Detection-Based Accented Speech Recognition Using Articulatory Features

Zhang Chao
CSLT, RIIT, Tsinghua University

Motivation

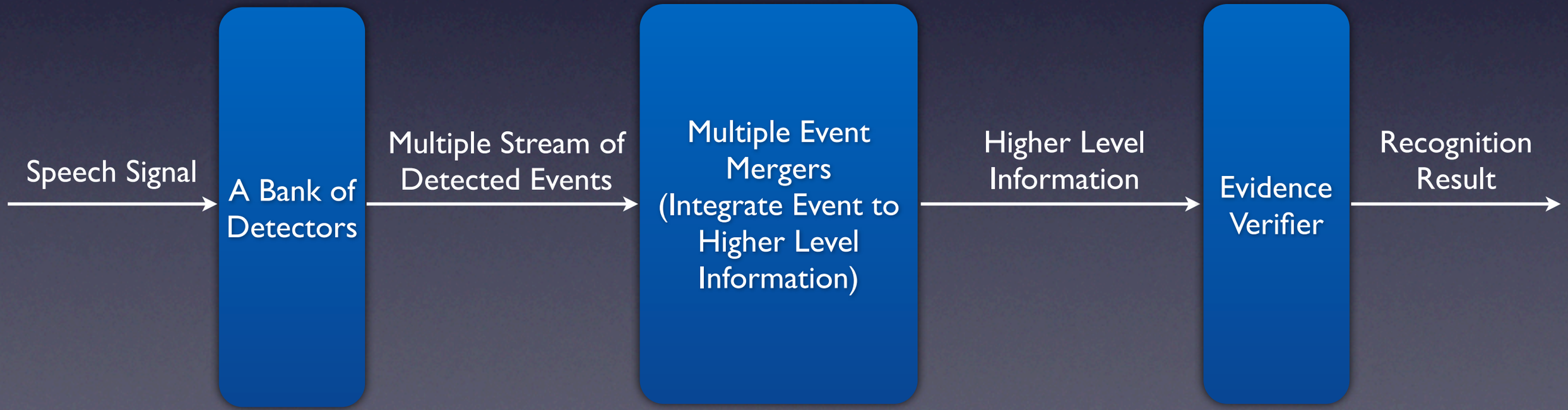
- Traditional approach for accented speech recognition covers hundreds of phoneme changes (e.g. zh_z, sh_s, ...) which is quite sophisticated.
- Linguistic rules for accent variations can be naturally explained with articulatory features.
- If we can cover accented phoneme changes succinctly by articulatory feature changes and shifts.

Motivation

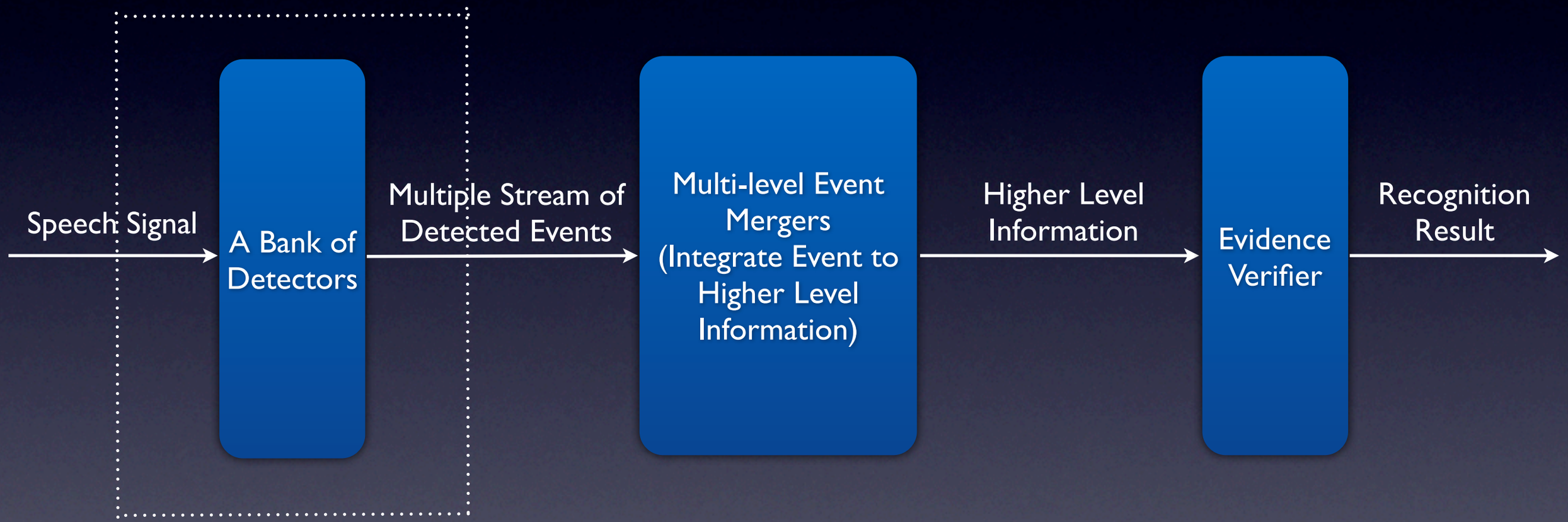
- Traditional approach isn't able to handle accent variations beyond phoneme changes.
    - The other types of accent variations: syllable structures, grammar and vocabularies, tones.
        - We focus on accented phoneme changes in this work.
- If we can cover all types of accent variations using a more advanced system structure.

Motivation

- Detection-based approach - ASAT (automatic speech attribute transcription) paradigm can take the advantage of linguistic information described by features and rules at different levels.

