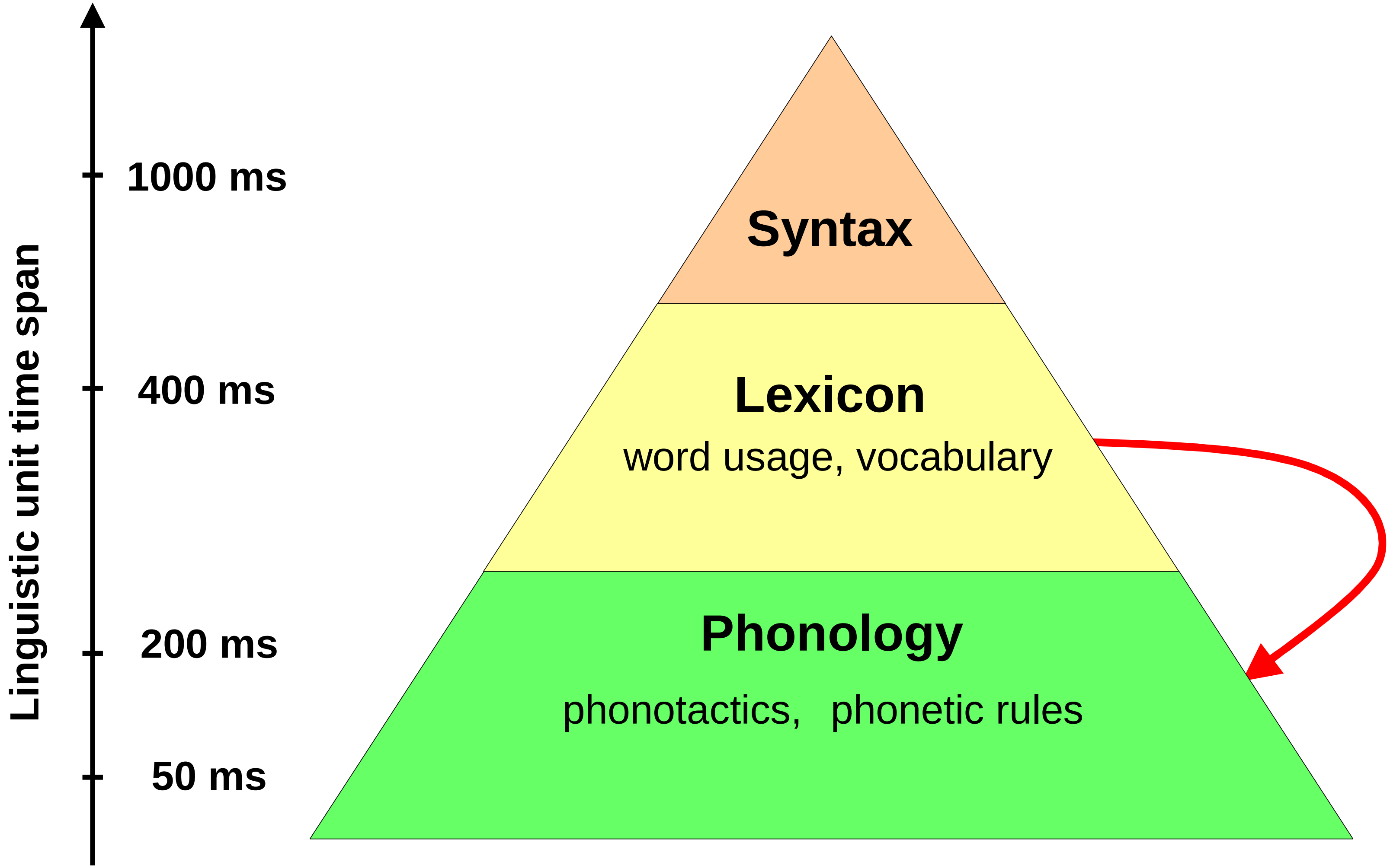


Proposed Model Improves Rule Recovery Rate



Word-Usage Differences

Word usage differences across Arabic dialects complicate evaluation of PPM



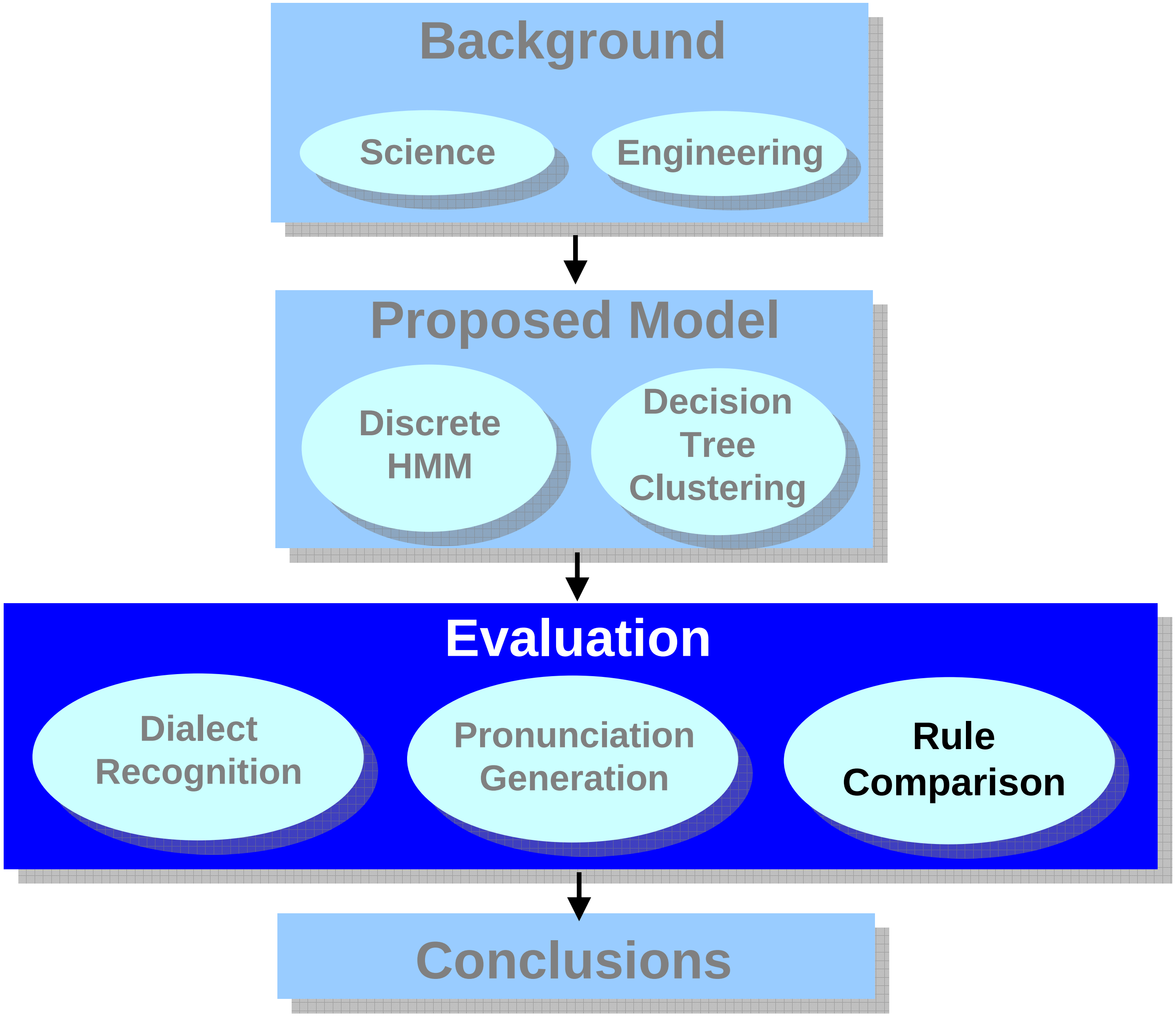
But…

Future Work Preview

- Phonetic Pronunciation Model (PPM) performs well on English corpora **w/o word-usage differences**

- ***Coming soon: Chen et. al, 2011 Interspeech***
  - Extensions of PPM
  - Multiple English corpora

Outline


Background
Science
Engineering
Proposed Model
Discrete HMM
Decision Tree Clustering
Evaluation
Dialect Recognition
Pronunciation Generation
Rule Comparison
Conclusions

Examples of Learned Rules from PPM

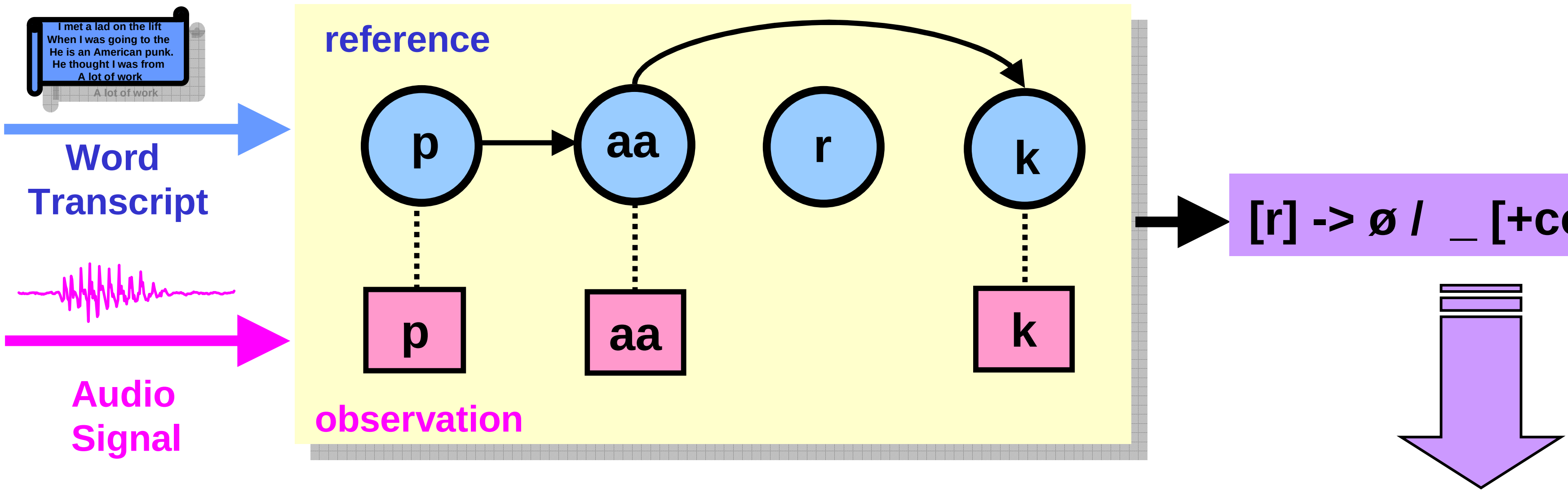
PPM quantifies occurrence frequency of rules

Literature		Proposed System		
Linguistic Description	**Dialect**	**Learned Rule**	**Prob**	**Dialect**
Palatal voiced affricate becomes palatal approximant	AE	[dZ] -> [j] /_ [+syl]	0.32	AE
Palatal voiced affricate becomes voiced stop	EG	[dZ] -> [d]	0.25	EG
Vowel [o] exists	IQ	[o:] -> [a]	0.28; 0.27; 0.32; 0.27	AE, EG, PS, SY
Interdental fricatives become stops	EG PS SY	[th] -> [t] /_ [-short]	0.60	EG
		[th] -> [t] / [-low] _ [+short]	0.59	EG
		[th] -> [t]	0.42; 0.43	PS, SY
		[dh] -> [d]	0.24; 0.29	PS, SY
		[dh] -> [d] / [-front] _	0.33	EG

Conclusions

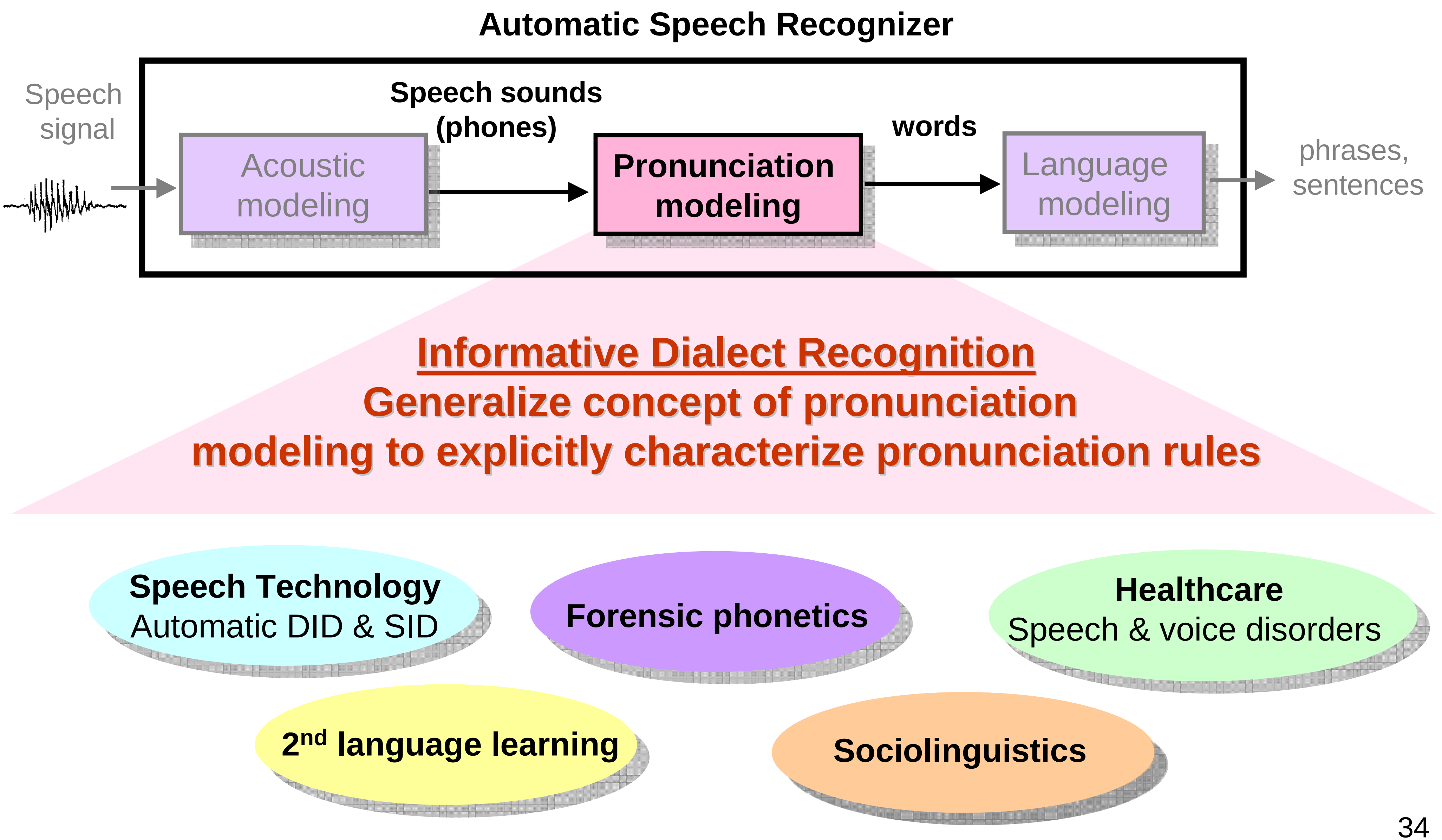
- Informative Dialect Recognition: ***automatic*** yet ***informative*** approach in analyzing dialects
- Mathematical framework characterizes phonetic transformations across dialects in ***explicit*** manner
- Proposed system postulates rules from ***large corpora*** to discover, refine, and ***quantify*** rules

Informative Dialect Recognition



Informative Dialect Recognition

Potential Applications



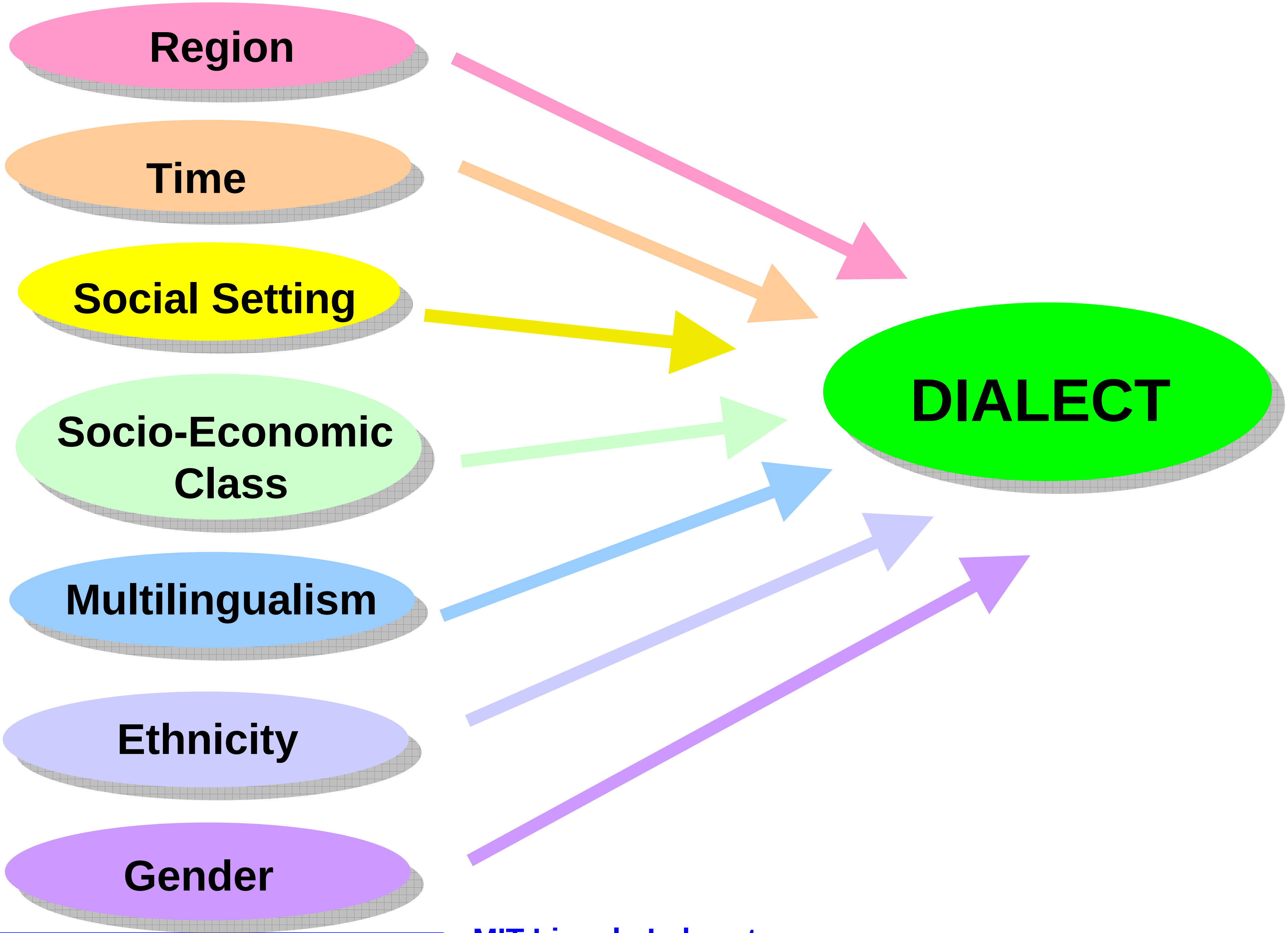
Back up

Outline

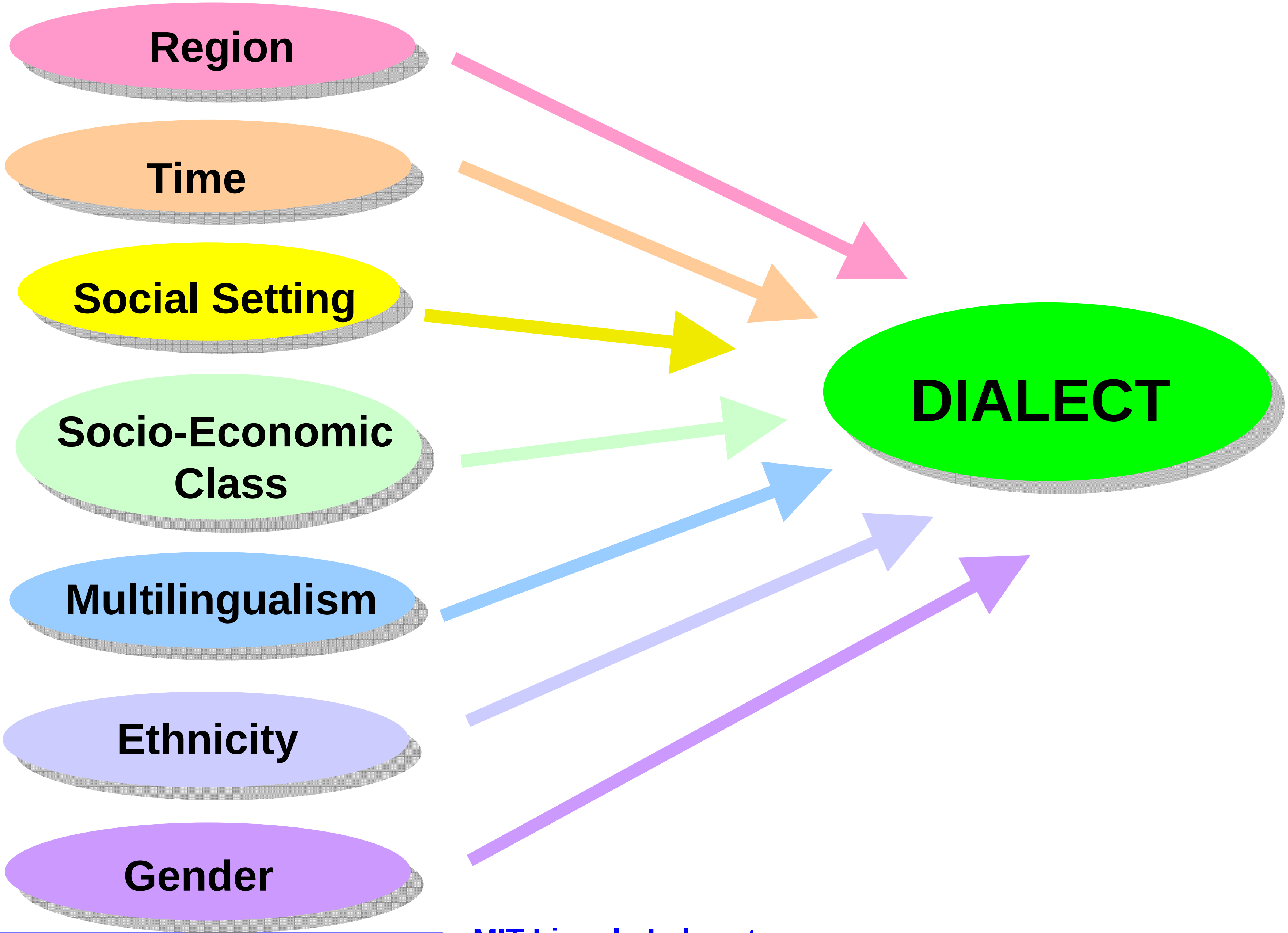
- Background
- Proposed Pronunciation Model
- Evaluation
  - Dialect Recognition
  - Generating Dialect-Specific Pronunciation
  - Rule Analysis: Interpretation and Quantification
- Conclusions

