



---

# Introduction

- Conditional Random Fields (CRFs) offer some benefits over traditional HMM models for sequence labeling
  - Direct model of the posterior probability of a label sequence given an observation
  - Make no assumptions about independence of observations
- The lack of an independence assumption make CRFs an attractive model for speech recognition
- We are interested in combining together arbitrary speech attributes to build a hypothesis of the observed speech

---

# Speech Attributes

- Two different types of speech attributes:
  - Phone classes are trained to indicate when a particular timeslice of speech is a particular phone (e.g. /t/, /v/ etc.)
  - Phonological feature classes are trained to indicate when a particular timeslice of speech exhibits a particular phonological feature

/t/

Manner: stop

Place of articulation: dental

Voicing: unvoiced

---

# Speech Attributes

- Two different types of speech attributes:
  - Phone classes are trained to indicate when a particular timeslice of speech is a particular phone (e.g. /t/, /v/ etc.)
  - Phonological feature classes are trained to indicate when a particular timeslice of speech exhibits a particular phonological feature

/t/  
Manner: stop  
Place of articulation: dental  
Voicing: unvoiced

/d/  
Manner: stop  
Place of articulation: dental  
Voicing: voiced

---

# Speech Attributes

- Two different types of speech attributes:
  - Phone classes are trained to indicate when a particular timeslice of speech is a particular phone (e.g. /t/, /v/ etc.)
  - Phonological feature classes are trained to indicate when a particular timeslice of speech exhibits a particular phonological feature