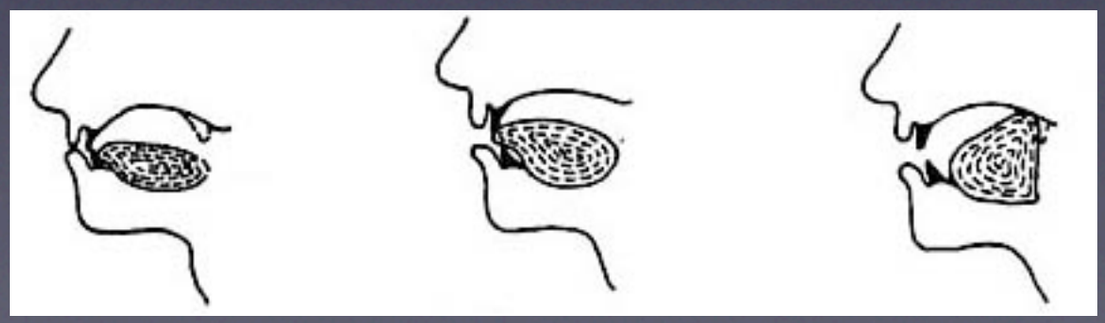


Our System for Accented Speech Recognition



Detectors for Our System

- Articulatory Features: symbolic indicators to characterize how phones are produced by related articulators and the airflow from the lungs.
- Can formulate linguistic knowledge for pronunciation changes caused by accent or coarticulation as context-dependent rules (Non-Linear Phonology).

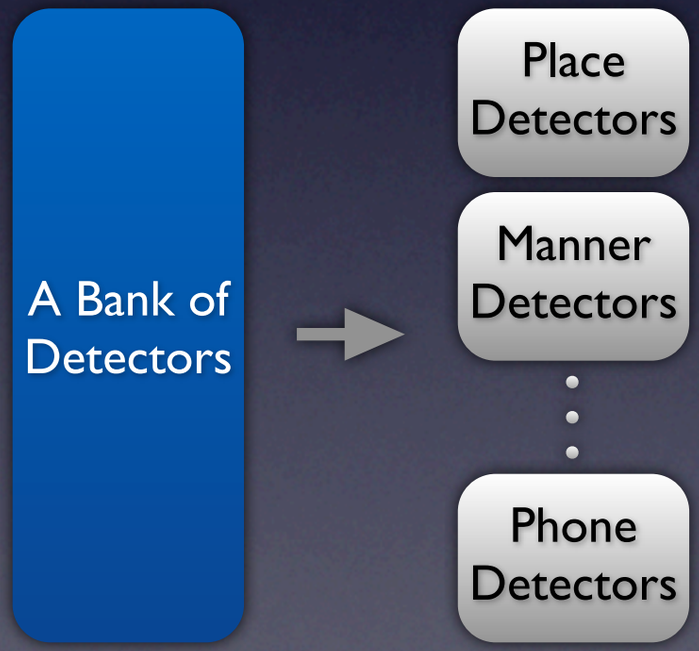


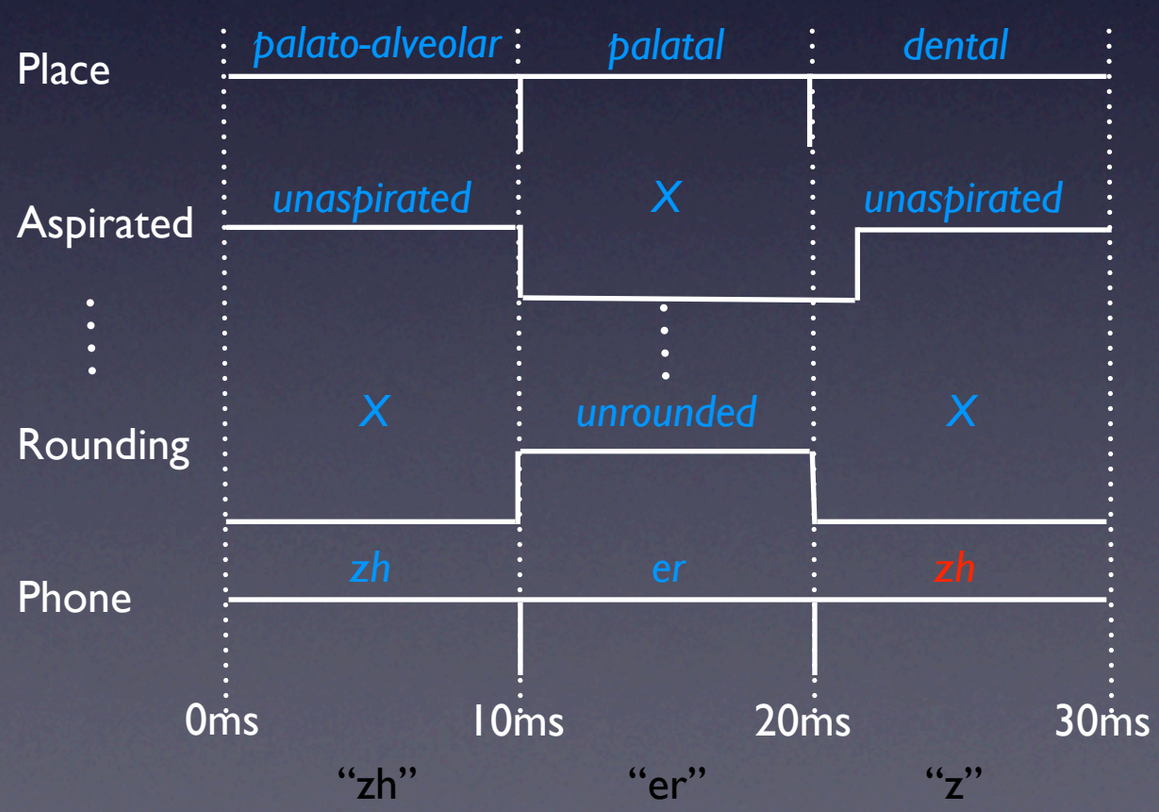
Detectors for Our System

- Articulatory features for Mandarin:
  - 31 attributes of 7 categories: Place, Manner, Aspirated, Voicing, Height, FrontEnd, Rounding
    - Vowel: Place, Height, FrontEnd, Rounding
      - *"er":<palatal,X,X,X,mid_low,central,unrounded>*
    - Consonant: Place, Manner, Aspirated, Voicing
      - *"z":<dental,affricative,unaspirated,unvoiced,X,X,X>*