Background

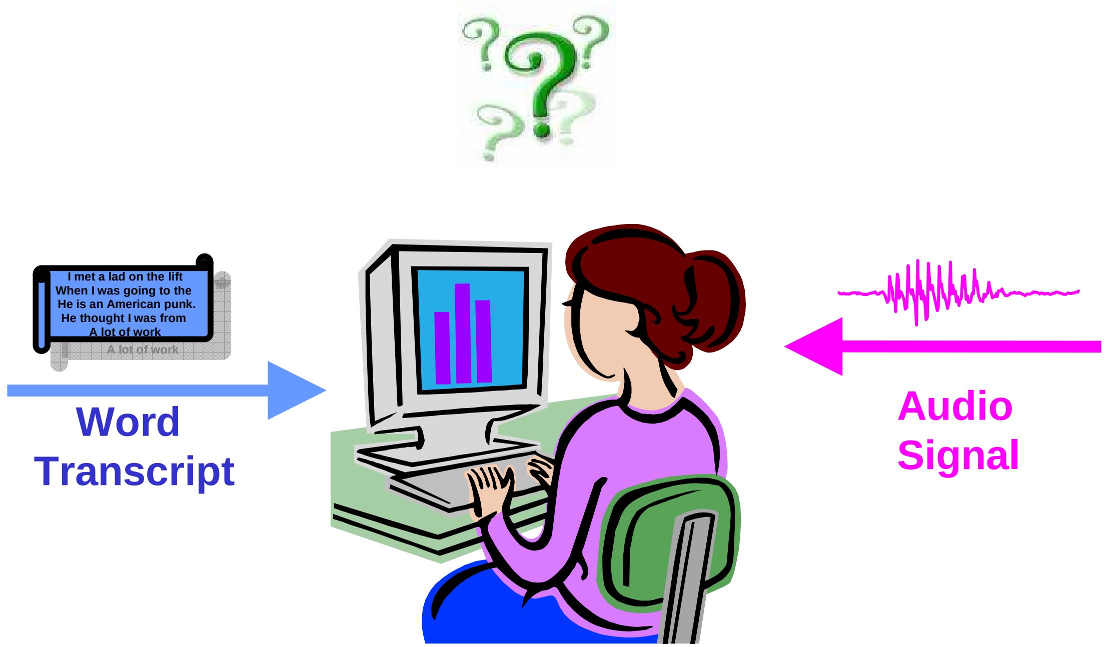
What Influences Dialects?



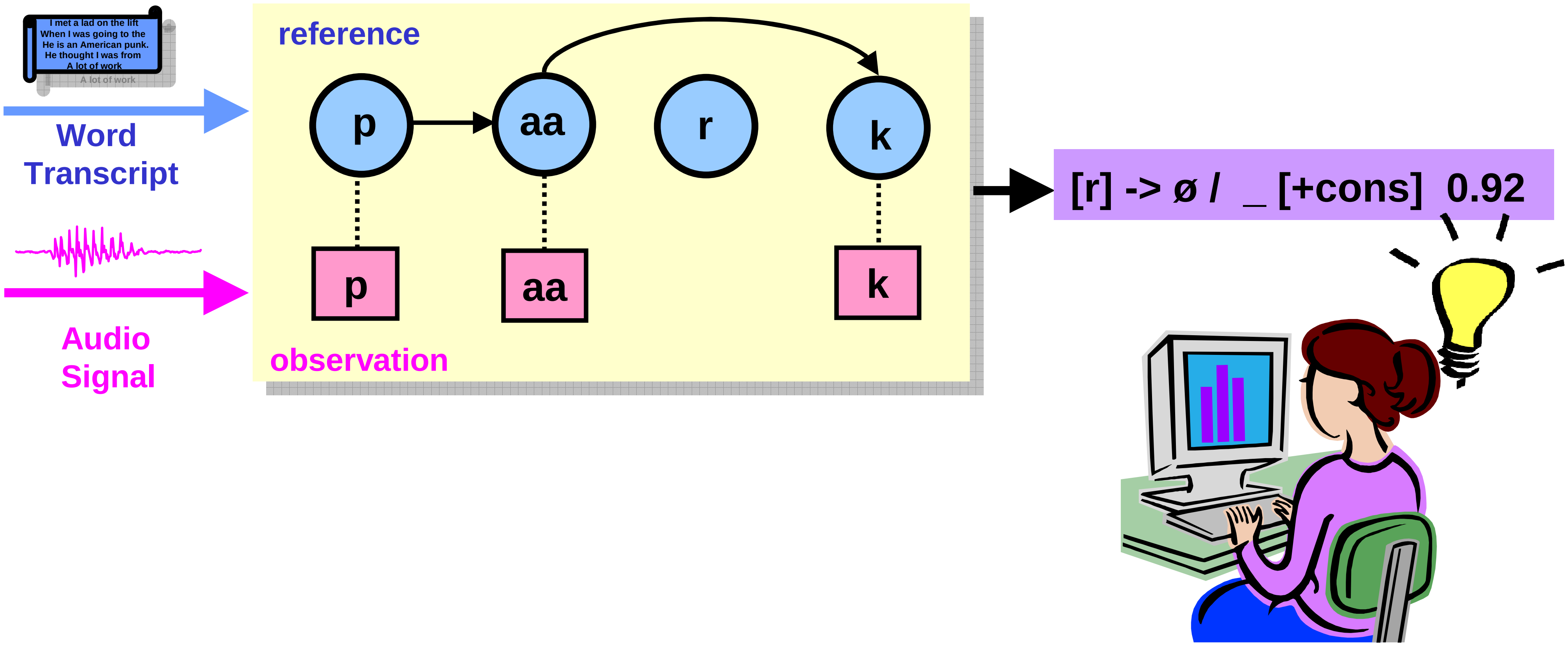
What Influences Dialects?



Traditional Analysis



Informative Dialect Recognition



Related Work

Our work generalizes ASR concepts to automatically learn rules

Primary Focus	Improve engineering performance	Automatically learn rules
	Automatic Dialect Recognition Richardson & Campbell (2009) Biadsy et al (2010)	*Informative Dialect Recognition* Chen et al (2010, 2011)

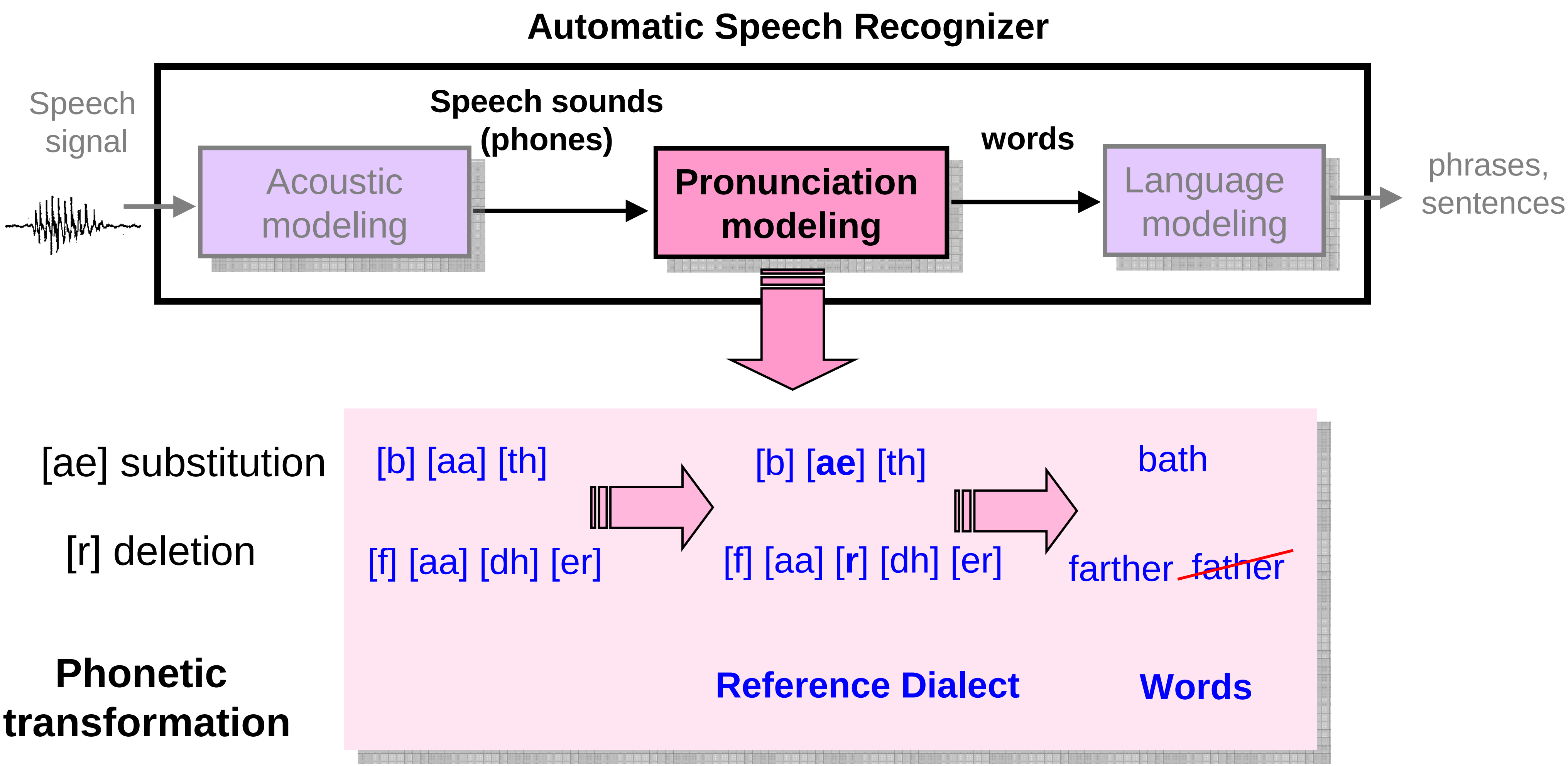
Related Work

Our work generalizes ASR concepts to automatically learn rules

purpose / method	**Improve speech analysis efficiency**	**Automatically learn rules**
Directly apply ASR tools	*Sociolinguistics* Evanini et al (2009) Yuan & Liberman (2009) *Computer-aided language learning* Kim et al (2004)	*Automatic Speech Recognition* Livescu & Glass (2000), Kim et al (2007)
Generalize ASR concepts	*Informative Dialect Recognition* Chen et al (2010, 2011)	*Informative Dialect Recognition* Chen et al (2010, 2011)

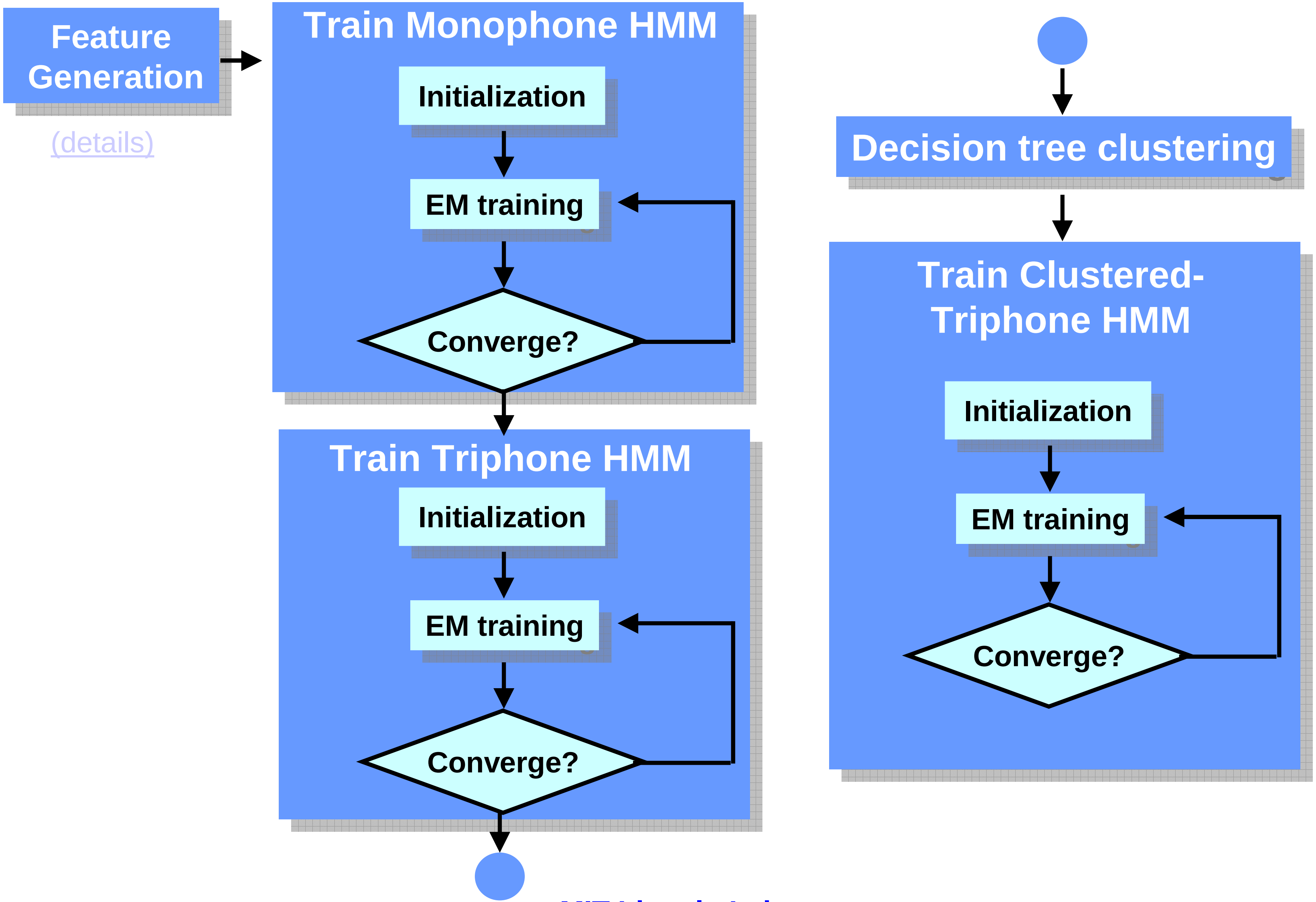
Applying Pronunciation Modeling to Dialect Analysis

Mapping between sound units and words



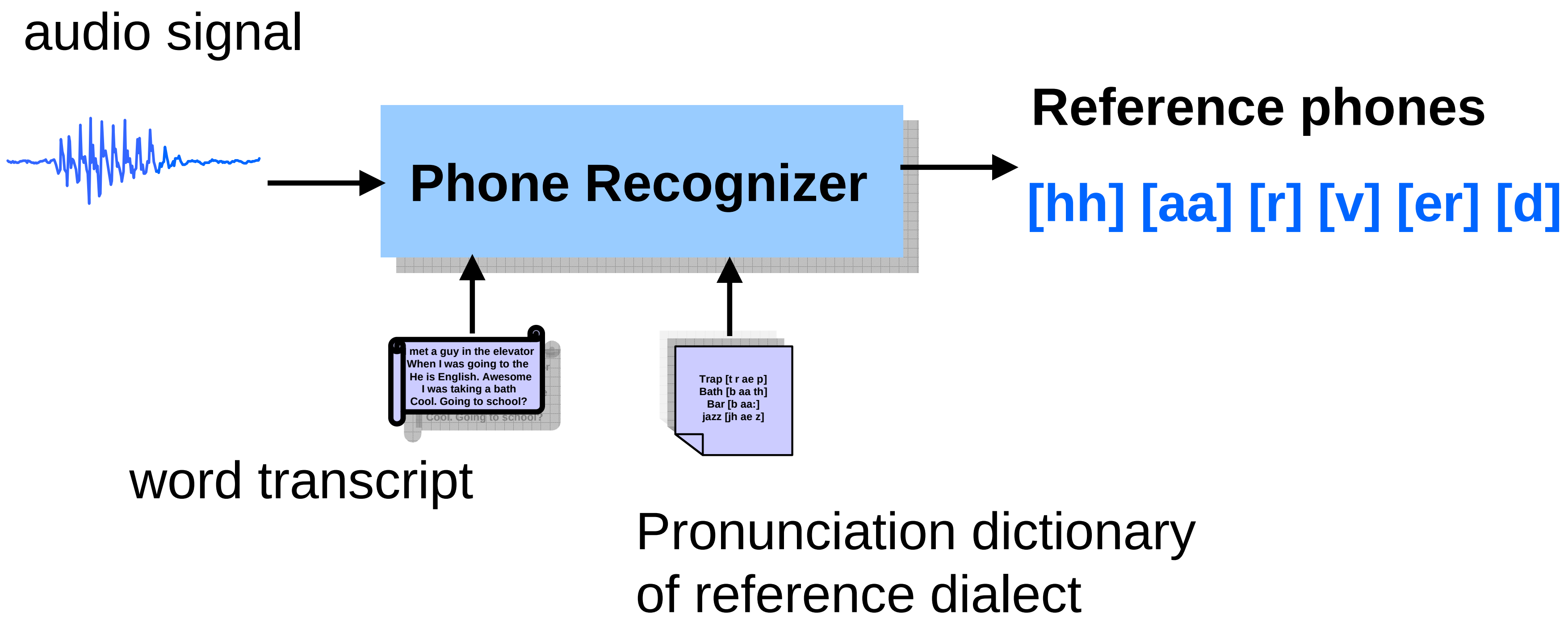
Implementation Details

Training
Phonetic Pronunciation Model (PPM)