/t/

Manner: stop

Place of articulation: dental

Voicing: unvoiced

/d/

Manner: stop

Place of articulation: dental

Voicing: voiced

/iy/

Height: high

Backness: front

Roundness: nonround

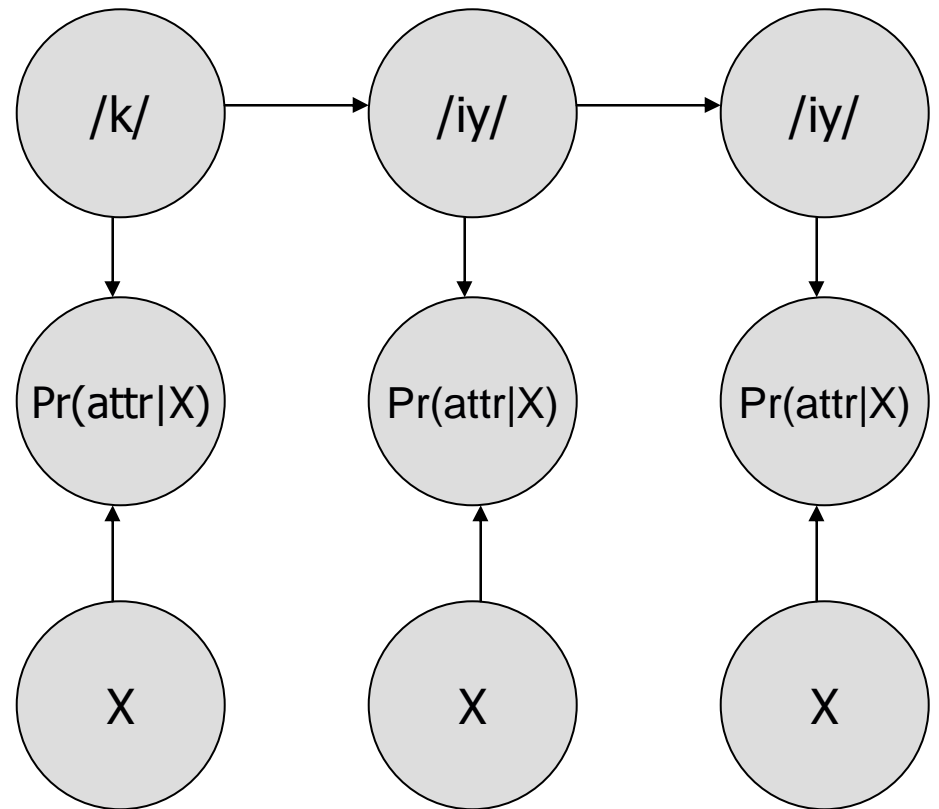
---

# Speech Attributes

- Attribute classifiers are trained using MLP neural networks that emit posterior probabilities –  $P(\text{attribute} \mid \text{acoustics})$ 
  - These posteriors can also be viewed as *indicator functions* for the given classes
  - Outputs are highly correlated with each other
- We want to combine the observations given by these indicator functions to get a hypothesis for the speech

# Tandem Systems

- HMM-based systems using neural network outputs as features (Hermansky and Ellis, 2000)
  - Neural network output is used to train an HMM
  - HMMs assume that the observed features are independent of each other
  - Features are decorrelated through principal components analysis (PCA) before training and testing

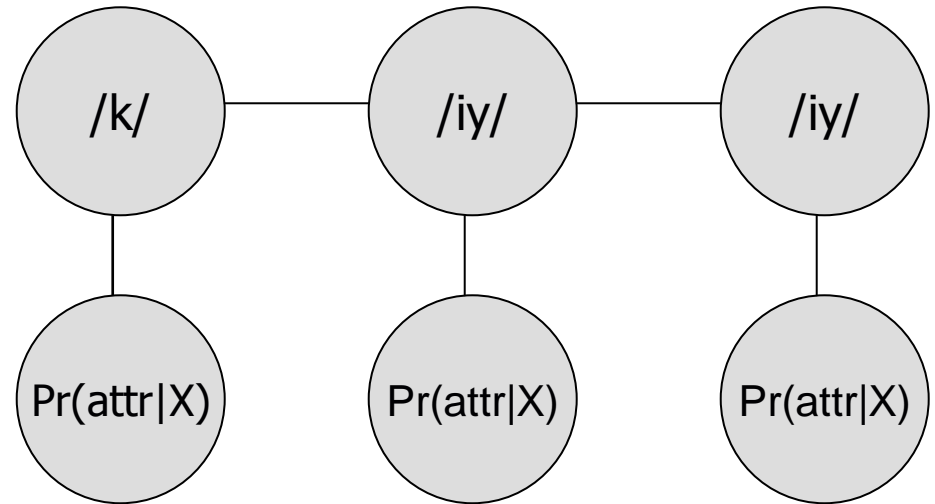


# CRF System

- We implement a CRF model using the neural network outputs as state feature functions

e.g:

$$s_{/d/}(y, \mathbf{x}, t) = \begin{cases} P(/d/ | x_t), & \text{if } y_t = /d/ \\ 0, & \text{otherwise} \end{cases}$$



- Compare the results to a Tandem system trained on the same features
- No PCA decorrelation is performed on the CRF inputs

---

# Phone Accuracy Results

	Tandem (triphones)	CRF (monophones)
Phones (all 61)	67.32%	66.89%
Phon. Features (all 43)	66.85%	63.84%
Combined (all 104)	66.80%	67.87%
Combined (top 39)	67.85%	

---

# Discussion

- The CRF model is much more conservative in its generation than the Tandem model
  - Many fewer insertions, many more deletions
    - All features CRF: 6500 deletions, 731 insertions
    - All features Tandem (top 39): 3184 deletions, 2511 insertions
- Label state space of the Tandem model is much larger than the CRF
- Transition information is currently unused
  - Adding transition feature functions built on observed data may improve results
- Benefit of this model over traditional Tandem model is that arbitrary features can be easily added
  - We want to explore adding arbitrary features to the model to see how performance changes (e.g. speaking rate, stress, pitch, etc.)



---

# Phone Precision Results

	Tandem (triphones)	CRF (monophones)
Phones	73.93%	79.22%
Phon. Features	73.62%	76.61%
Combined	73.33%	79.49%
Combined (top 39)	74.43%	