Detectors for Our System

- A bank of detectors
  - Articulatory features (context-dependent HMMs)
  - Phone (monophone HMMs) (as redundancy)





Accent Changes on Articulatory Features

- Mandarin detector performance
  - Categorized tens of phoneme changes into several confusing articulatory features

Category	*Guanhua*	*Yue*
Place	79.54	68.09
Manner	84.61	83.30
Aspirated	83.85	82.59
Voicing	88.26	88.20
Height	85.31	83.44
FrontEnd	85.59	83.30
Rounding	85.70	85.12

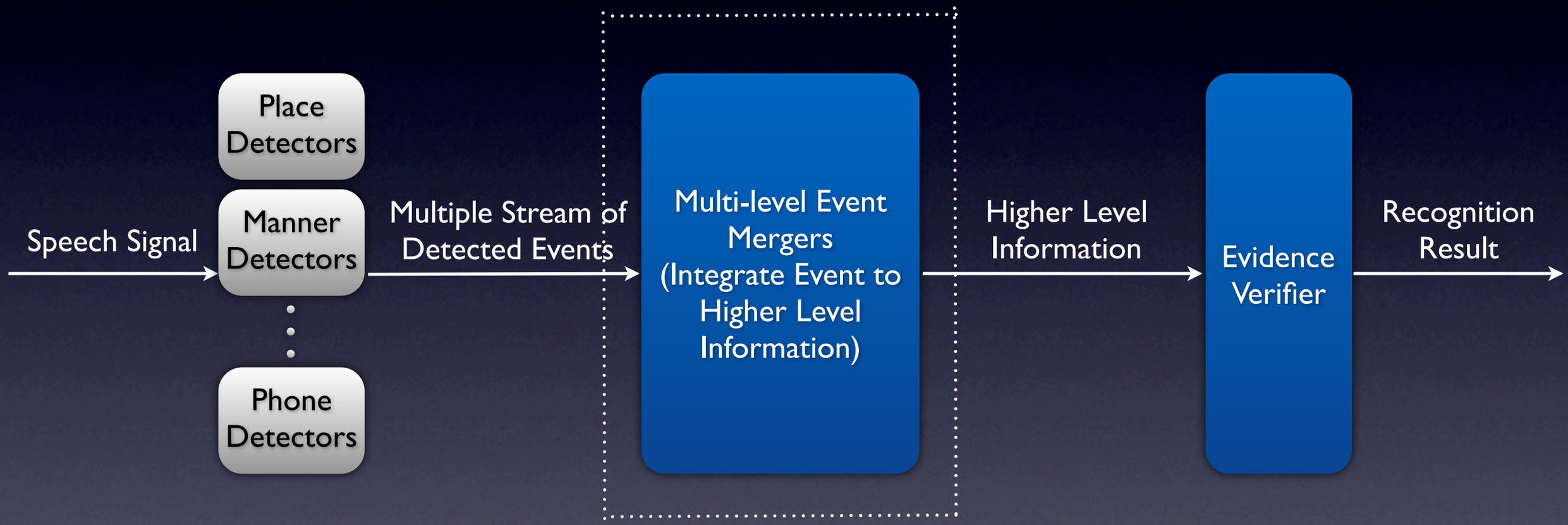
'Place' Attribute	*Guanhua*	*Yue*
bilabial	81.8	79.0
labiodental	92.2	90.9
dental	82.3	76.7
alveolar	75.4	72.3
palato-alveolar	79.3	71.0
palatal	80.7	80.2
velar	80.9	77.3
retroflection	90.8	66.1

zh_z, ch_c, sh_s, ...