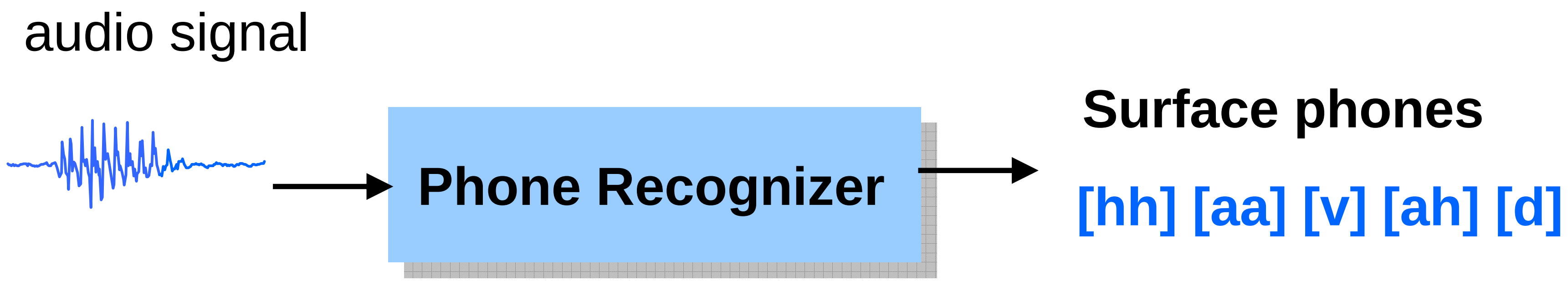
Feature Generation

Reference phones generated through force alignment



Feature Generation

Surface phones generated by phone recognition decoding

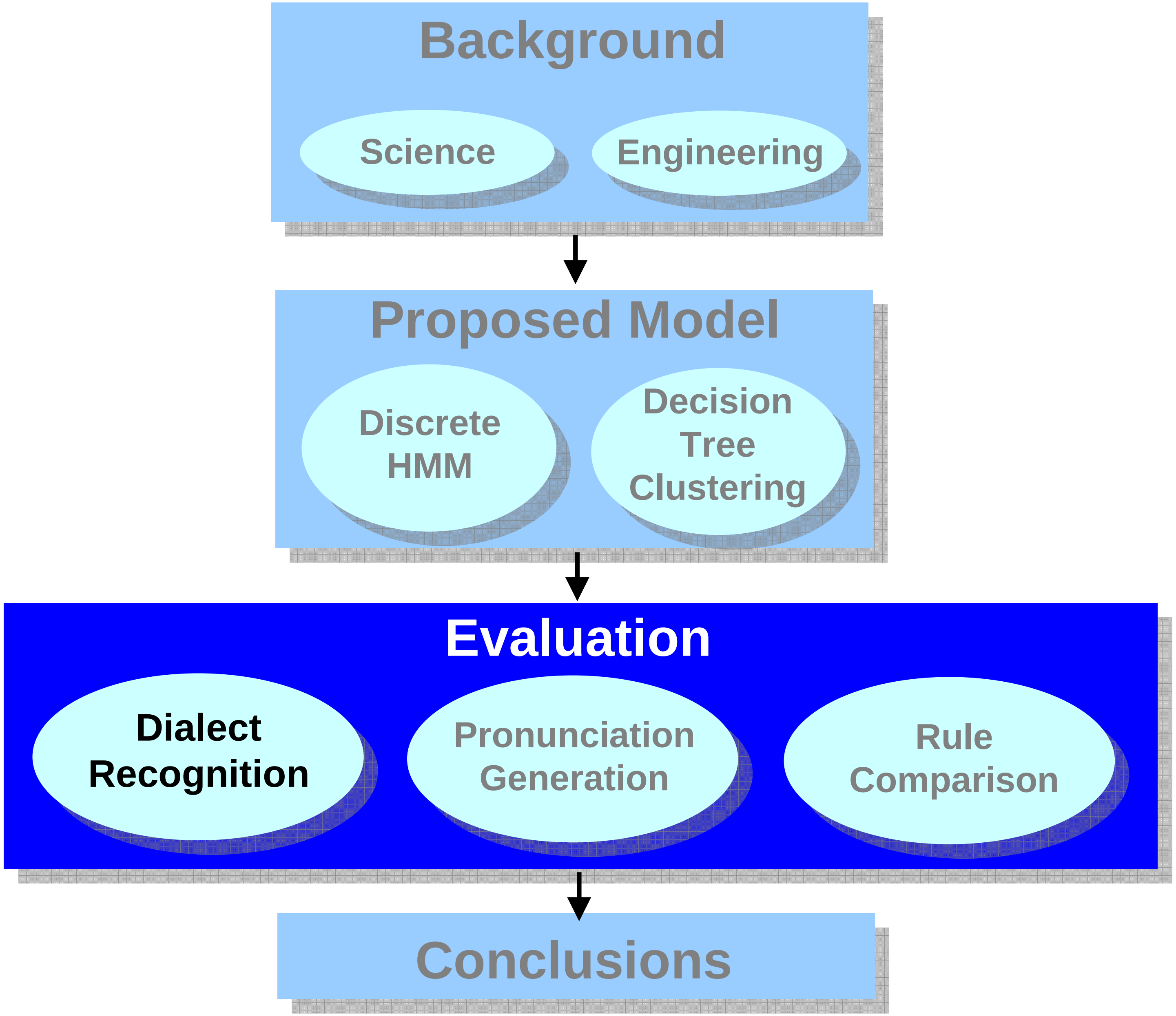


Current tying mechanism

- State clustering
  - Emissions of each state are used to train decision trees
  - Arcs and emissions are shared if the triphone states are tied
- Transition arc constraints
  - Deletion & typical arcs are destination independent
  - Insertion arcs are destination dependent by definition
  - Consecutive deletion not allowed, while consecutive insertions are allowed
- Assumptions
  - The phone before the deleted phone goes through phonetic transformation
    - Example: /park/ -> [p a: k]
  - The phone after the deleted phone does not characterize the deletion

Dialect Recognition

Outline


Background
Science
Engineering
Proposed Model
Discrete HMM
Decision Tree Clustering
Evaluation
Dialect Recognition
Pronunciation Generation
Rule Comparison
Conclusions

Experimental Setup

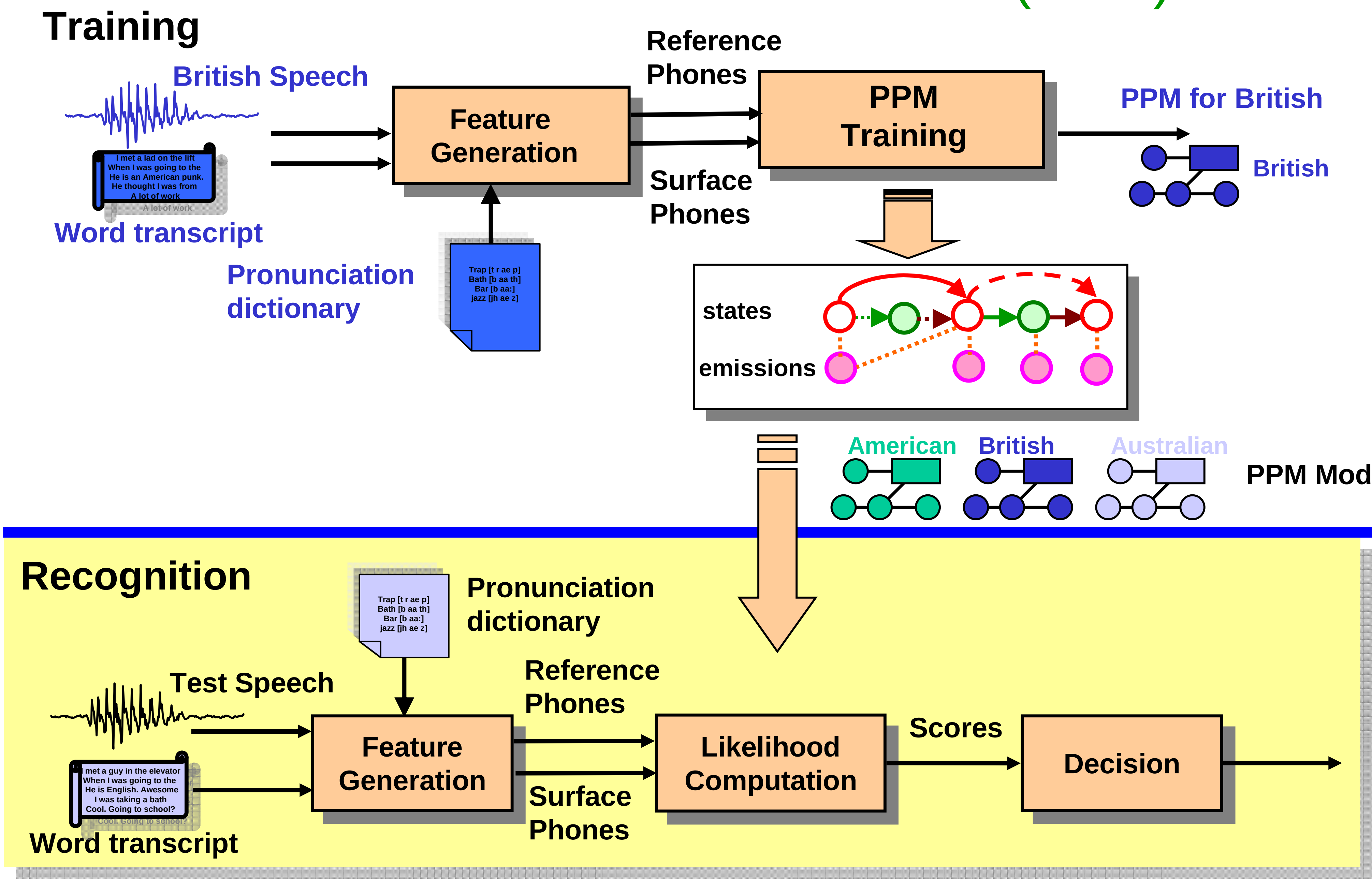
- Assumption
  - If model learned rules well, it can do dialect recognition
- 5 dialects regions
  - UAE (AE), Egypt (EG), Iraq (IQ), Palestine (PS), Syria (SY)
  - Conversational telephone speech
- IQ: reference dialect

Data set	Speaker number	Duration
Train	276	46.25 hr
Dev	83	13.9 hr
Test	88	14.75

5 Dialect Regions

Dialect Recognition System

Phonetic Pronunciation Model (PPM)

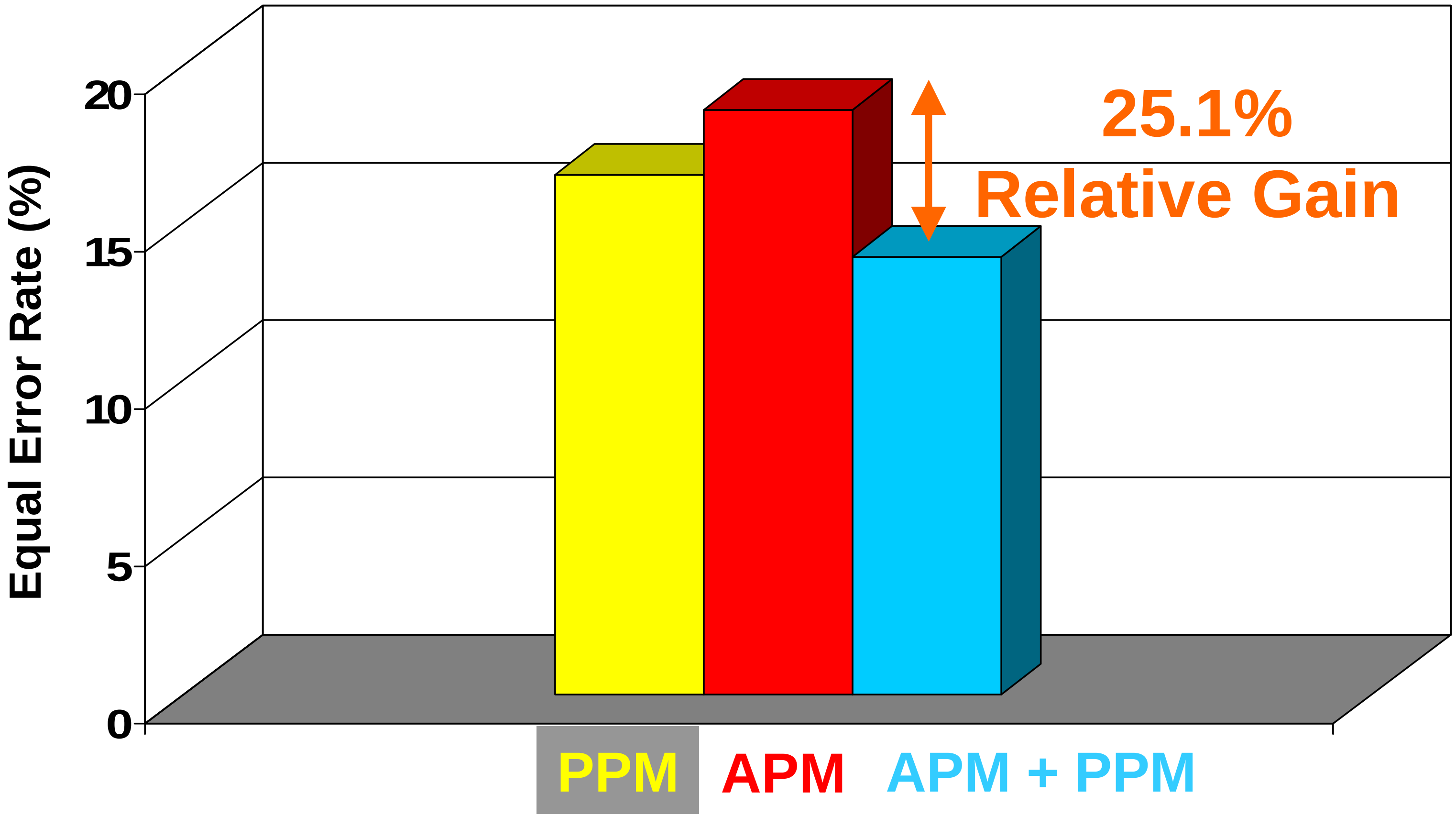


PPM Fuses Well with APM

- APM: Acoustic Phonetic Model (Shen et al, 08)
  - Acoustic, monophone counterpart of PPM
  - Each phone is a GMM
  - Phonetic categorizations determined by forced-alignment

PPM Fuses Well with APM

- APM: Acoustic Phonetic Model (Shen et al, 08)
  - Acoustic, monophone counterpart of PPM
  - Each phone is a GMM
  - Phonetic categorizations determined by forced-alignment



Dialect Recognition Experiment

- Proposed System: Phone-based Pronunciation Model
  - **PPM-1**:
    - Surface phones obtained through **force-alignment** using pan-Arabic pronunciation dictionary
  - **PPM-2**:
    - Surface phones obtained through **direct decoding** using an Iraqi phone recognizer

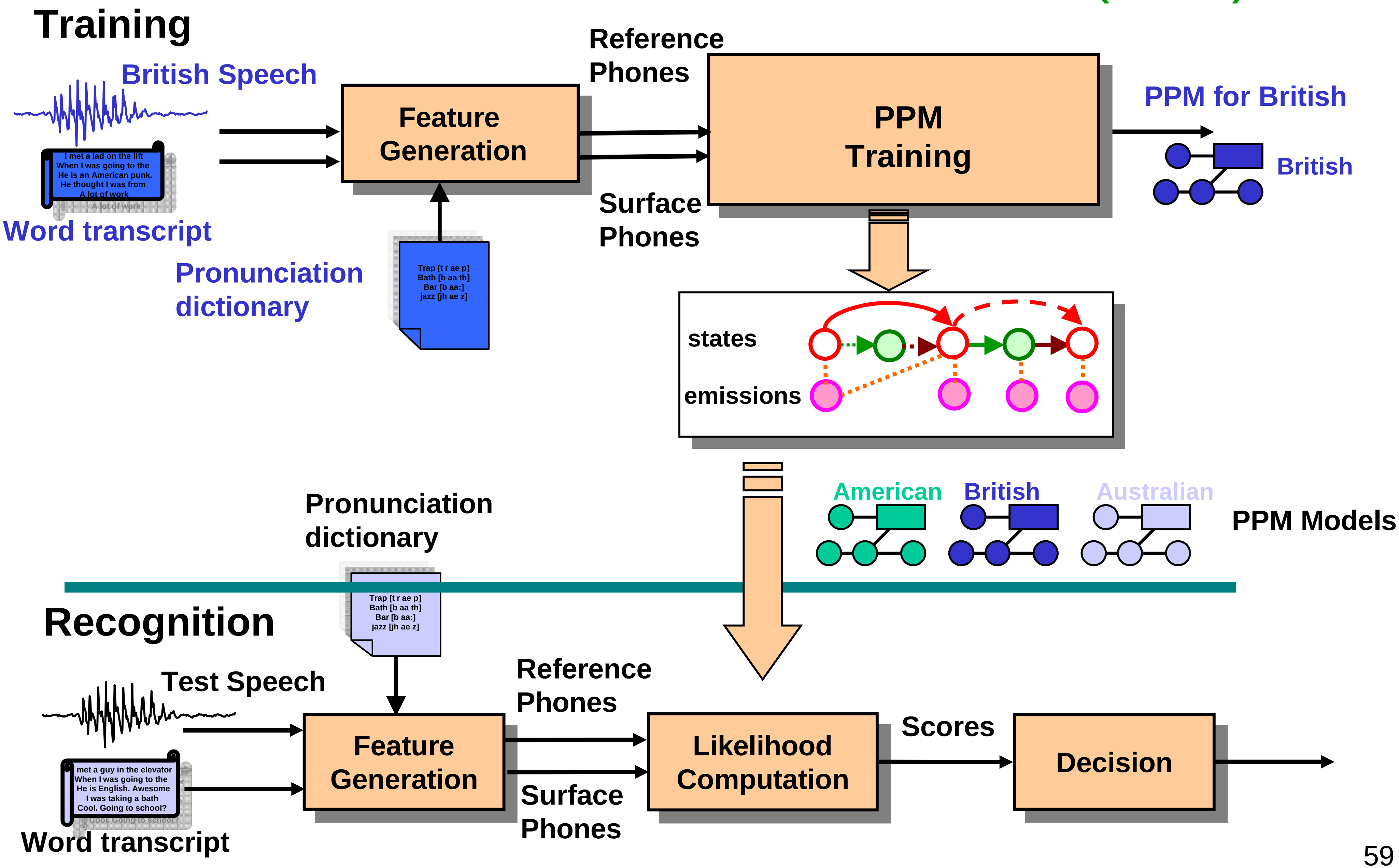
Dialect Recognition Experiment

- Proposed System
  - Phone-based Pronunciation Model (PPM)
    - Reference phones:
      - **forced-alignment** using word transcripts & Iraqi pronunciation dictionary
    - Surface phones:
      - **direct decoding** using Iraqi phone recognizer

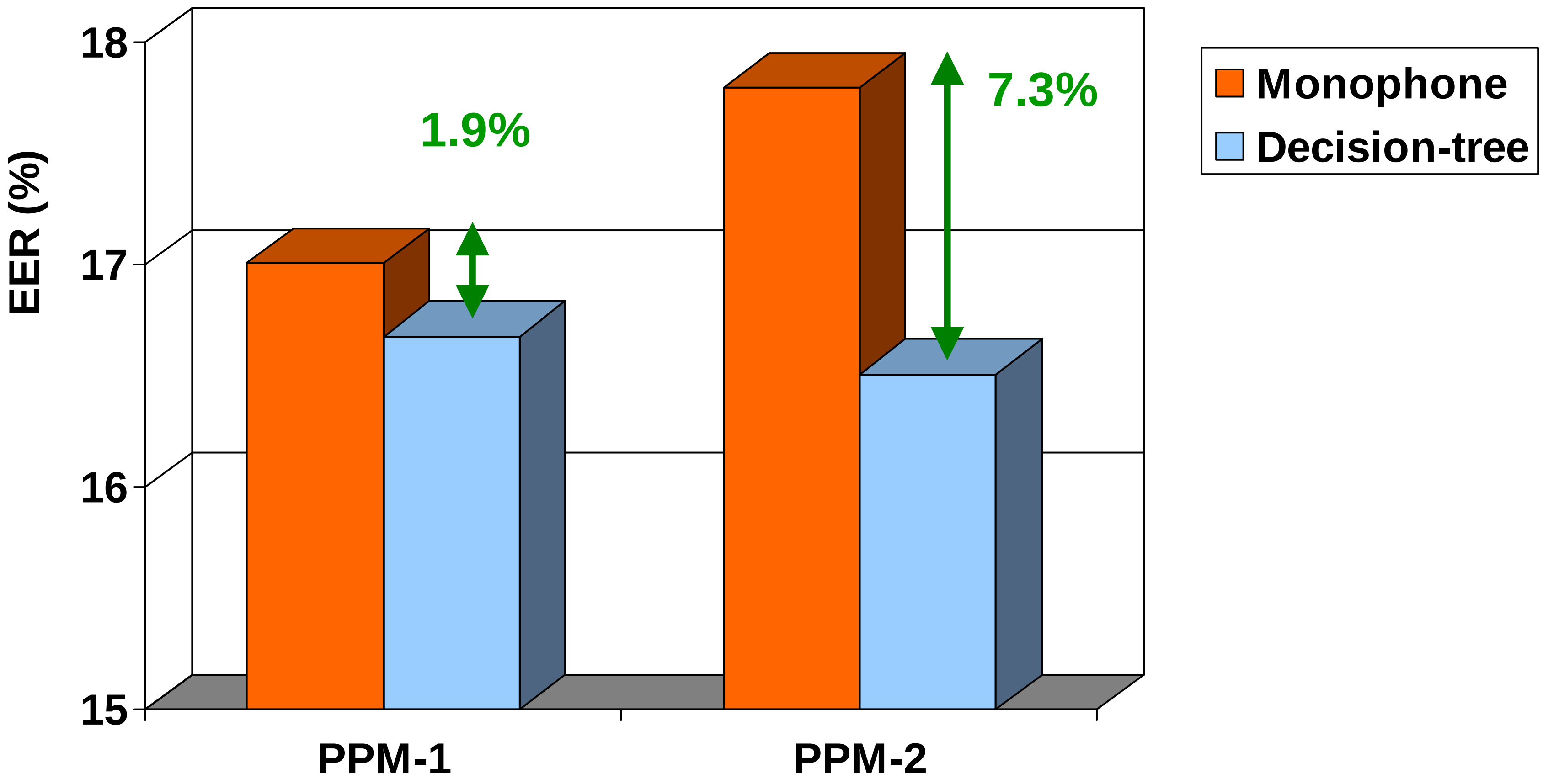
- Baseline System
  - Acoustic Phonetic Model (APM)
    - Each phone is a GMM
    - Phonetic categorizations determined by forced-alignment

Proposed Approach

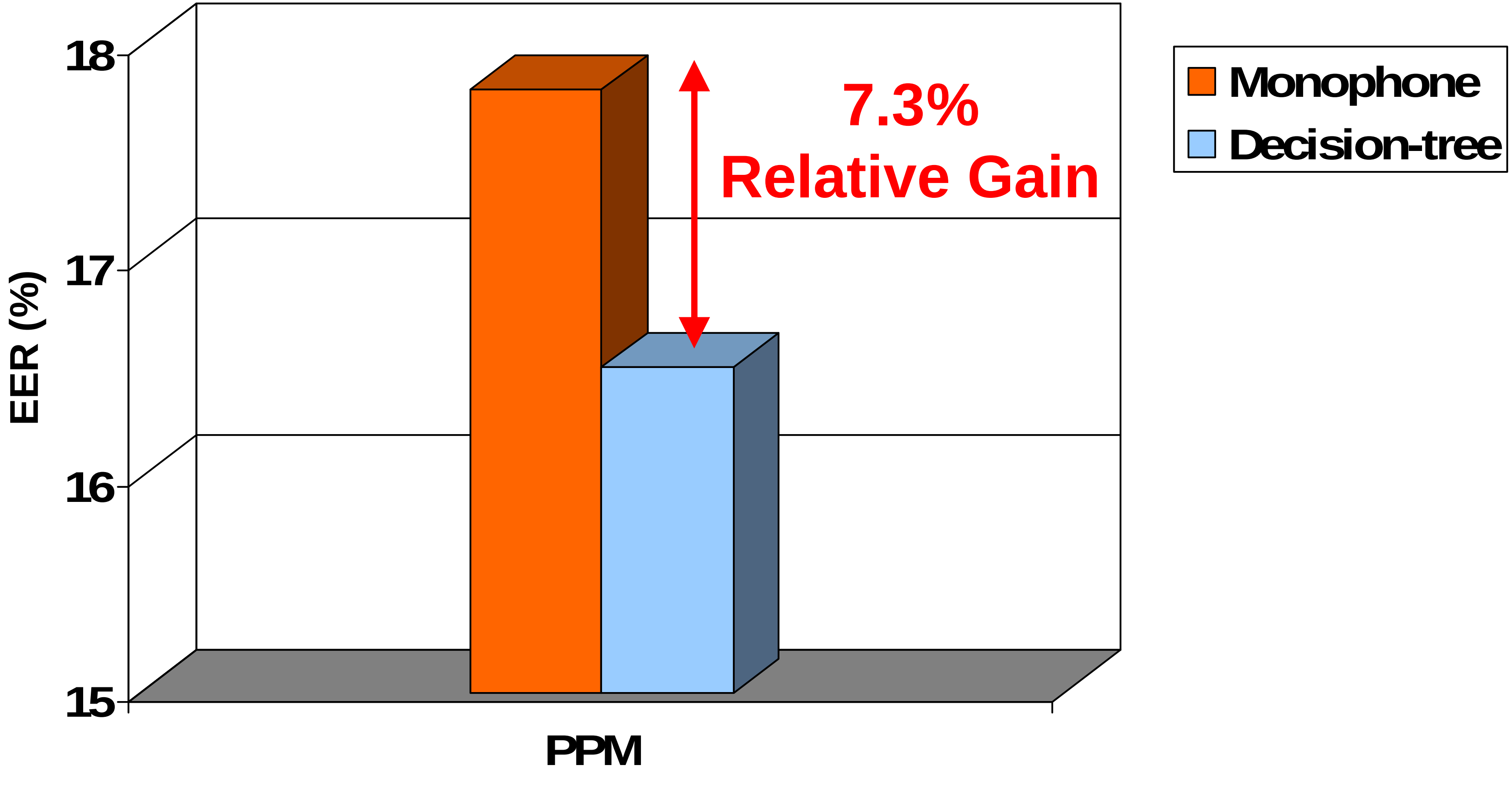
Phonetic-based Pronunciation Model (PPM)



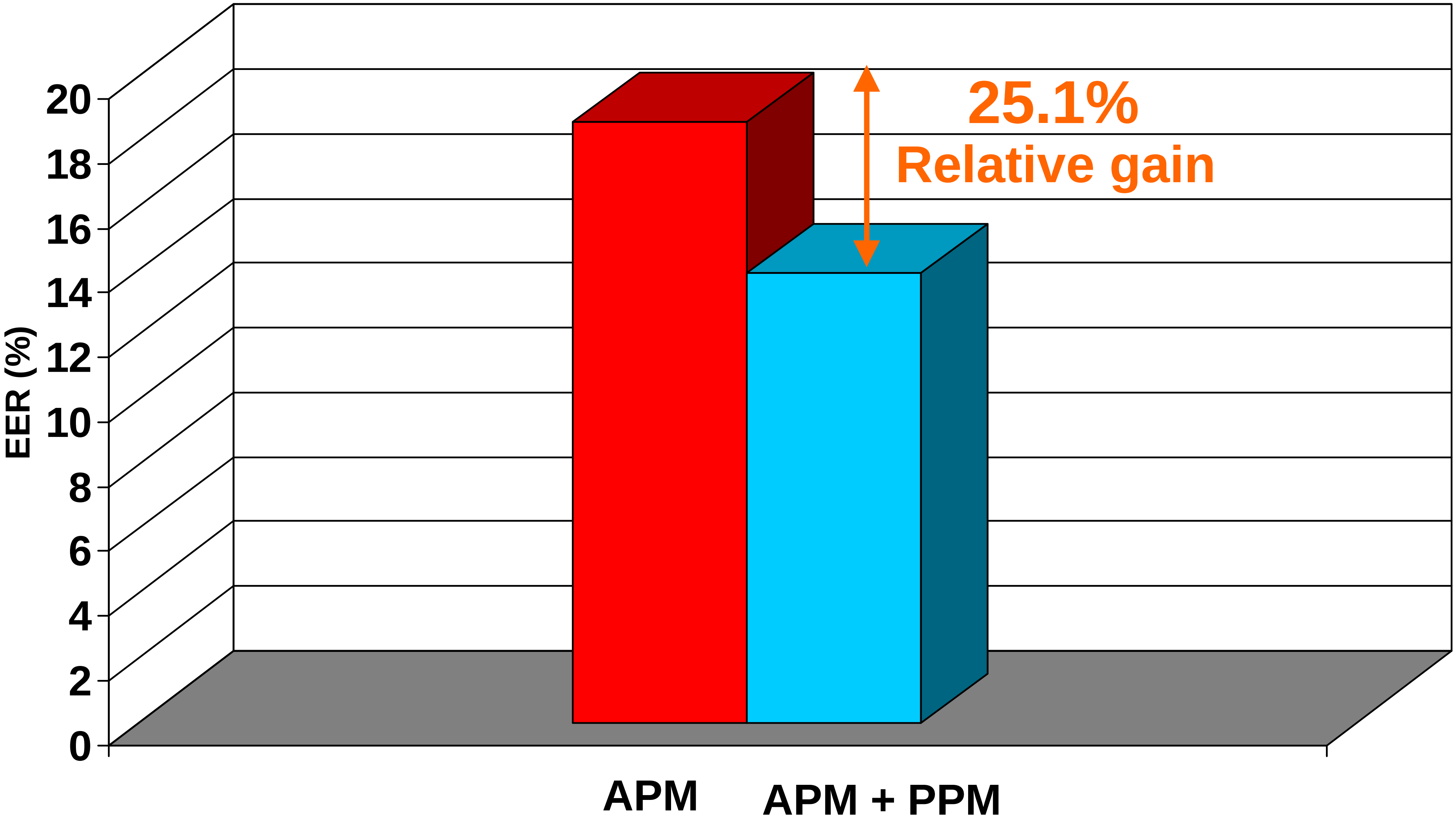
- Phonetic context improves dialect recognition performance
- Performance: Decision Tree PPM-2 > Decision Tree PPM-1



Phonetic Context Helps Characterize Dialects

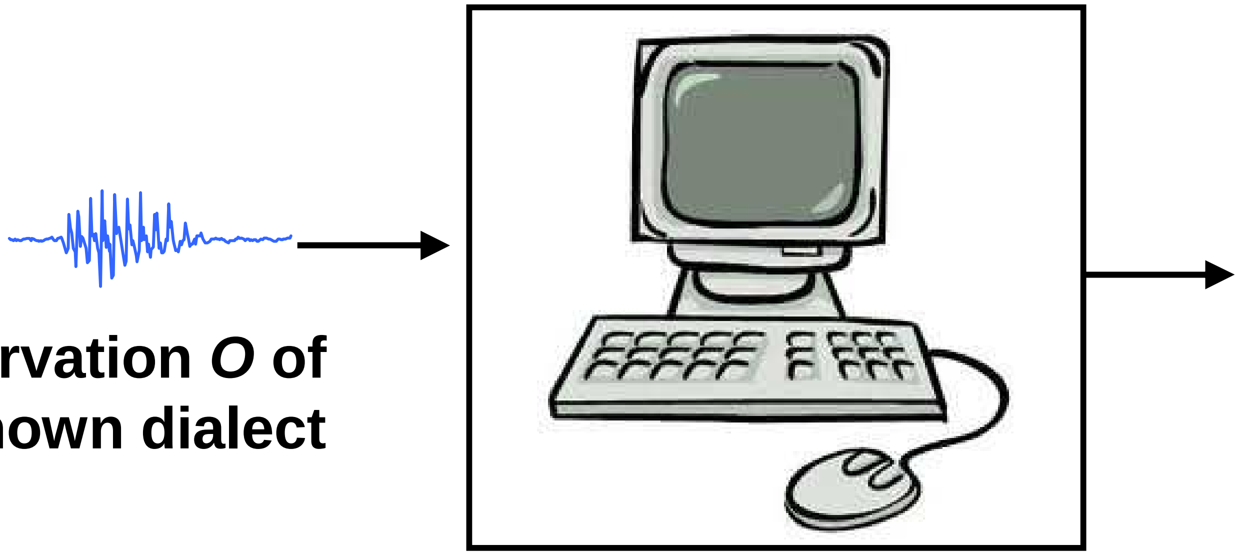


PPM fuses well with APM



Automatically Identifying Dialects

- **Dialect recognition is an identification task**
- **Likelihood ratio is used to make decisions**

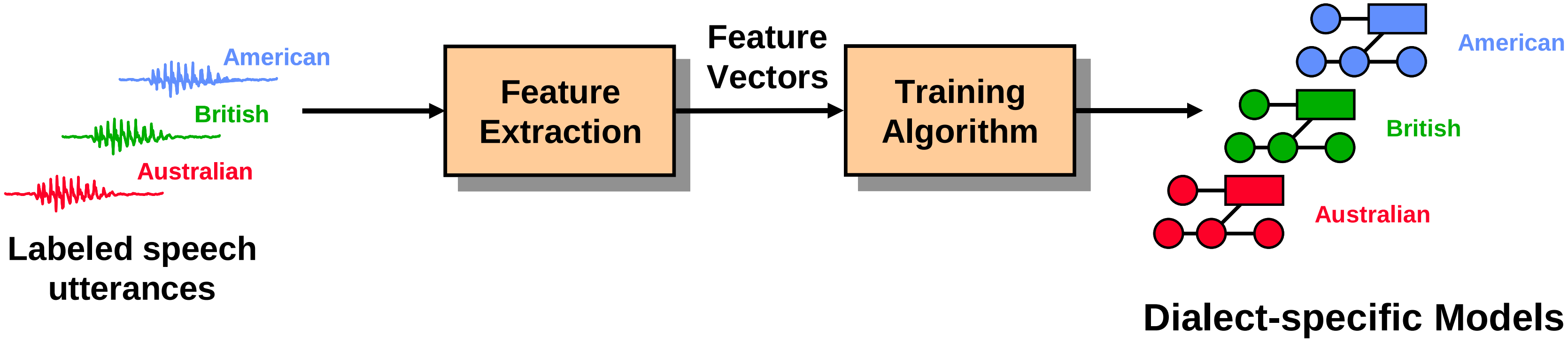


$$\log \frac{P(O|\lambda_d)}{\sum_{i \neq d} P(O|\lambda_i)}$$

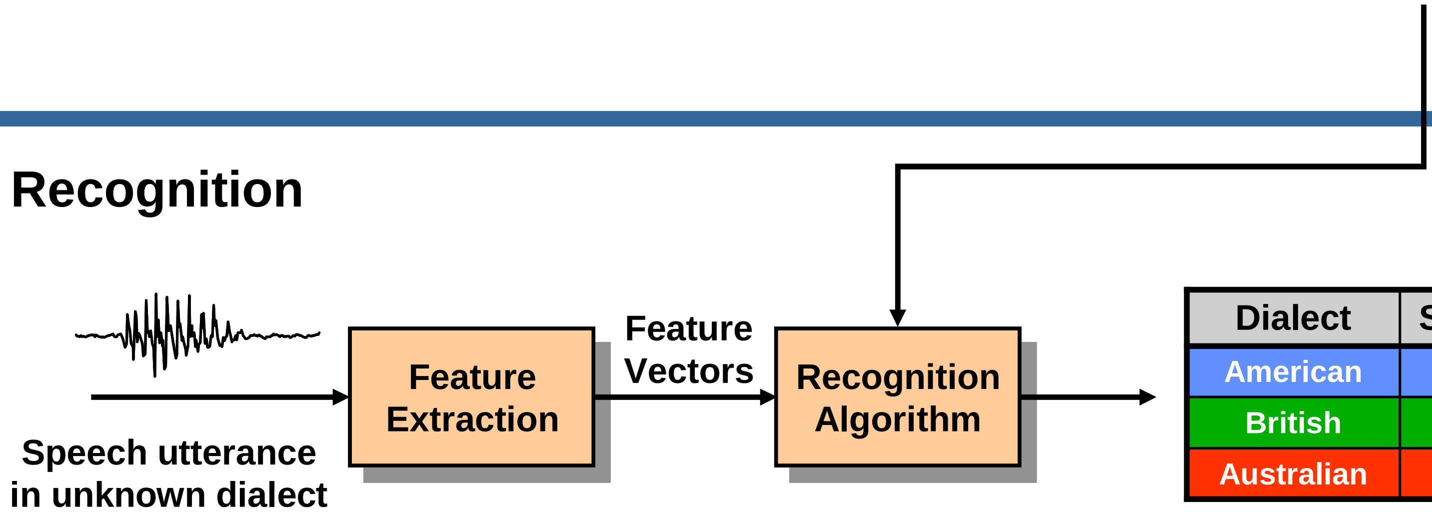
Automatic Dialect Identification

System Architecture

- **Training**



- **Recognition**



Dialect	Score
American	0.66
British	0.93
Australian	0.23

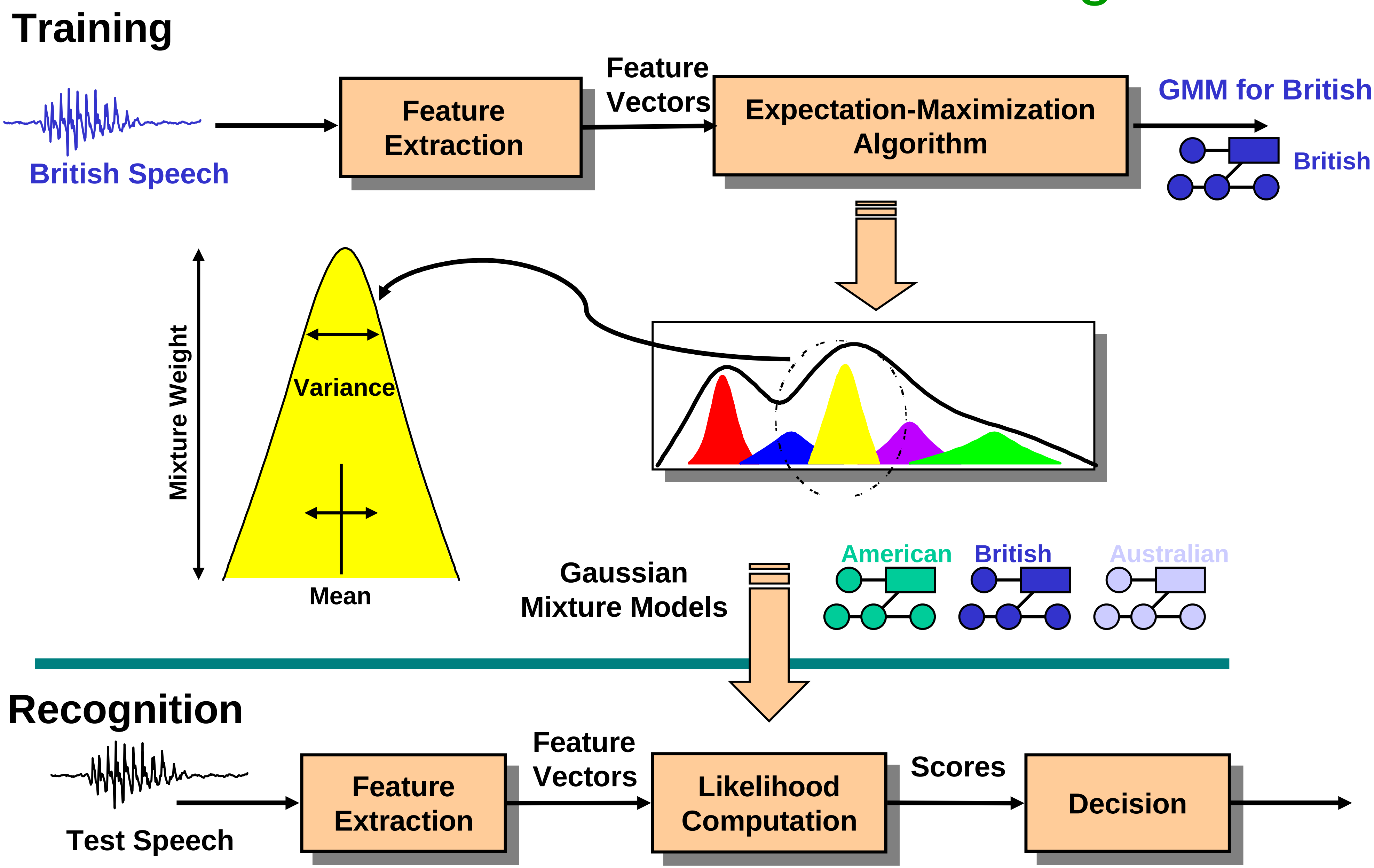
$$\log \frac{P(O|\lambda_d)}{\sum_{i \neq d} P(O|\lambda_i)}$$

Dialect Recognition Experiment (4 dialects)

- IQ: reference dialect
- Baseline Systems
  - **APM-1**: adapted phonetic model (Shen et al, 08)
    - Acoustic segmentation determined by phone recognition
  - **APM-2**:
    - acoustic segmentation determined through force-alignment with word transcripts
  - **SDC-GMM**: shifted-delta-cepstra Gaussian mixture model (Torres-Carrasquillo et al, 2004)
  - **PRLM**: phone recognition followed by language modeling (Zissman et al, 1996)