er_e

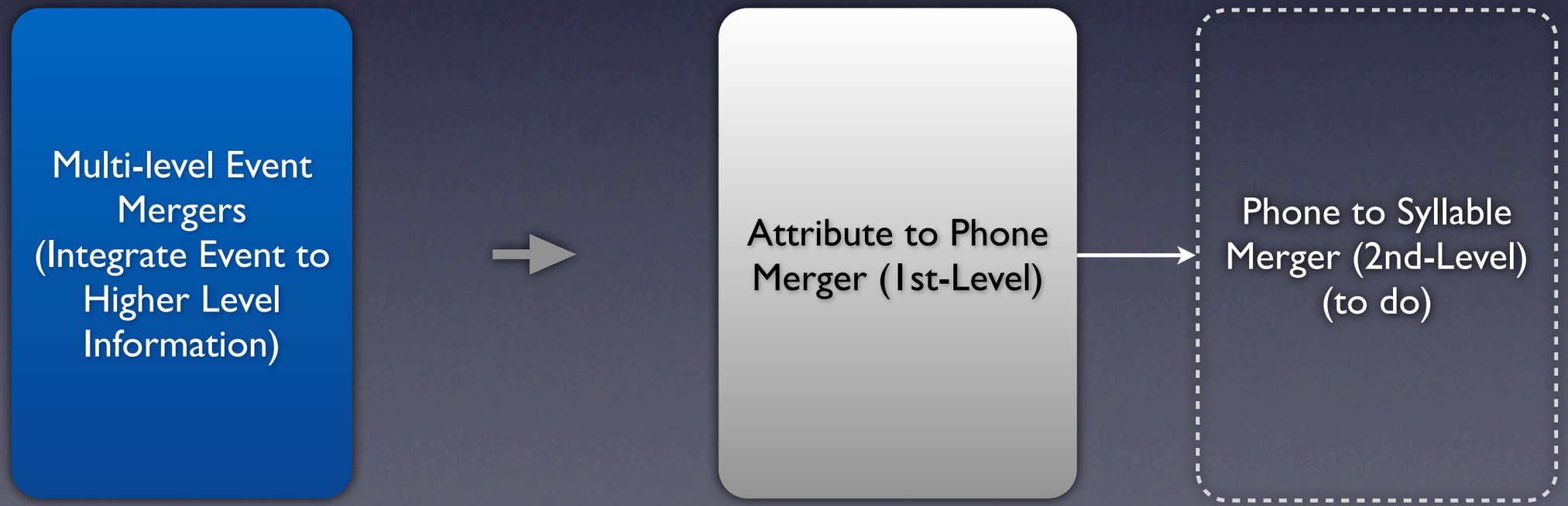
DEL er

Our System for Accented Speech Recognition



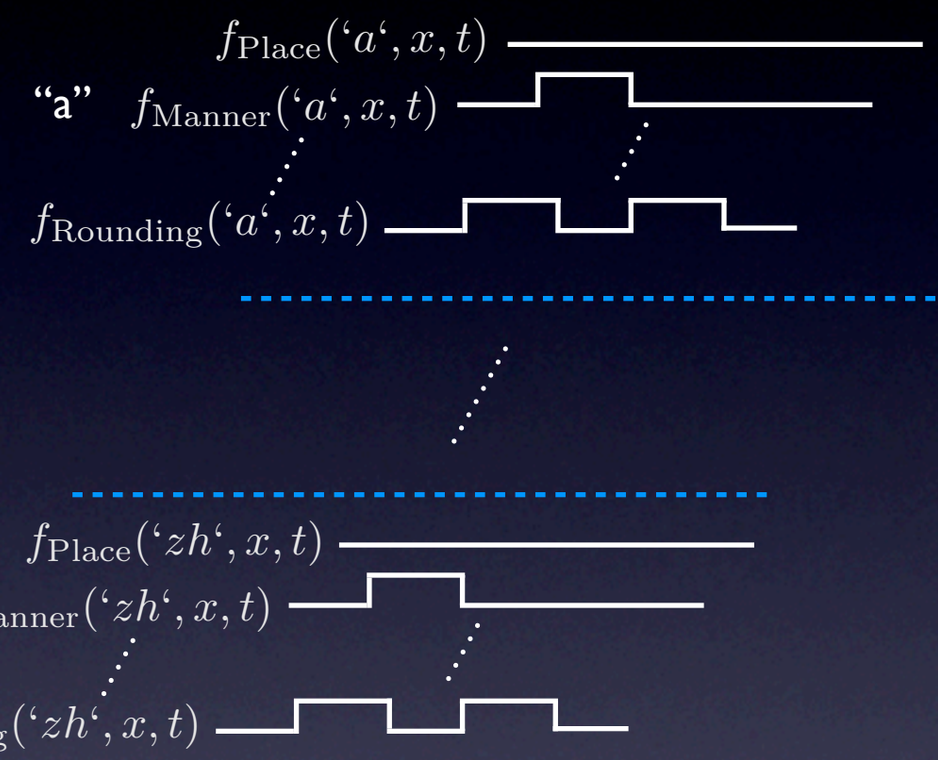
Merger for Our System

- Attribute to Phone Merger
  - Combine the detected event streams using different weights into a probabilistic phone lattice according to linguistic rules from the underlying data.
    - Conditional Random Fields (CRF)
- Phone to Syllable Merger (to do)

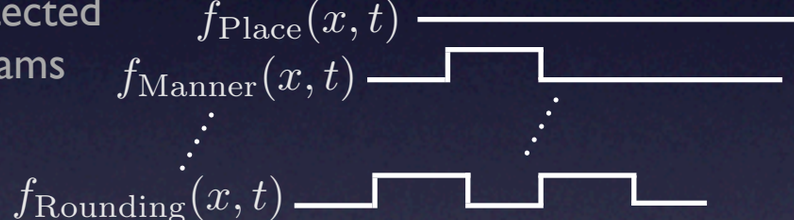


Merger for Our System

- Attribute to Phone Merger
  - How does CRF work?