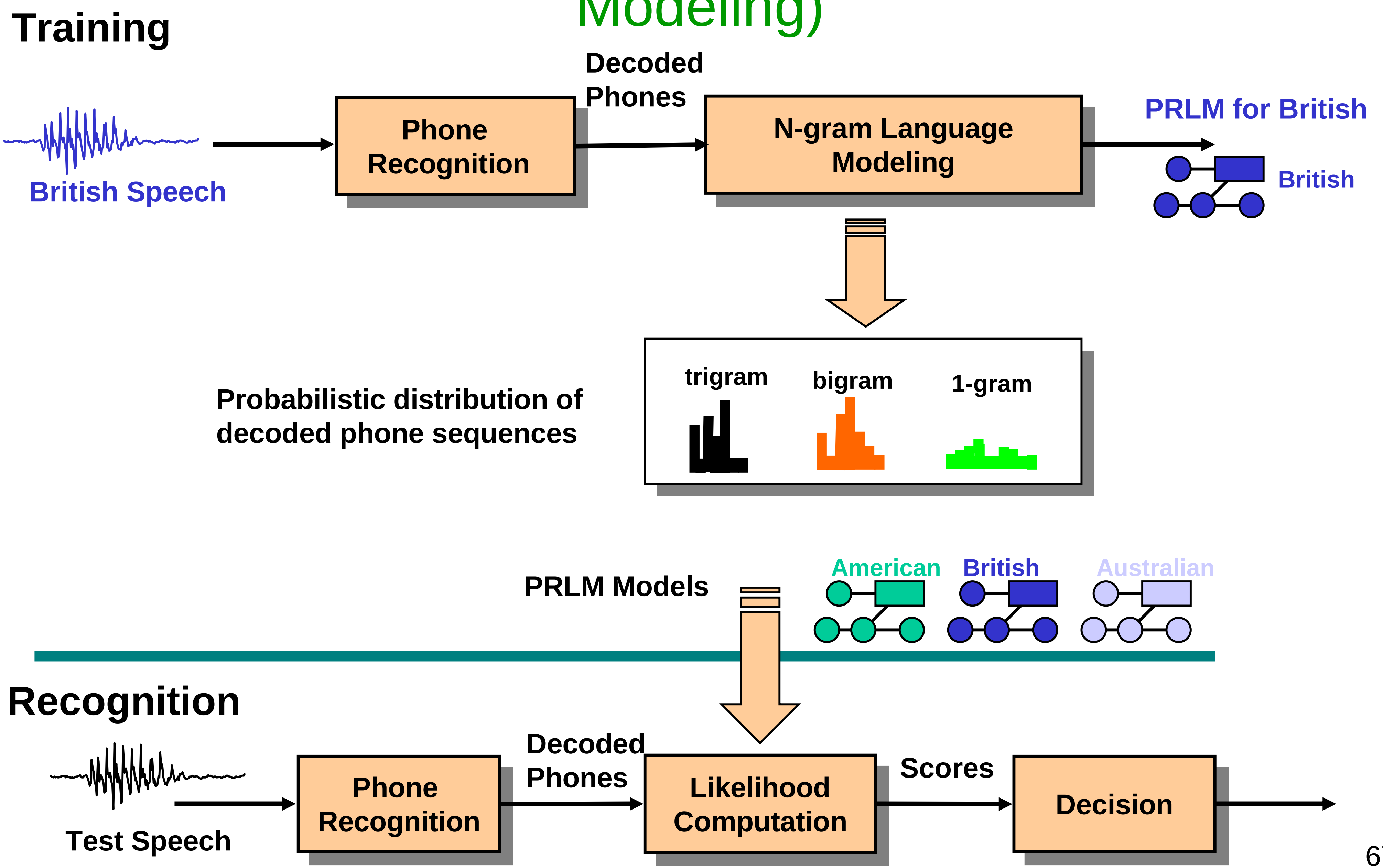
Existing Approach

Gaussian Mixture Modeling



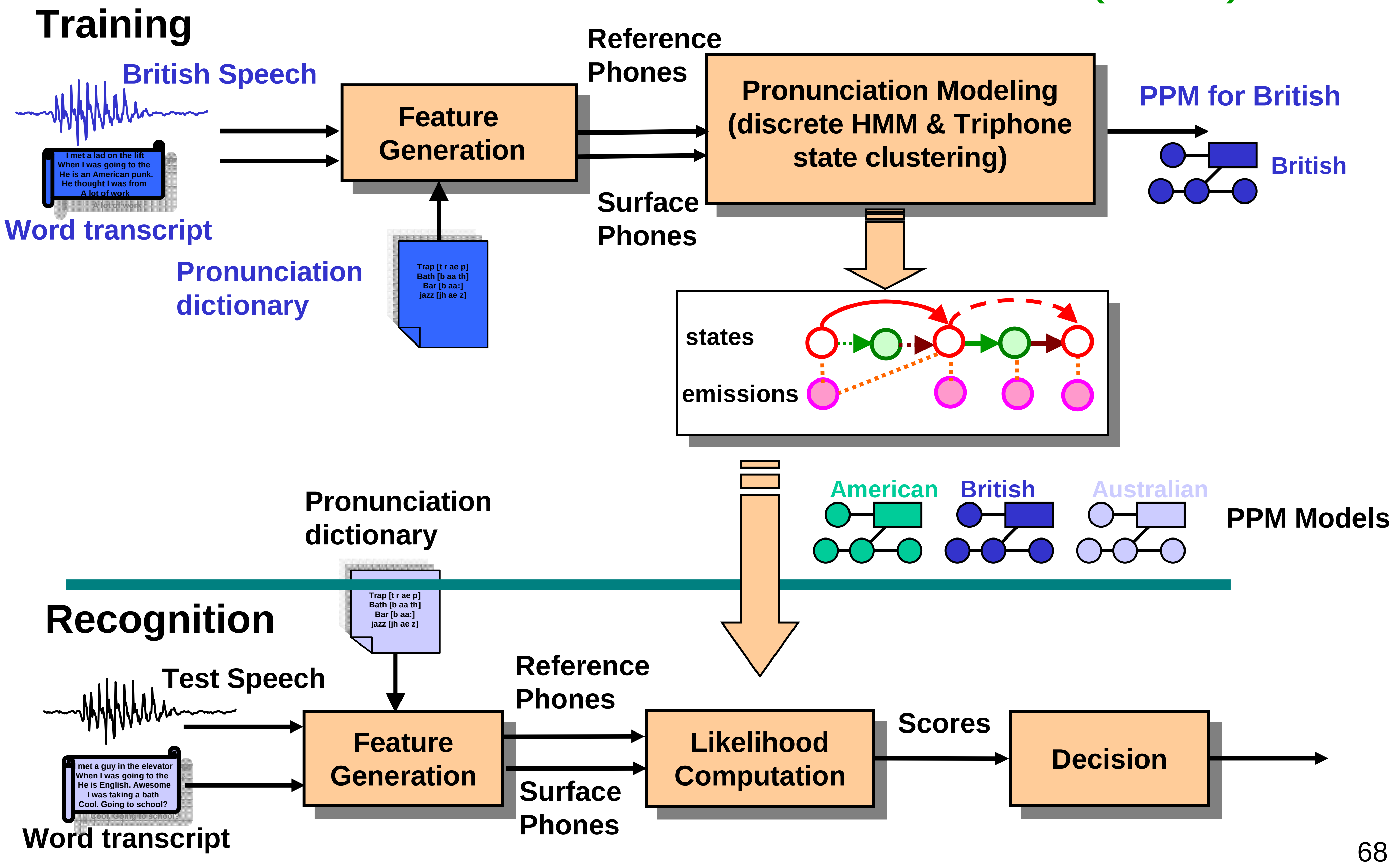
Existing Approach

PRLM (Phone Recognition followed by Language Modeling)



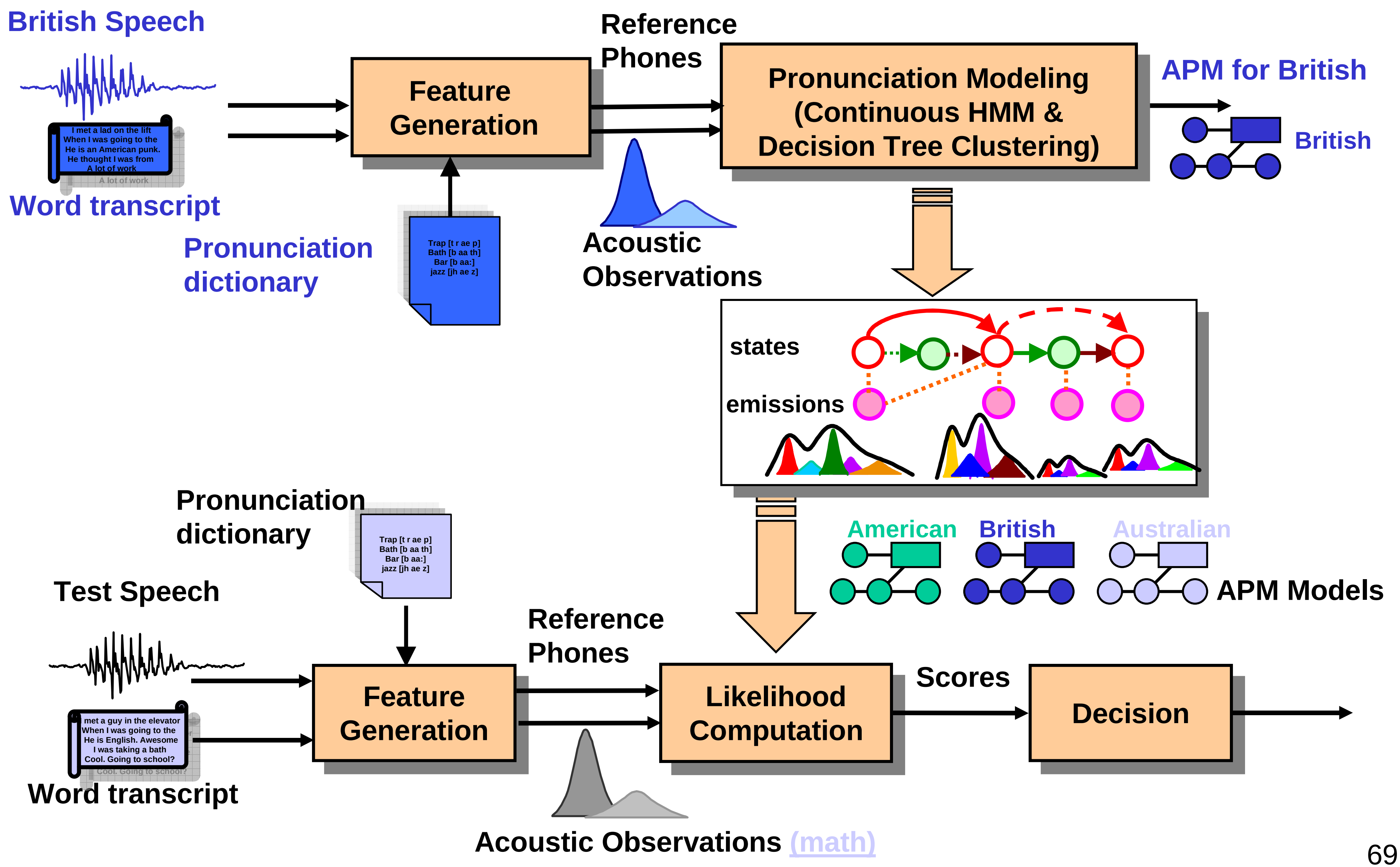
Proposed Approach

Phonetic-based Pronunciation Model (PPM)



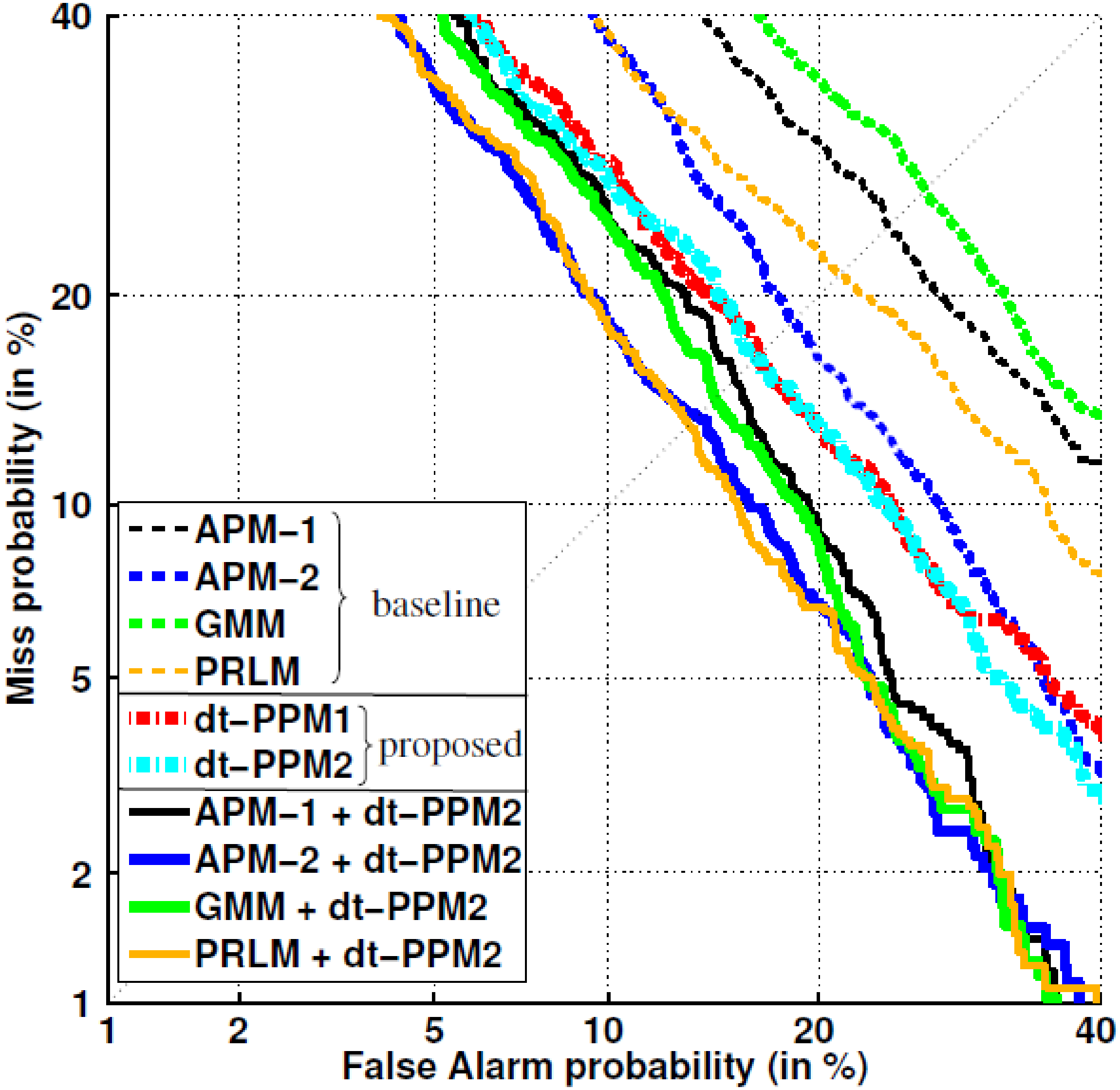
Proposed Approach

Acoustic-base Pronunciation Model (APM)



Detection Error Trade-off

Error rate: fused systems < Proposed PPM < baselines



Dialect Recognition Summary

- PPM: Decistion Tree outperforms Monophone
- DT PPM-2 outperforms DT PPM-1
  - PPM-2: Learned rules not limited to pronunciation dictionary
- DT PPMs fuse well with baseline systems

Pronunciation Generation Experiment

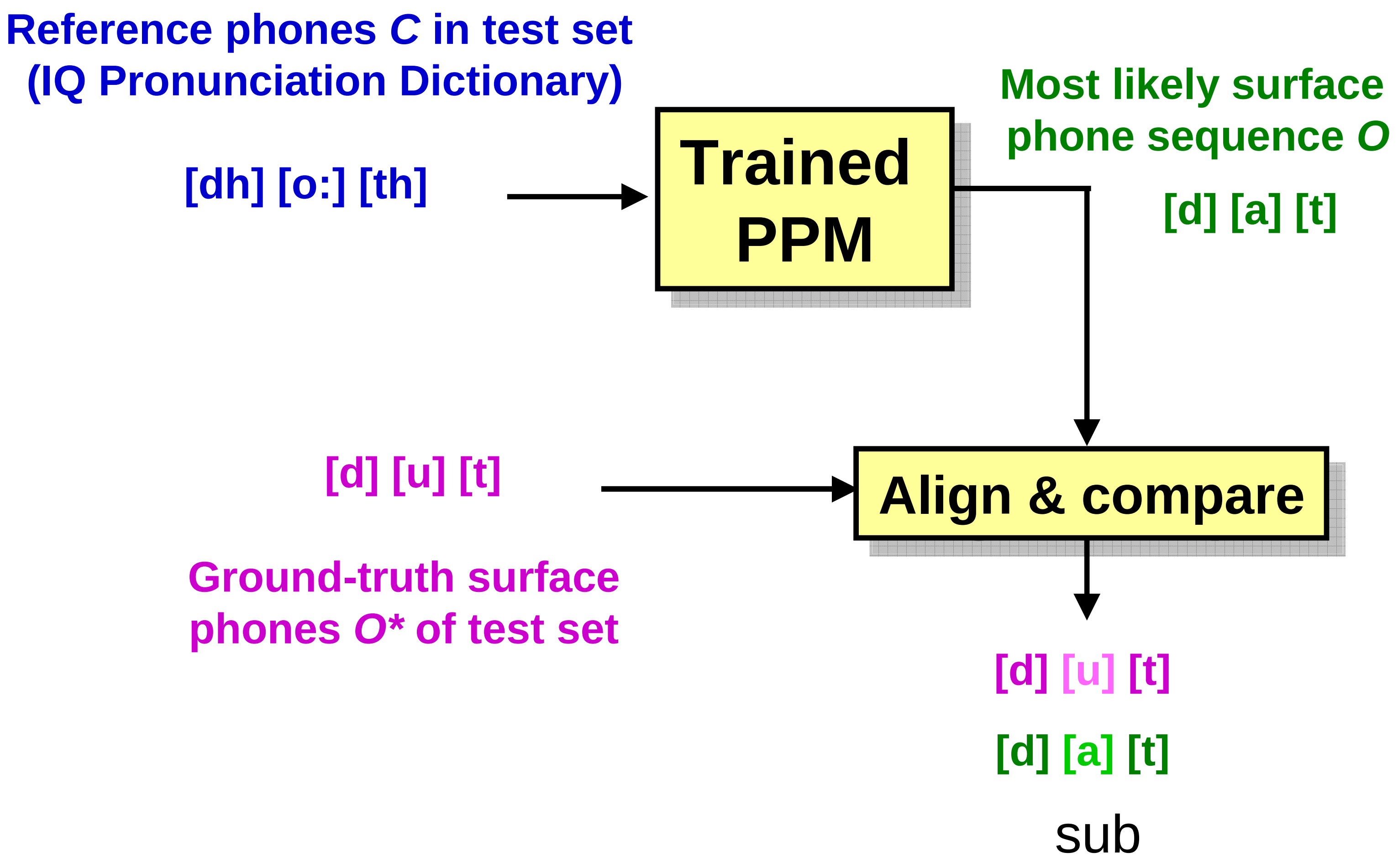
Assumptions

1. All pronunciation variations across dialects are governed by underlying phonetic rules

2. The phonetic transcriptions provided by WSJ-CAM0 are ground-truth surface phones $O*$

3. Ability to predict ground-truth surface phones $O*$ from the trained pronunciation model given the reference phones indicates how well the phonetic rules are learned from the pronunciation model algorithms

Experimental Setup

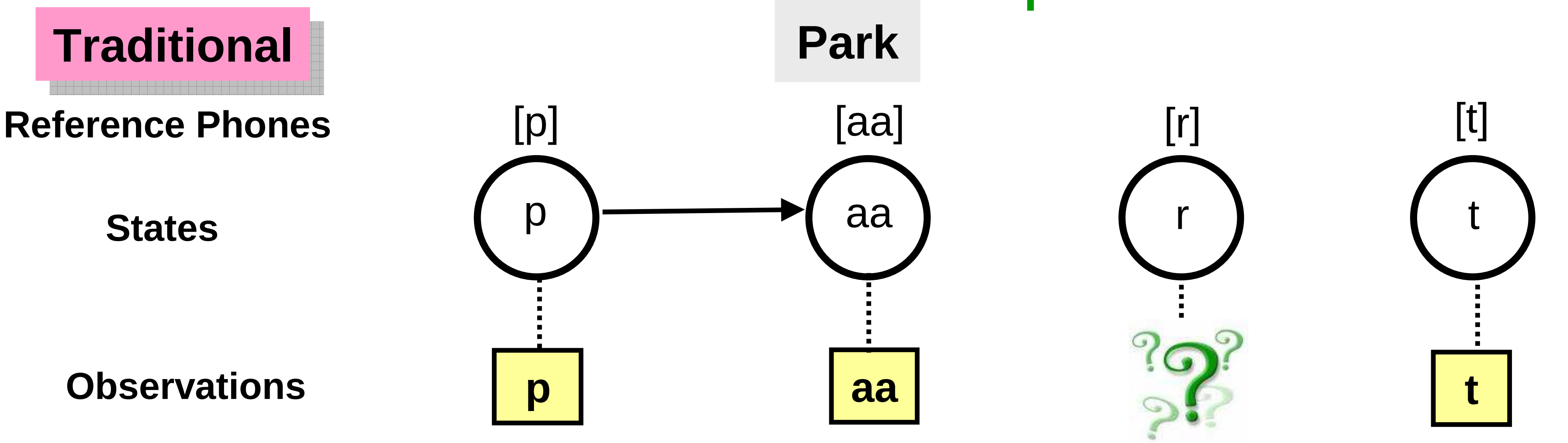


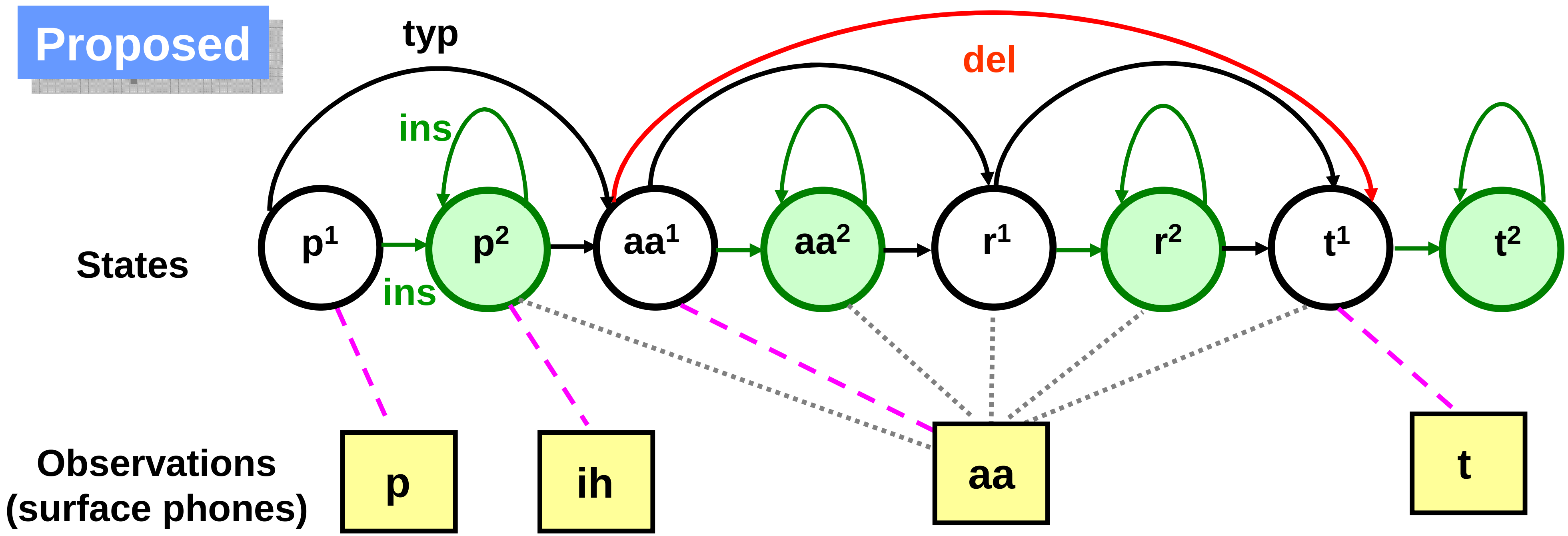
Phone error rate (PER): 33%

HMM

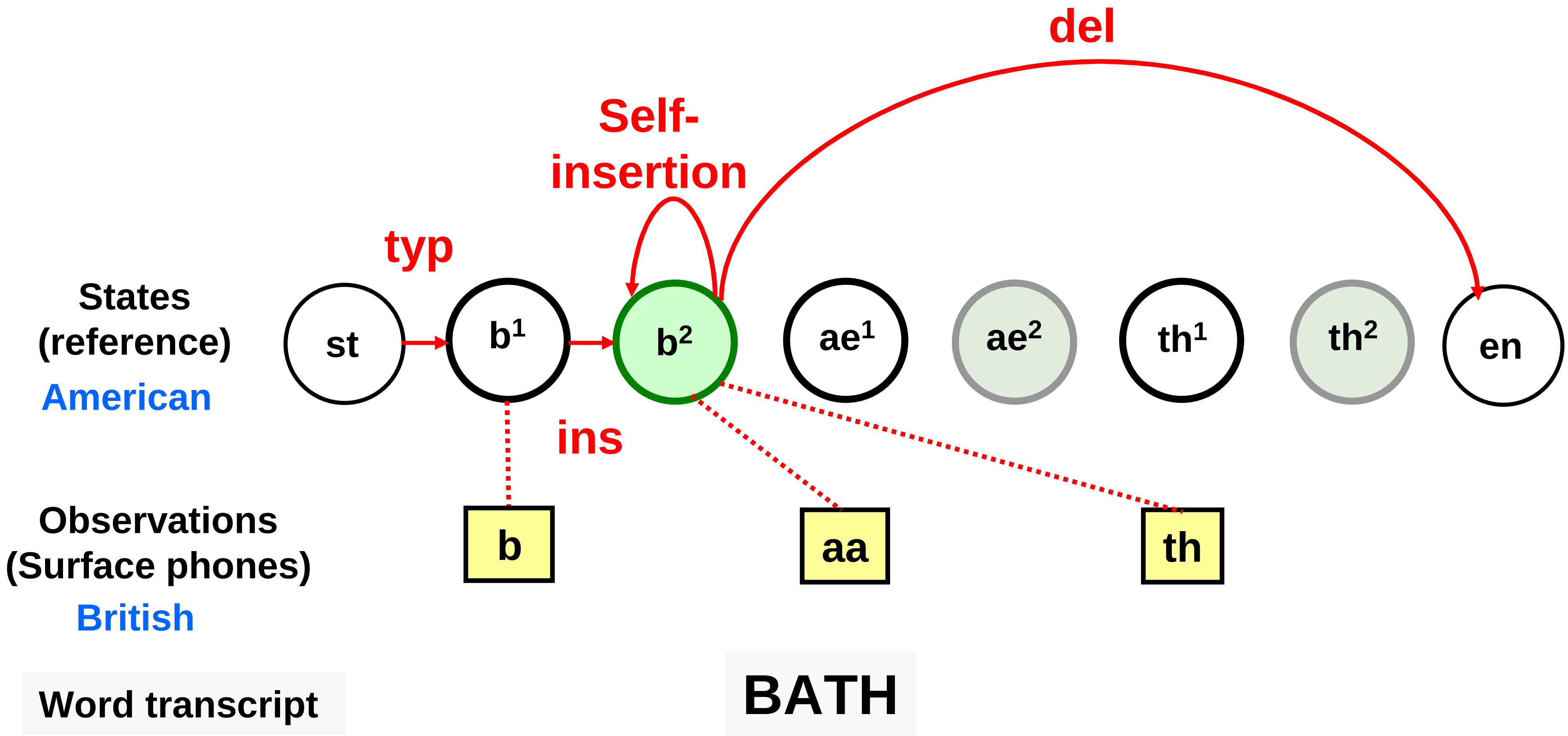
Hidden Markov Model (HMM)

Traditional vs. Proposed

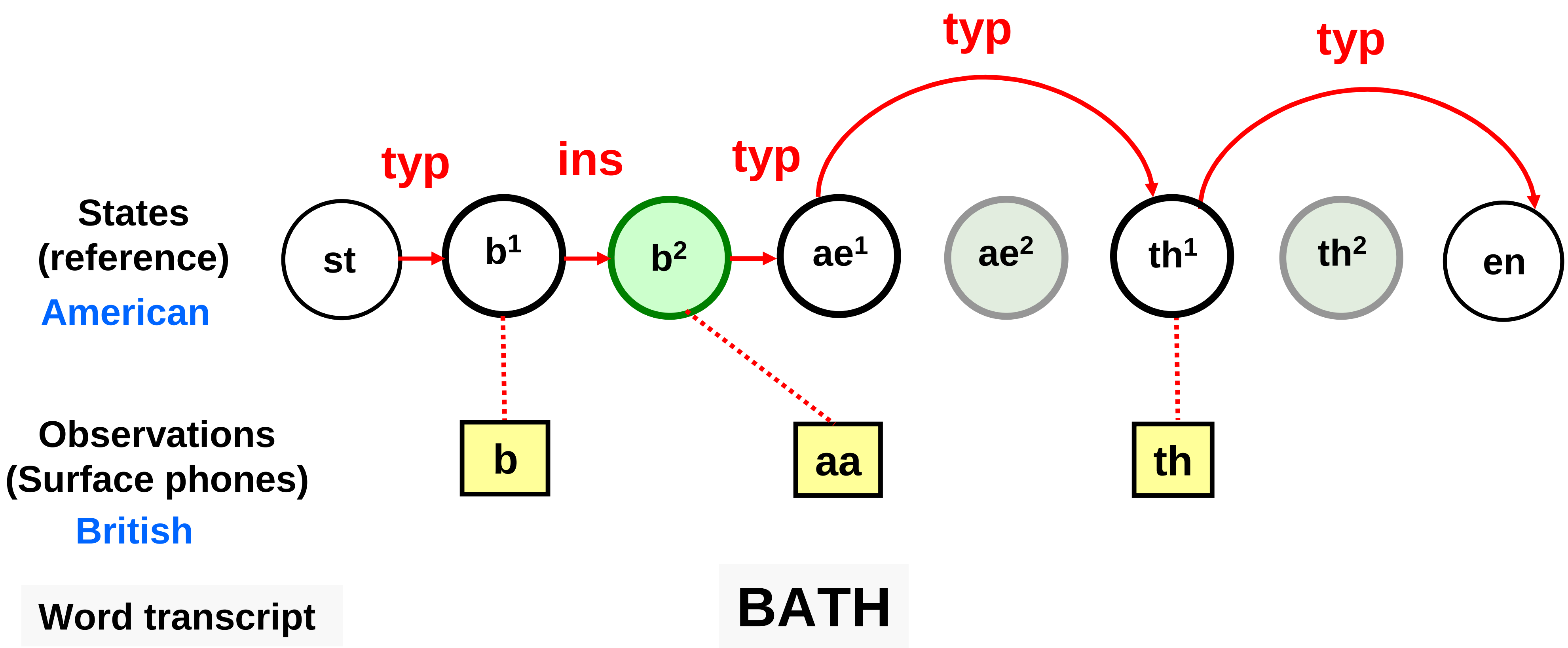




Alignment Example 1

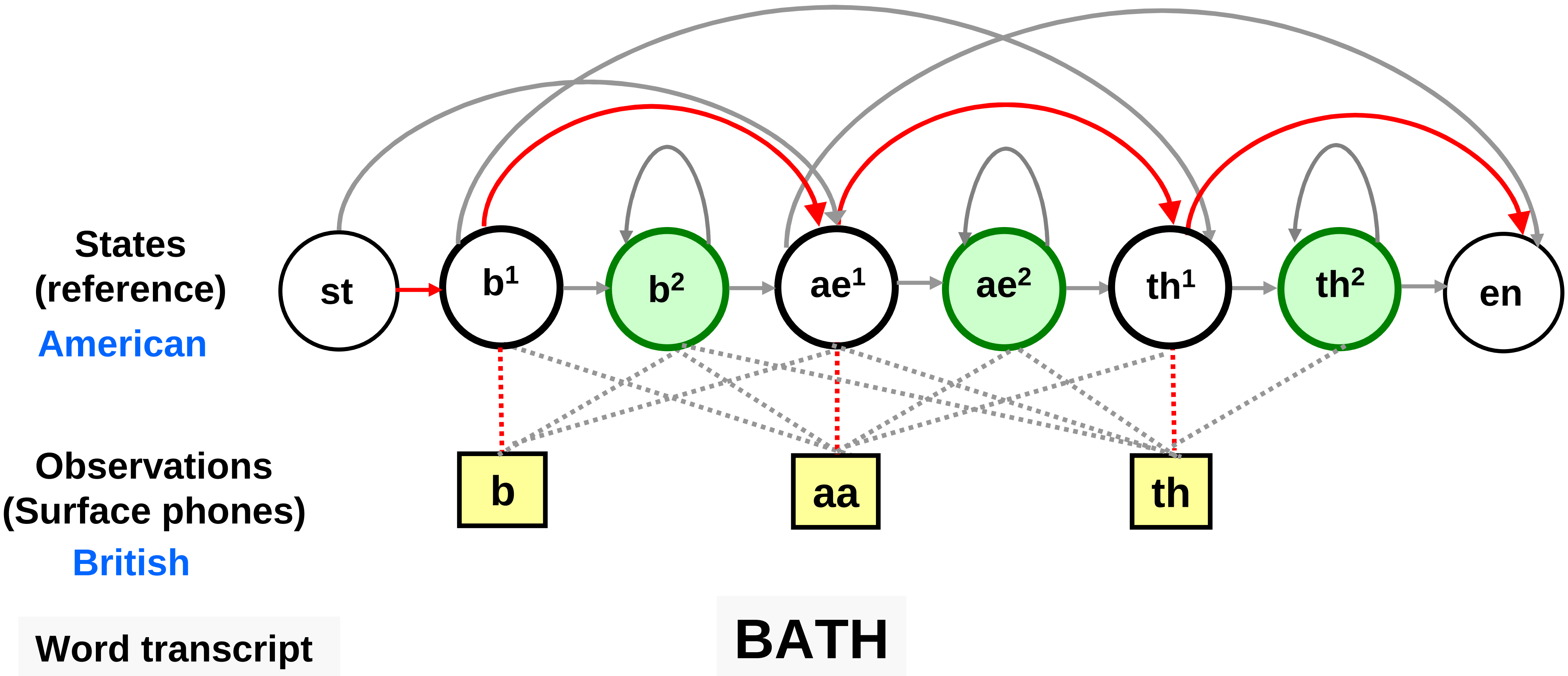


Alignment Example 2



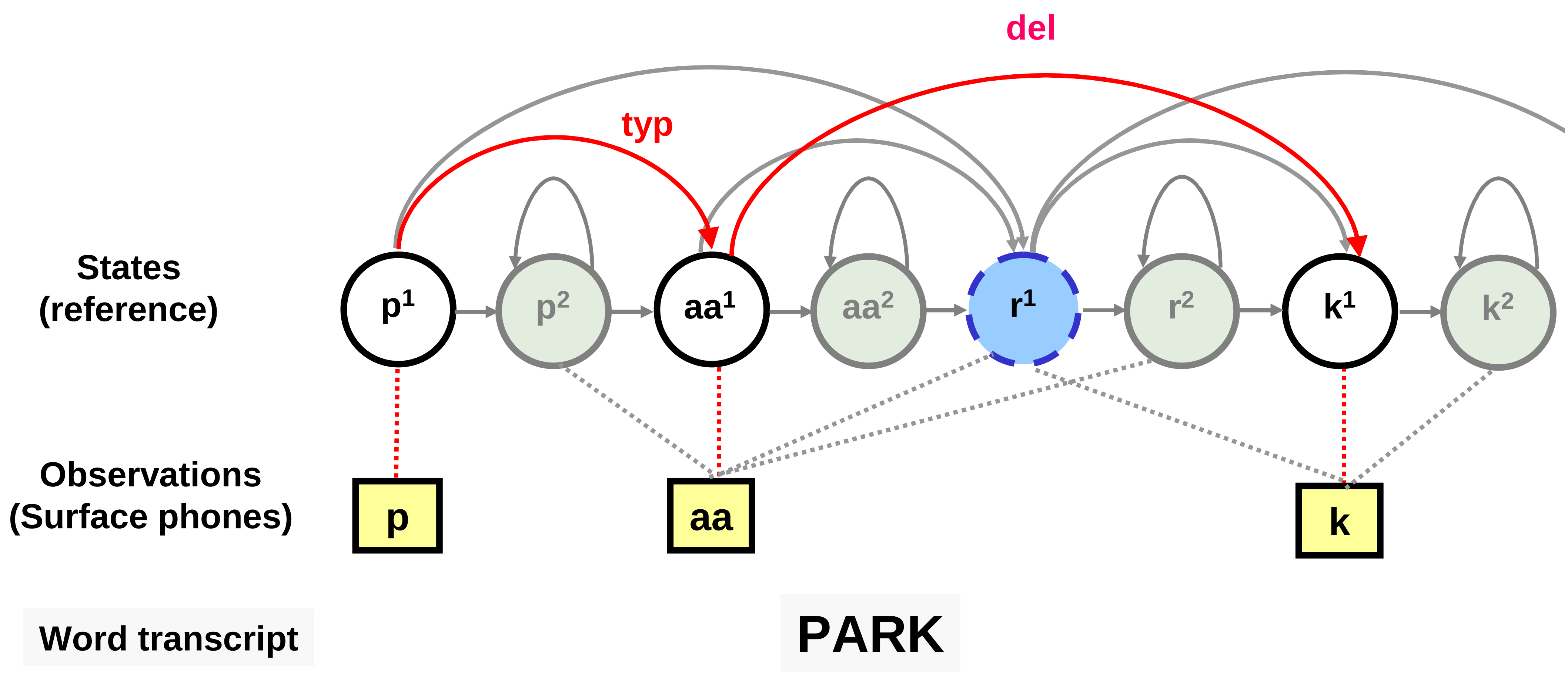
Alignment Example 3
/ae/ substitution rule

All possible alignments, given the states and observations



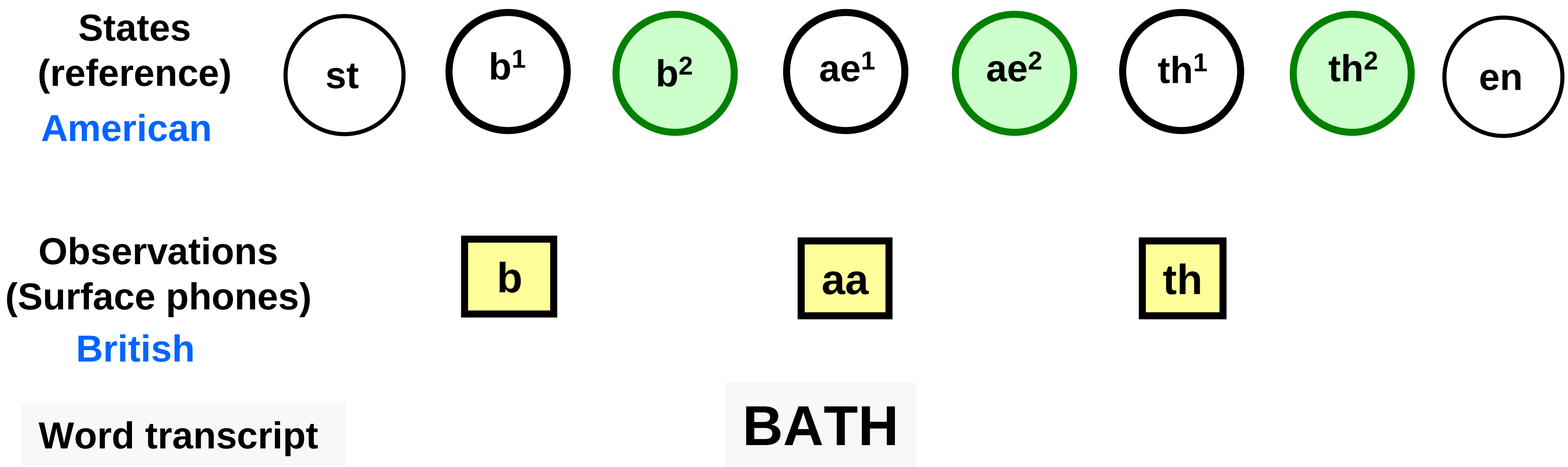
Non-Rhoticity Rule

The most likely alignment for the word *park*



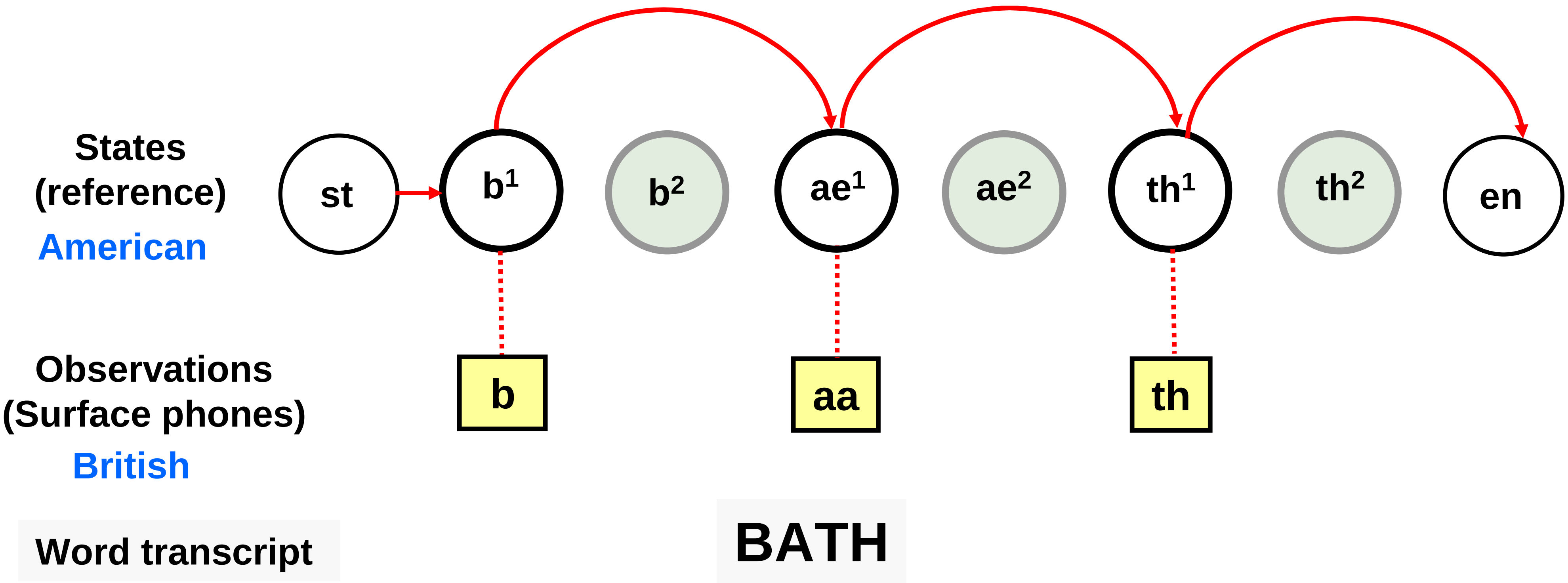
Alignment Example 1

Given the reference phones and surface phones, what are the possible alignments?