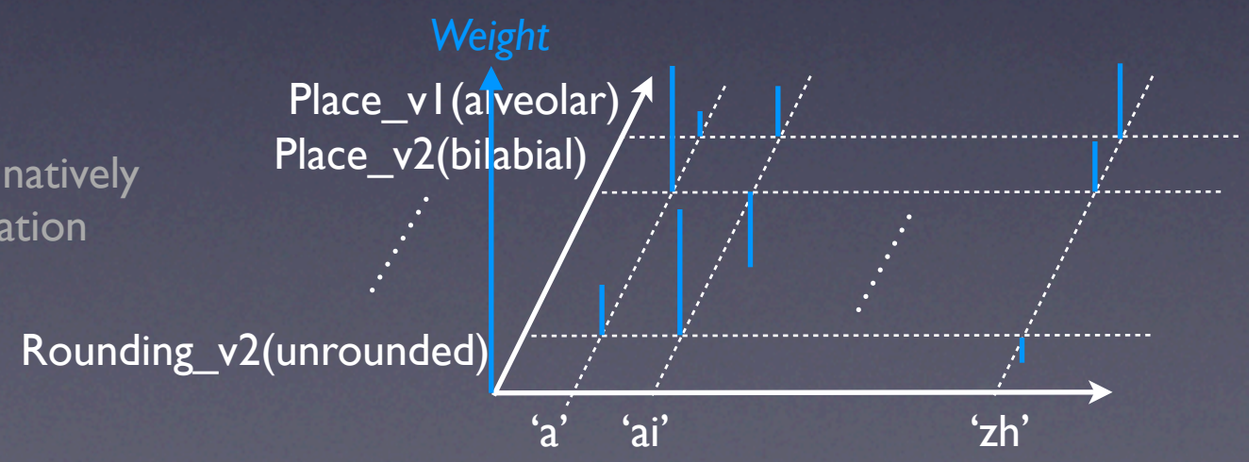


Multiple Detected Event Streams



Feature Extension for CRF

Automatically and Discriminatively Linear Feature Combination



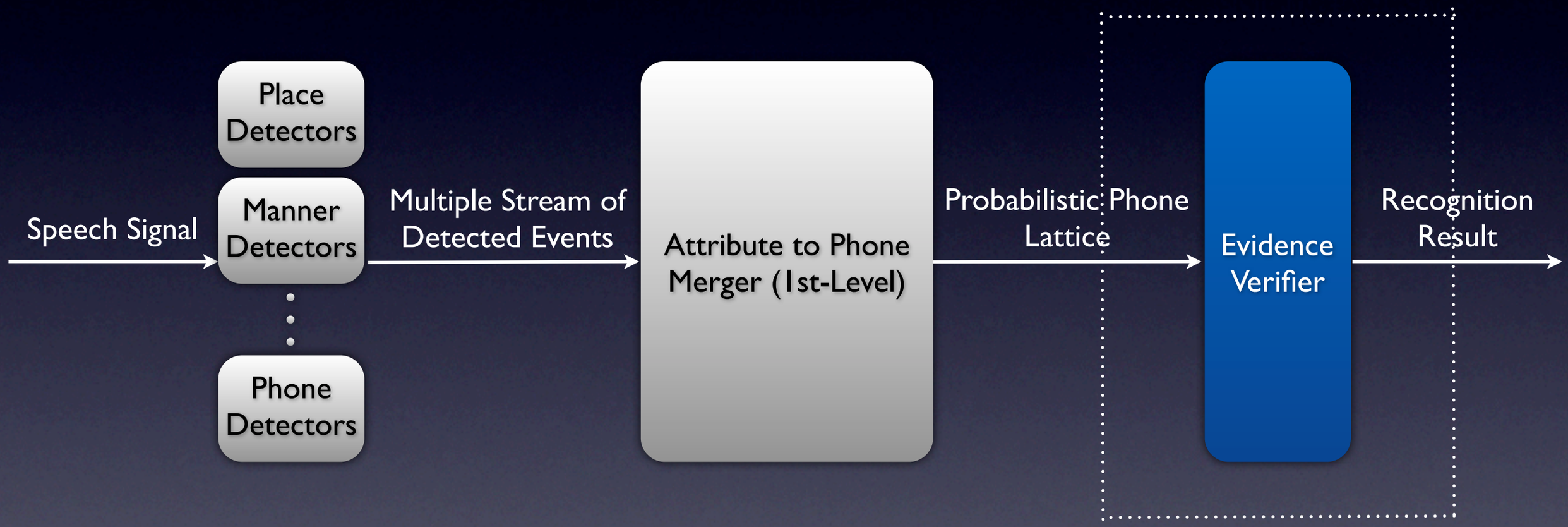
Merger for Our System

- Attribute to Phone Merger
  - Feature types: binary features
  - Adopted features:
    - Presence Feature: absence of each feature.
    - Distinctive Feature: linguistic knowledge for distinguishing each phone unit.
    - Window Feature: Presence Feature and Distinctive Feature for previous 2 and next 2 frames.

Merger for Our System

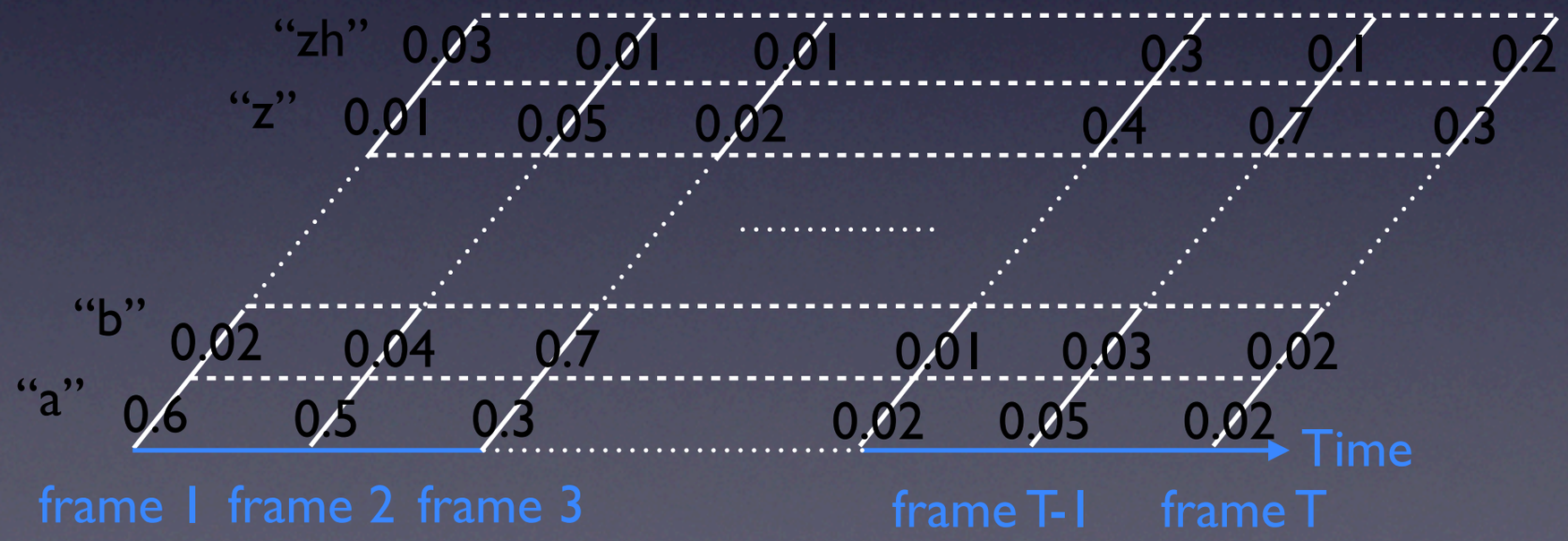
- Attribute to Phone Merger
  - Previous work uses CRF on such task found CRF causes phone undergeneration problem which hurts system performance.
    - We use no transition features (i.e. no CRF language models), thus, CRF regresses to discriminatively trained Logistic Regression.
    - CRF output *probabilistic phone lattice* instead of 1-best phone-level recognition result.

Our System for Accented Speech Recognition



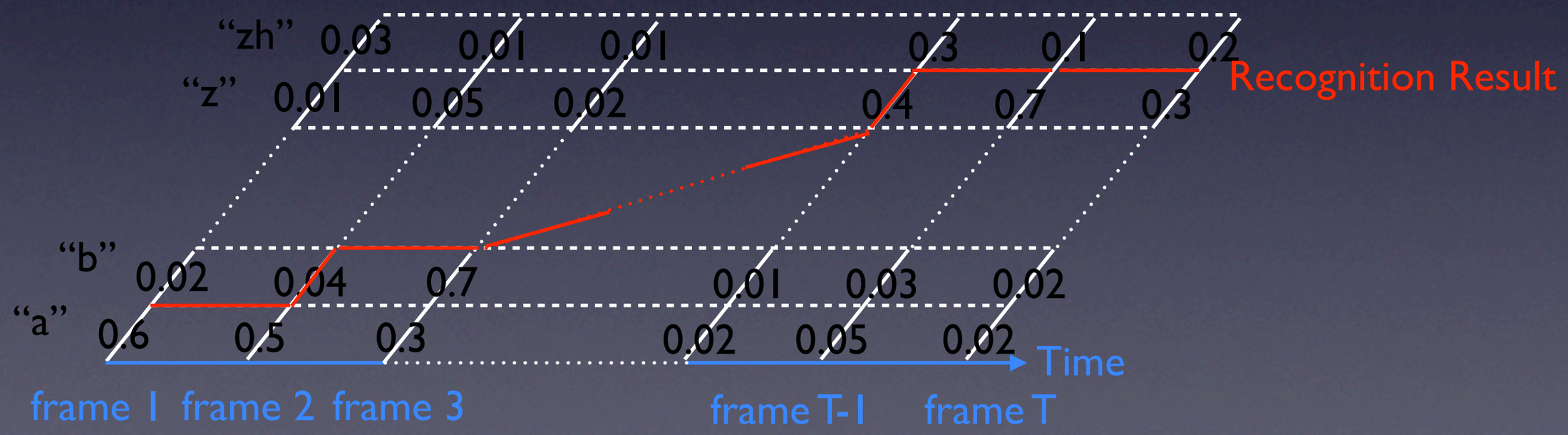
Evidence Verifier

- Probabilistic Phone Lattice
  - Probabilities of each phone at each frame



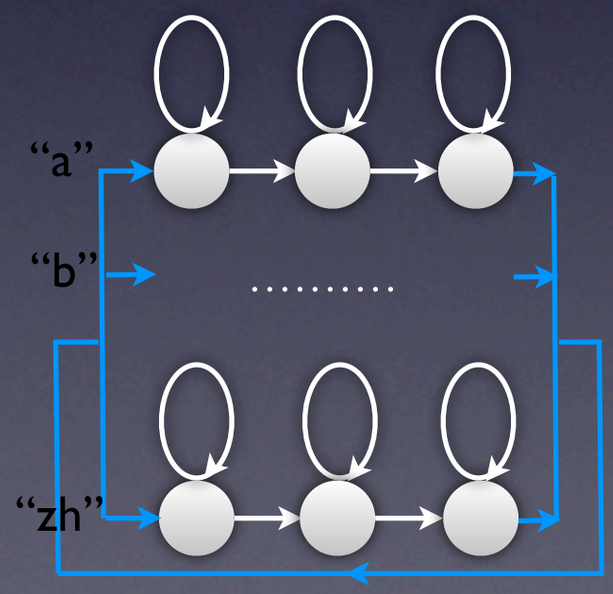
Evidence Verifier

- Probabilistic Phone Lattice
  - Probabilities of each phone at each frame
  - To find the best possible path



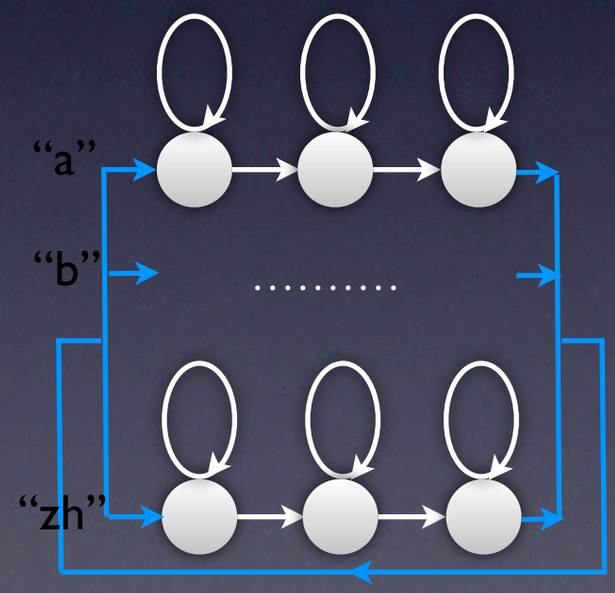
Evidence Verifier

- Various decoding techniques can be used to get the best path.
  - We build free grammar decoding network.
- We build an HMM for each phone.
- The emission probability is obtained by looking up the table (the probabilistic phone lattice).