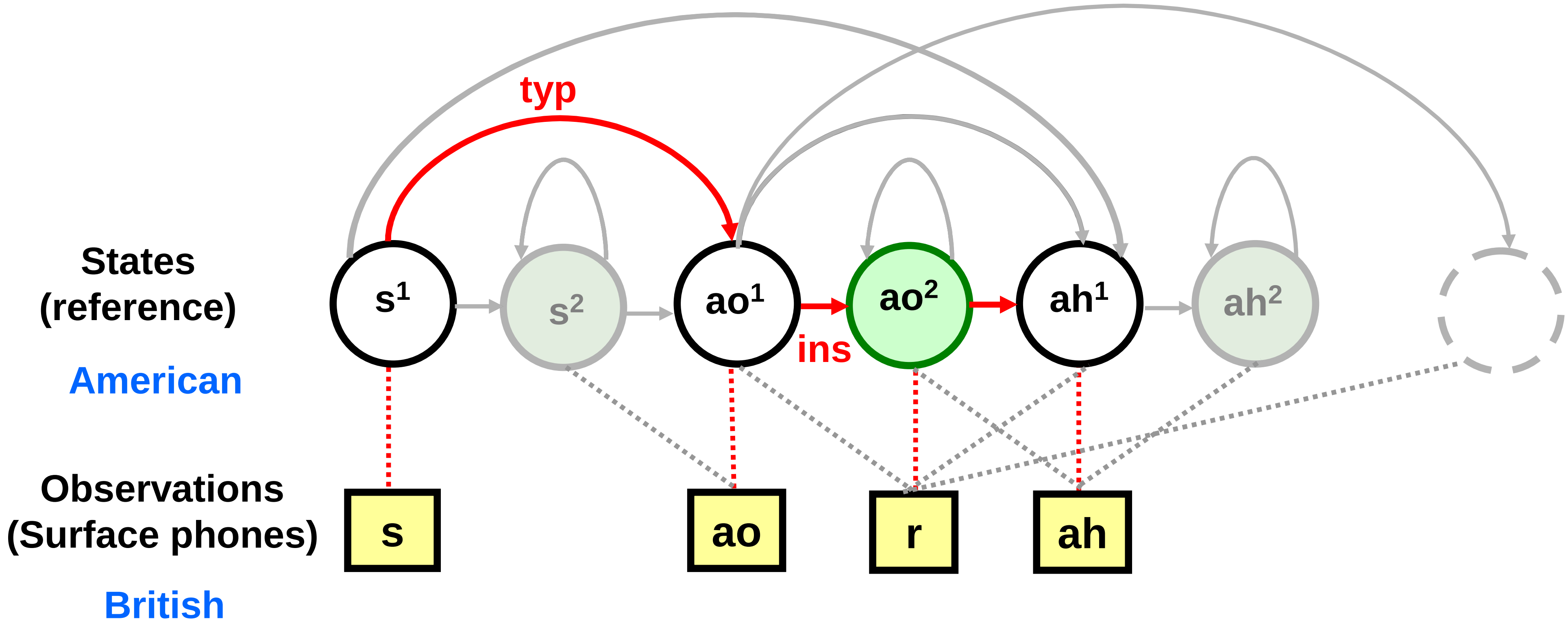
Alignment Example 1
/ae/ substitution rule

We expect the most likely alignment to be something like this.



Intrusive r example

The most likely alignment for the phrase *saw a film*



Word transcript

SAW A (FILM)

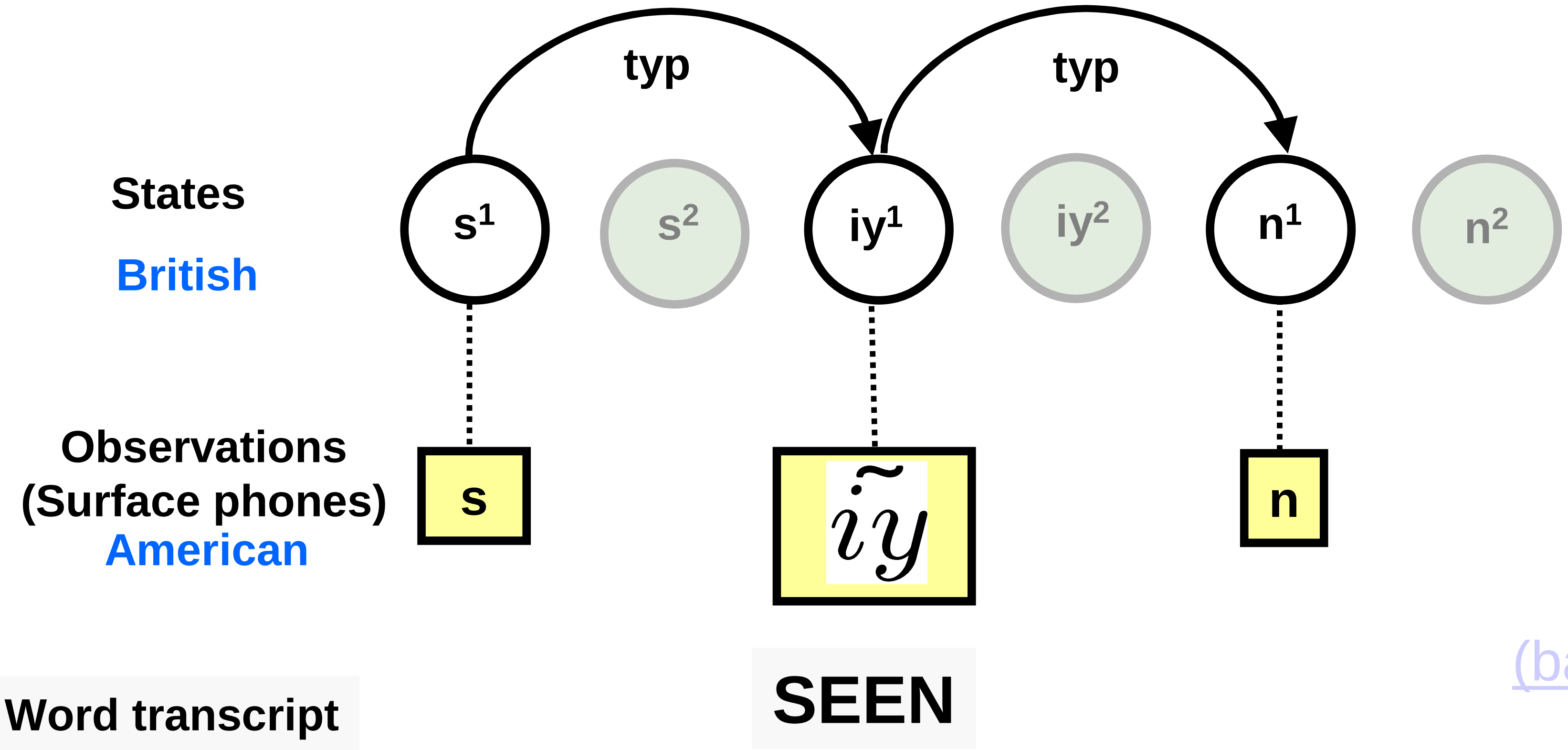
Limitations in Learning Sub & Ins Rules

Cannot fully model right-context driven rules

Rule: [+vowel] -> $[\text{+vowel}]_{\text{nasalized}}$ / _ [+nasal]

[iy] -> $[\text{iy}]_{\text{nasalized}}$ / _ [+nasal]
[ah] -> $[\text{ah}]_{\text{nasalized}}$ / _ [+nasal]
[ay] -> $[\text{ay}]_{\text{nasalized}}$ / _ [+nasal]
…..

Learned rules are less general



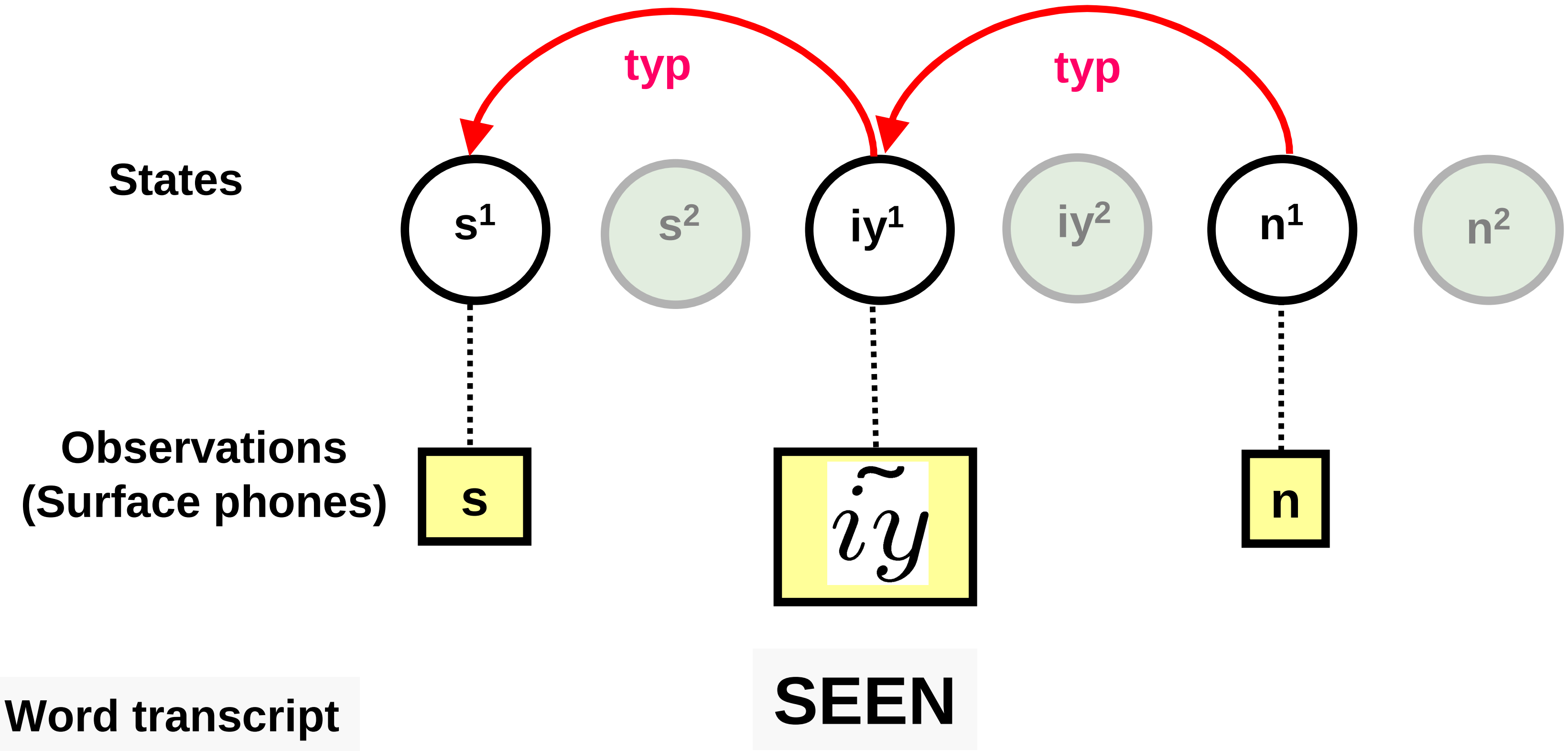
(backup)

Modeling Rules Driven by Right-Context

Rule: [+vowel] -> [+vowel]$_{\text{nasalized}}$ / _ [+nasal]

Learned Rule: [+vowel] -> [+vowel]$_{\text{nasalized}}$ / _ [+nasal]

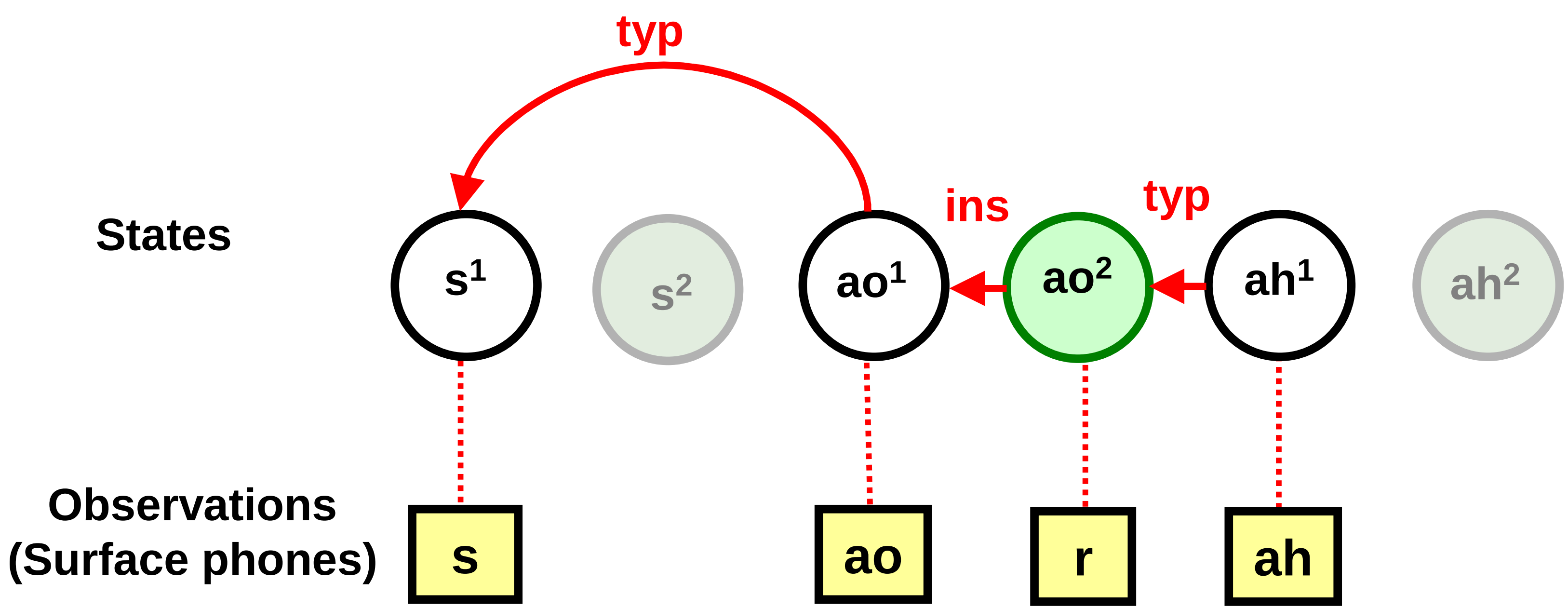
Reverse source and target!



(back)

Modeling Rules Driven by Right-Context

Reverse source and target!



Word transcript **SAW A (FILM)**

(back)

Monophone vs. Triphone

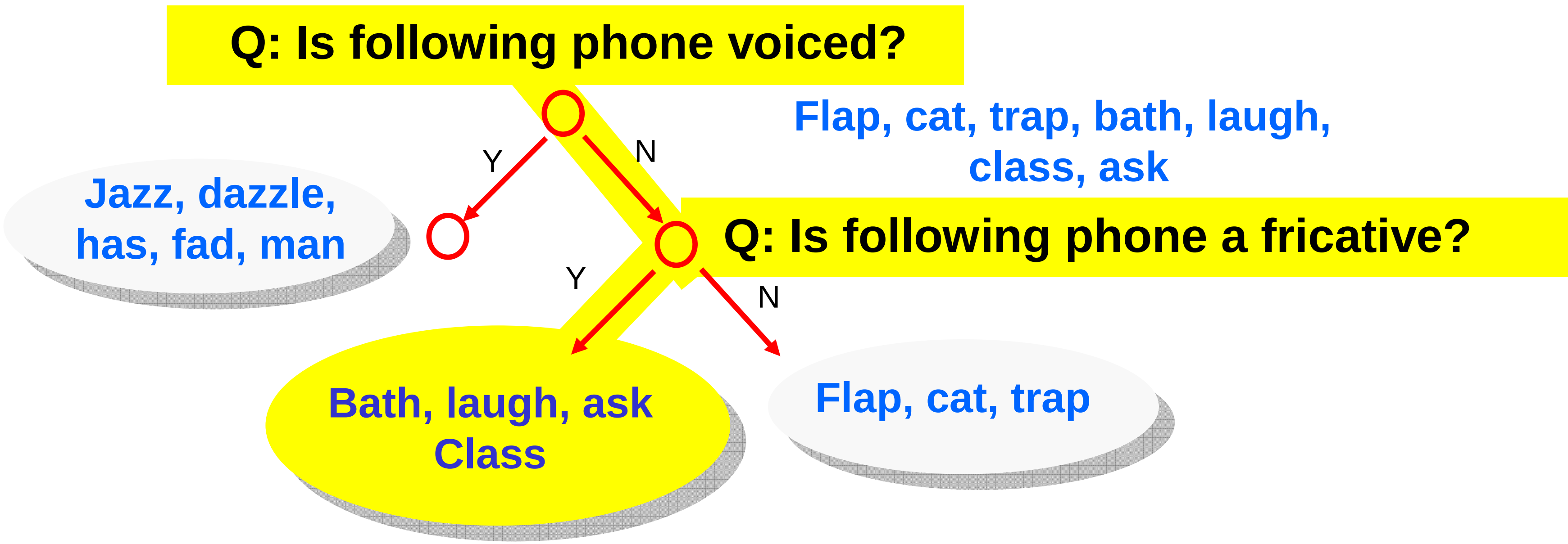
text		**Park**		
Monophones	**[p]**	**[aa]**	**[r]**	**[t]**
Monophone States	p	aa	r	t
Triphone States	p + aa	p - aa + r	aa - r + t	r - t

Decision Tree Clustering

Decision Tree Clustering

Find a set of features that best describe how [ae] is realized in British English

bath, jazz, laugh, dazzle, has, fad, man, cat, class, flap, trap, ask



back

Learned Rule: [ae] -> [aa] / _ [-voiced, +fricative]
-> triphone states (* - ae + [-voiced, +fricative]) are clustered

(math)

Generalizing Rules

What is the underlying rule of [ae] transforming to [aa]??

[ae] -> [aa] / _ [th] bath

[ae] -> [aa] / _ [s] class

[ae] -> [aa] / _ [f] laugh

Generalizing Rules

What is the underlying rule of [ae] transforming to [aa]??

[ae] -> [aa] / _ [th] bath

[ae] -> [aa] / _ [s] class

[ae] -> [aa] / _ [f] laugh

[ae] -> [aa] / _ [-voiced, +fricative]

Decision Tree Clustering
Rule Learning

- Phonetic transformation
  - Bath, class, laugh

- What is needed

1. A list of questions

(linguistic characterization)

- Is following phone voiced?
- Is previous phone voiced?
- Is following phone a stop?
- Is following phone a fricative?
- Is previous phone a nasal?
- …….

2. An objective splitting criteria

3. A threshold to stop splitting

Decision tree clustering

Rule Learning

Find a set of features that best describe how [ae] is realized in British English

(math)

Rule Learning

Find a set of features that best describe how [ae] is realized in British English

bath, jazz, laugh, dazzle, has, fad, Man, cat, class, flap, trap, math, hassle, ask

(math)

Decision tree clustering

Rule Learning

Find a set of features that best describe how [ae] is realized in British English

bath, jazz, laugh, dazzle,
has, fad, Man, cat, class, flap,
trap, math, hassle, ask

Q: Is following phone voiced?

(math)

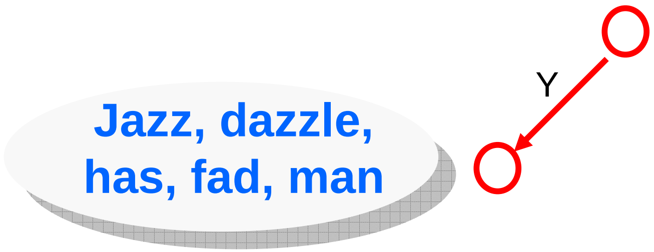
Decision tree clustering

Rule Learning

Find a set of features that best describe how [ae] is realized in British English

**bath, jazz, laugh, dazzle,
has, fad, Man, cat, class, flap,
trap, math, hassle, ask**

Q: Is following phone voiced?