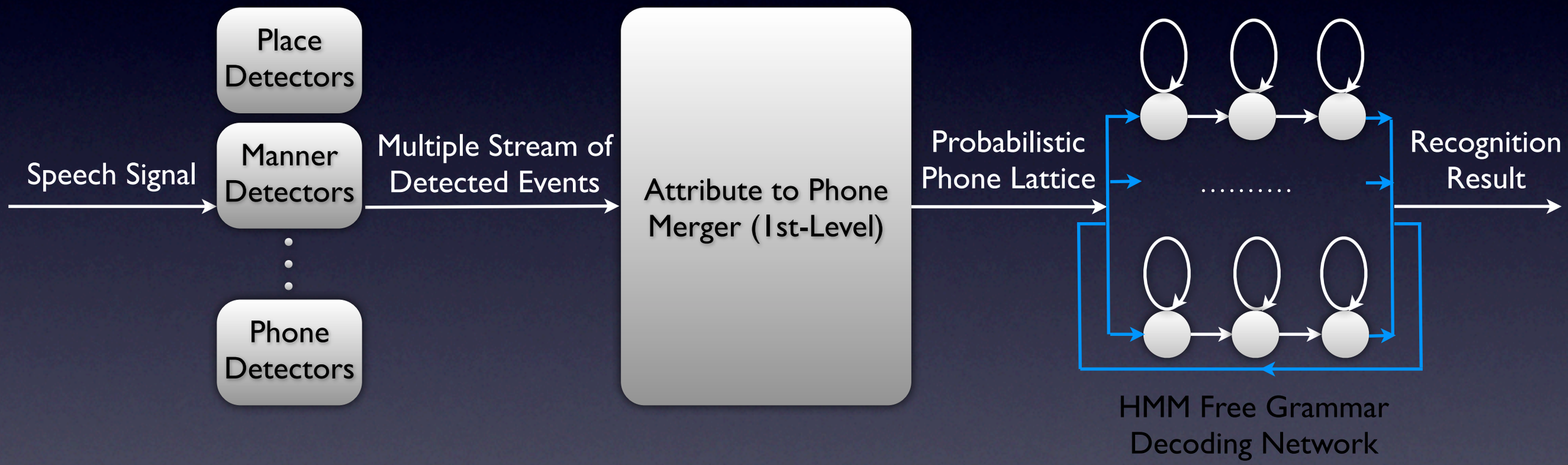
Evidence Verifier

- We solve the phone undergeneration problem by setting a suitable word penalty score.
- We use this free grammar decoding network as the evidence verifier.



HMM Free Grammar
Decoding Network

Our System for Accented Speech Recognition



Evaluation of Our Systems

- All systems are trained and tested on the same accent.
- Our system is 5.71 times faster than triphone HMMs.

System	*Accent*	*Phone Acc.%*
Our Systems	Guanhua	64.48
	Yue	63.06
	Wu	63.59
Monophone HMMs	Guanhua	59.44
	Yue	58.38
	Wu	57.53
Triphone HMMs	Guanhua	66.17
	Yue	64.79
	Wu	64.71

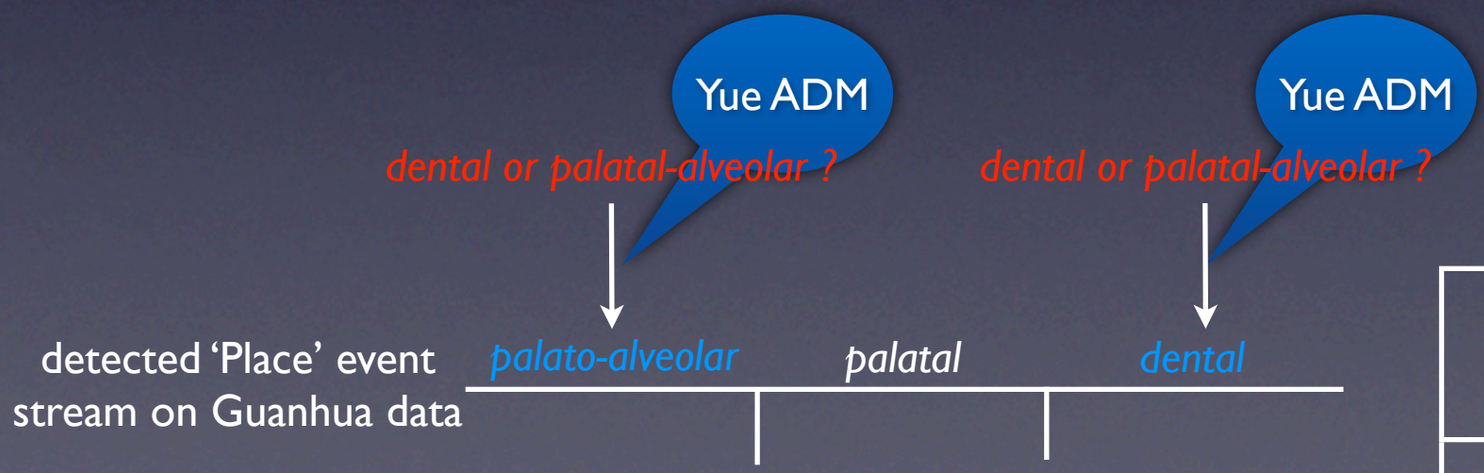
Accent Related Attribute Discrimination Module

- We have shown accent variations can be presented as articulatory feature changes.
- If we can cover the changes explicitly that is similar to traditional pronunciation modeling.
- We propose accent related attribute discrimination module (ADM) to achieve this goal.

Accent Related Attribute Discrimination Module

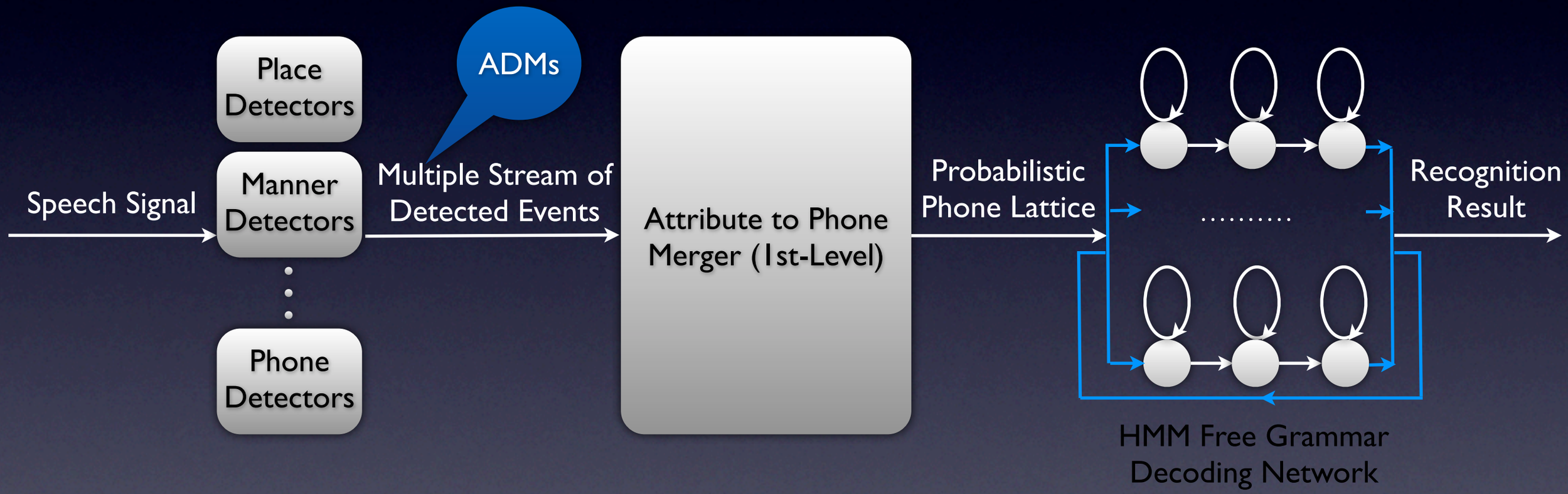
- We use an SVM trained on Yue accented data to reclassify the detected events belong to 'palato-alveolar' and 'dental' attributes of Guanhua system.