(math)

Decision tree clustering

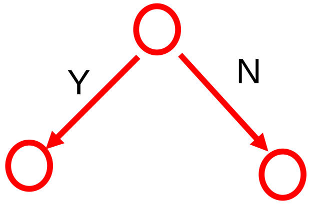
Rule Learning

Find a set of features that best describe how [ae] is realized in British English

**bath, jazz, laugh, dazzle,
has, fad, Man, cat, class, flap,
trap, math, hassle, ask**

Q: Is following phone voiced?

**Jazz, dazzle,
has, fad, man**



**Flap, cat, trap, bath, laugh,
class, math, hassle, ask**

(math)

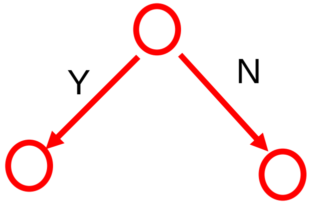
Decision tree clustering

Rule Learning

Find a set of features that best describe how [ae] is realized in British English

**bath, jazz, laugh, dazzle,
has, fad, Man, cat, class, flap,
trap, math, hassle, ask**

Q: Is following phone voiced?



**Jazz, dazzle,
has, fad, man**

**Flap, cat, trap, bath, laugh,
class, math, hassle, ask**

Q: Is following phone a fricative?

(math)

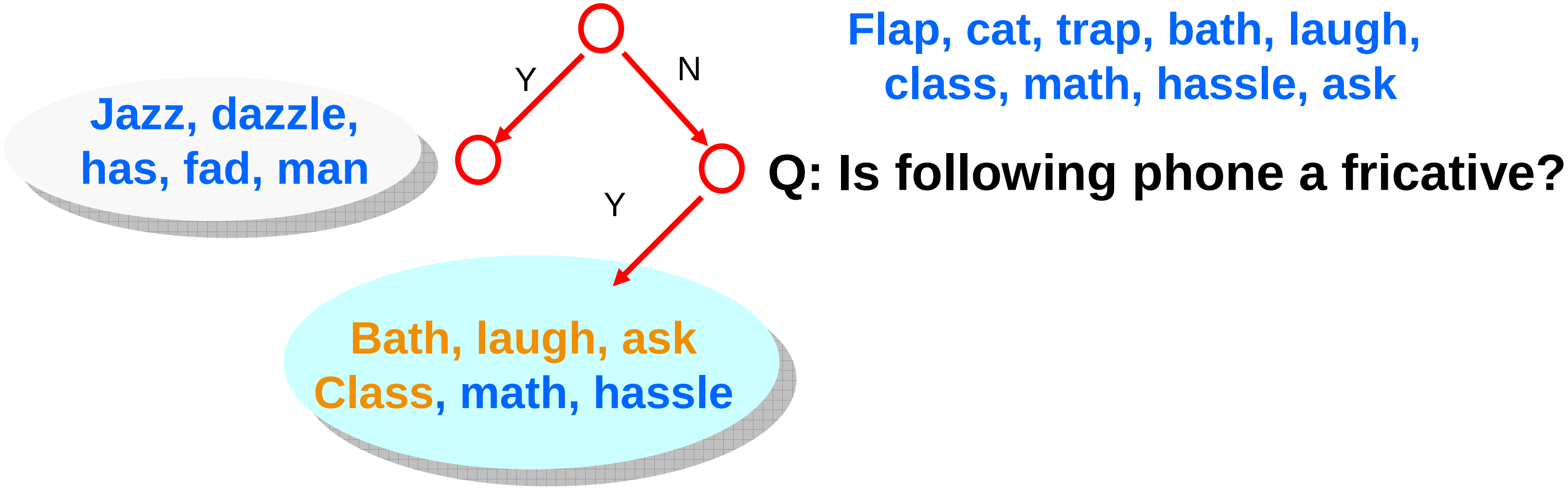
Decision tree clustering

Rule Learning

Find a set of features that best describe how [ae] is realized in British English

**bath, jazz, laugh, dazzle,
has, fad, Man, cat, class, flap,
trap, math, hassle, ask**

Q: Is following phone voiced?



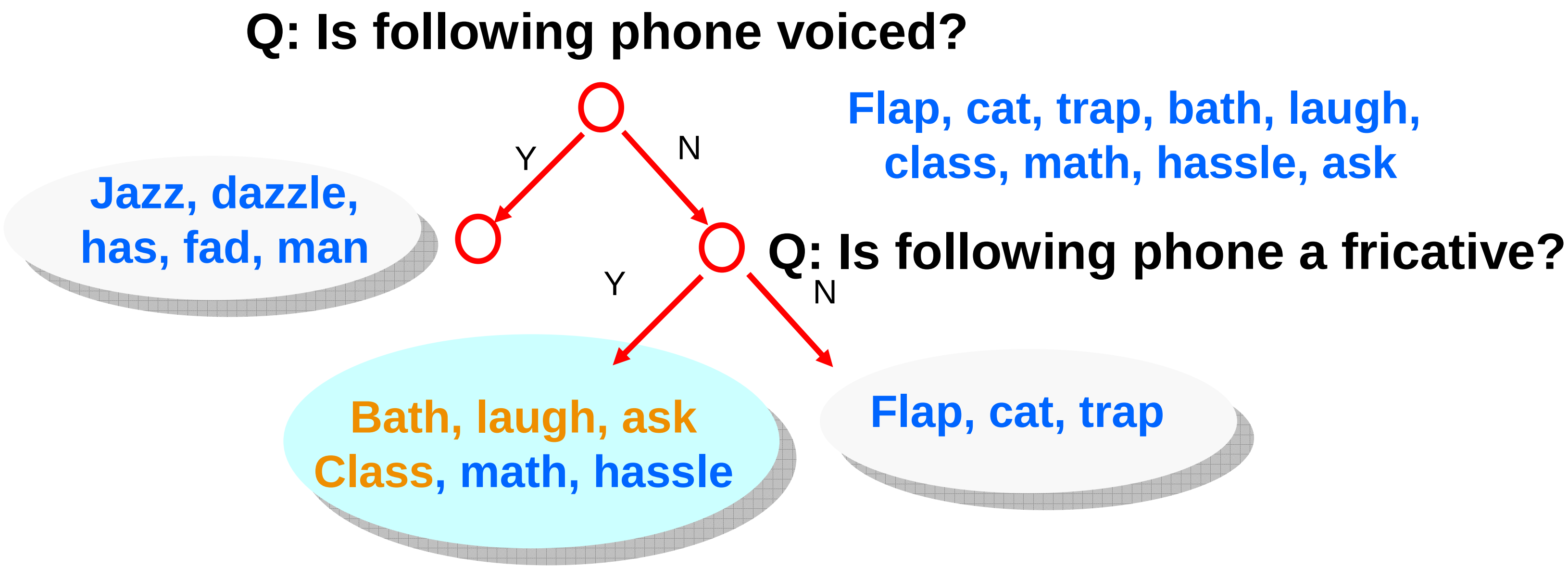
(math)

Decision tree clustering

Rule Learning

Find a set of features that best describe how [ae] is realized in British English

bath, jazz, laugh, dazzle, has, fad, Man, cat, class, flap, trap, math, hassle, ask



(math)

Rule Analysis: Interpretation & Quantification

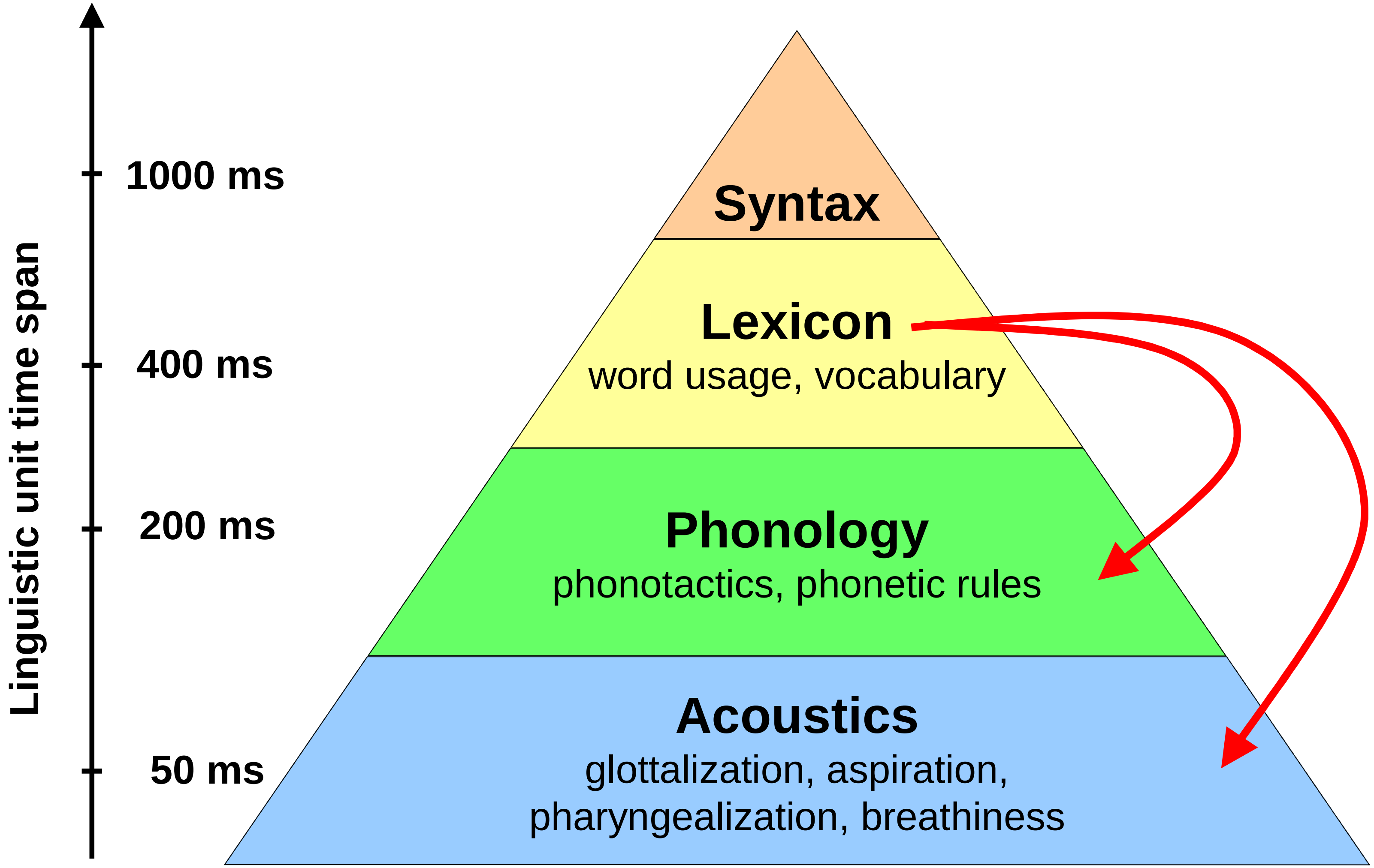
Examples of learned rules from PPM-1

Literature		Proposed System		
Rule	**Dialect**	**Learned Rule**	**Prob**	**Dialect**
Interdental fricatives become stops	EG PS SY	[th] -> [t] / _ [+long]	0.79 0.70 0.87	EG PS SY
		[dh] -> [d] / [-back] _	0.57	EG
		[th] -> [t] / [-short] _ [-long]	0.62	
Vowel [o] exists (usually only [a], [i], [u] exist)	IQ	[o:] -> [u:] / _ [+fricative, -voiced]	0.68	EG
		[o:] -> [a] / _ [+fricative, +voiced]	0.51	

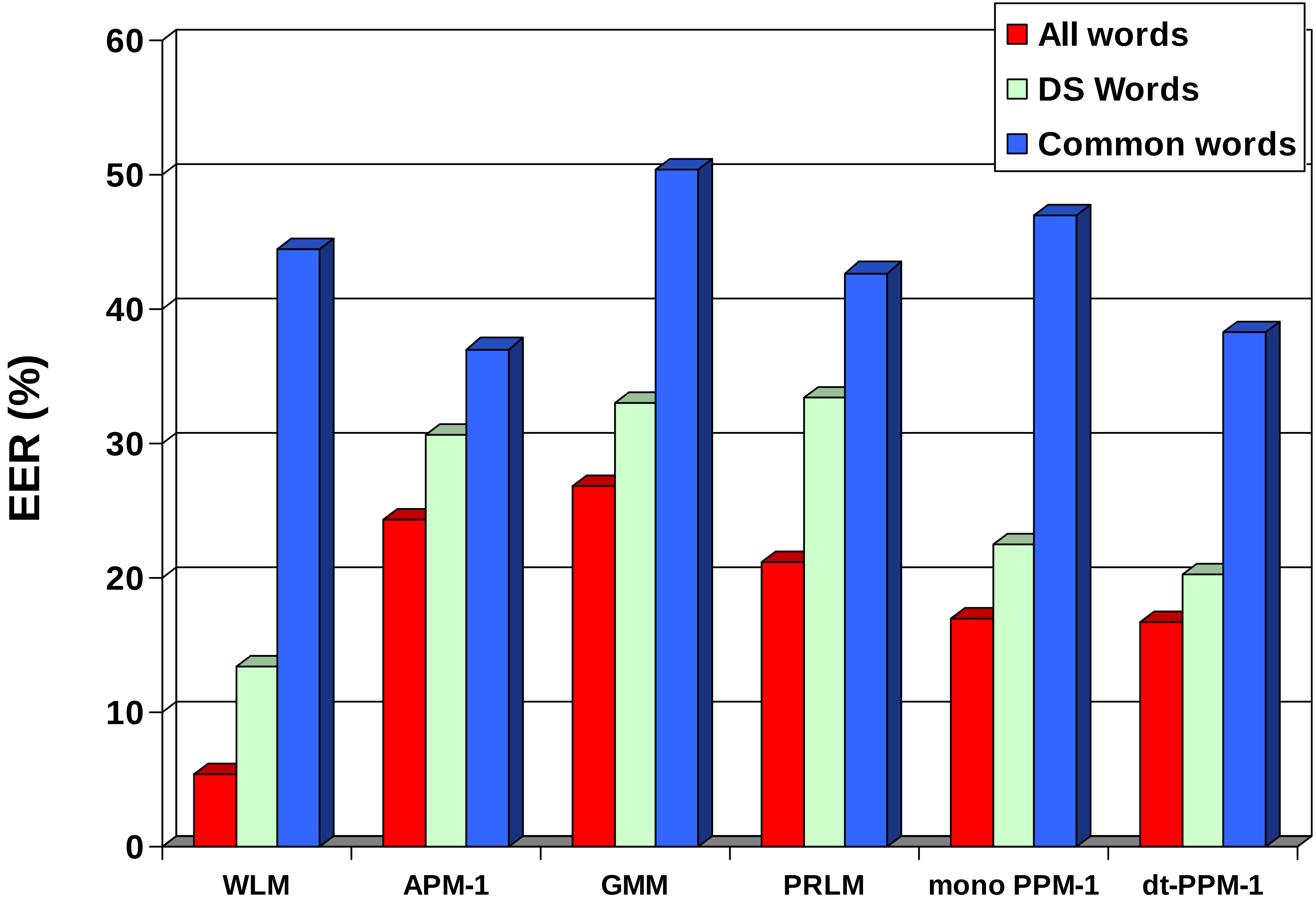
Word Usage Difference Complication

Lexical differences

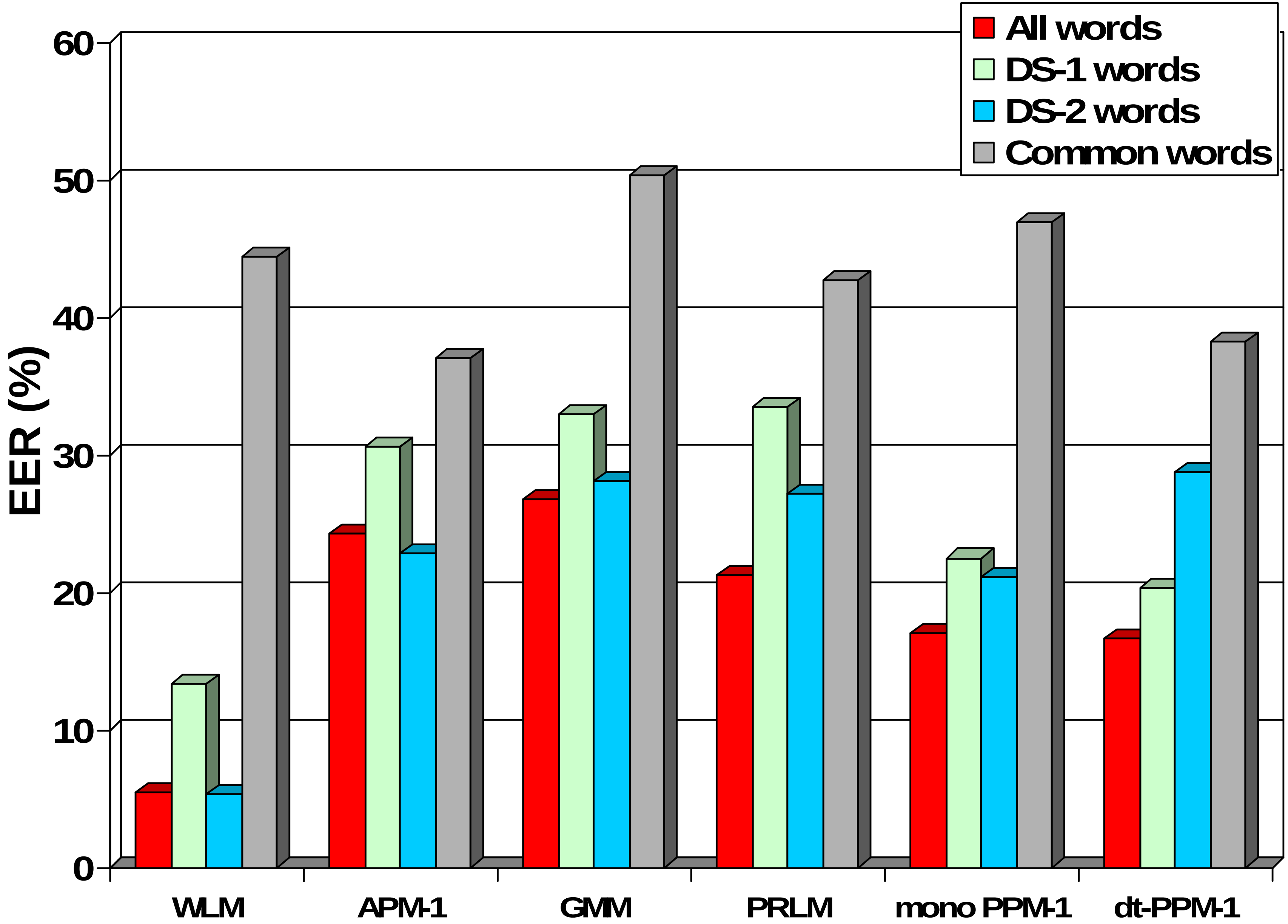
Word usage differences across dialects complicates phonetic characterization



Dialect-Specific (DS) Words Complicate Pronunciation Analysis



Gains from word usage difference



Gains from word usage difference

- DS-1: Words only specific to one single dialect
- DS-2: Words specific to more than one dialect
- Common words:
  - Words that do not help dialect recognition in word language model (WLM)

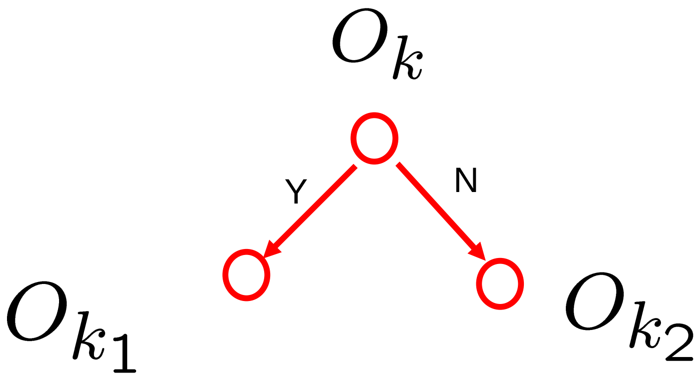
EER performance (%) scoring different types of words

System	All words	DS-1 words	DS-2 words	Common words
1-gram WLM	5.43	13.37	5.35	44.44
APM	22.45	30.65	22.8	37.04
GMM	27.92	33.00	28.14	50.38
PRLM	27.57	33.46	27.17	42.71
Mono. PPM-1	24.03	22.5	21.11	46.97
DT PPM-1	31.28	20.29	28.77	38.26

- All systems score better on dialect-specific words than common words

Math

Decision Tree Clustering



$$\Delta \log L = \log \frac{L(O_{k_1} | x \in H_f) L(O_{k_2} | x \notin H_f)}{L(O_k | x)}$$

$$\widehat{H}_f = \arg\max_{H_f} \Delta \log_L$$

(back)