- Confusions between 'palato-alveolar' and 'dental':
  - "zh","ch","sh" to "z","c","s"



System	*Guanhua Sys.*	*Guanhua Sys. + Yue ADM*
Phone Acc.%	52.15	59.21

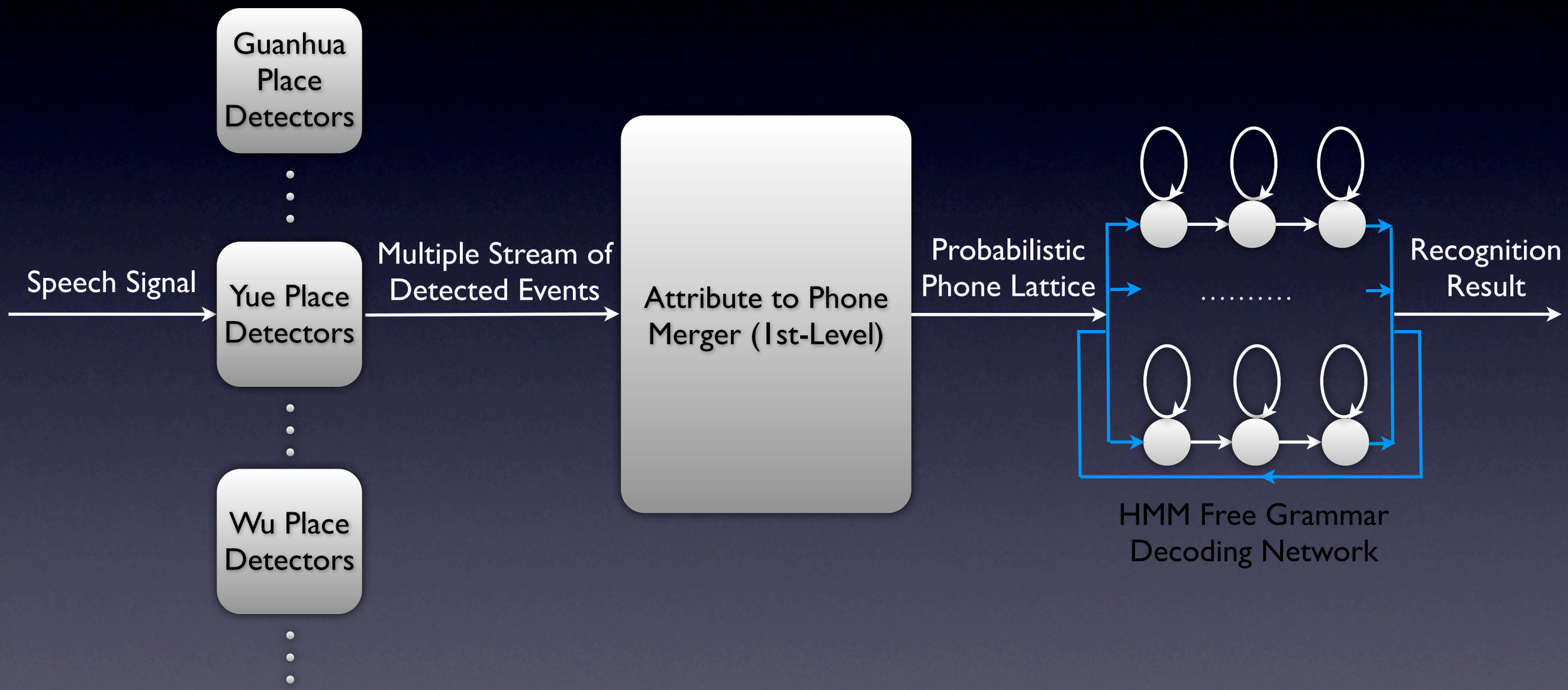
Our System with ADM



Our Robust Multi-Accent System

- Detectors trained on different accents display different regularities on the same accent.
    - It is possible to extract patterns for multi-accent changes given by detectors.
- We use banks of detectors of Guanhua, Yue, and Wu accents together to build a multi-accent robust system.
- Interactive Feature: combinations of the same attributes of different accents .

Our Robust Multi-Accent System



Our Robust Multi-Accent System

- Our System is comparable to triphone HMMs system on phone accuracy.
- Our System performs 3.71 times faster.

System	*Accent*	*Phone Acc.%*
Our Multi-Accent System	Guanhua	65.39
	Yue	66.39
	Wu	67.09
Triphone HMM System	Guanhua	65.40
	Yue	66.34
	Wu	65.87

Conclusions

- We have shown we can present and cover accent variations by articulatory feature changes.
- With ASAT paradigm, more accent changes rather than only phoneme inventory changes could be modeled.
- We proposed a attribute discrimination module which covers accent changes without retraining any model of the system.
- Our proposed systems perform comparable to triphone HMM systems on phone accuracy at a much faster recognition speed.

Thanks for your listening~!