

Speech Technology Days Workshop University of Zaragoza

Where are the Speech Production Models in Our Speech Processing Systems?

Richard Rose
November, 2006

McGill University
Dept. of Electrical and Computer Engineering

OUTLINE

- Motivating Articulatory Based Models for ASR
- Sounds to Words – Problems with Pronunciation Dictionaries
- Phonological Distinctive Features (PDF) for ASR
- Integrating PDF's in ASR

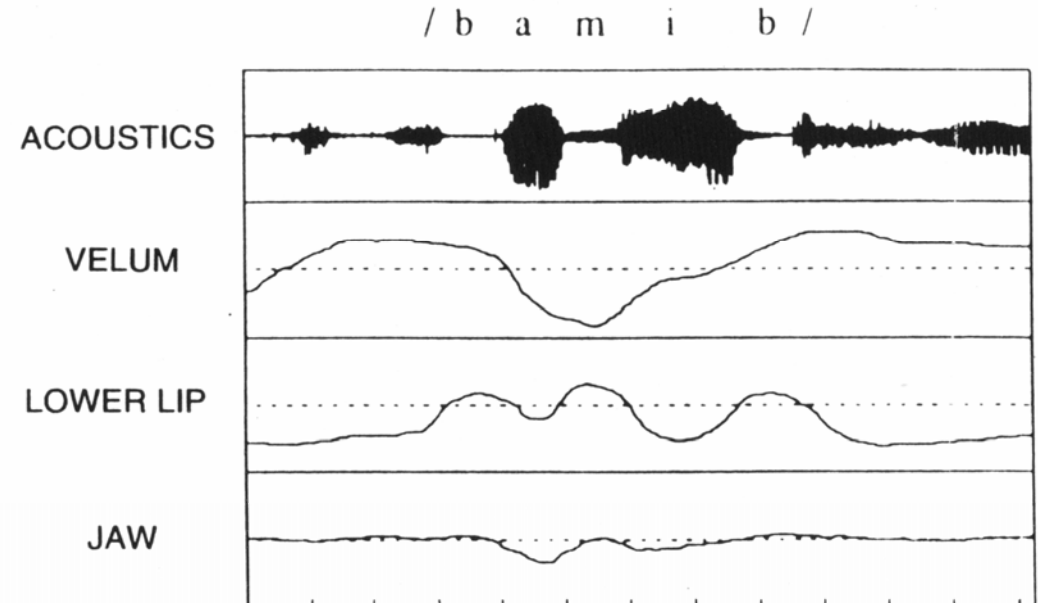
Motivating Articulatory-Based Models for ASR

- A case for Articulatory Representations
 - Speech as an organization of articulatory movements
 - Critical articulators – Invariance in the articulatory space
 - Evidence for usefulness of articulatory knowledge

The Organization of Articulatory Movements

- Speech production can be described by the motion of loosely synchronized articulatory gestures
- Motivates the use of multiple streams of semi-independent phonological features in ASR
- Suggests that segmental, phonemic models are problematic

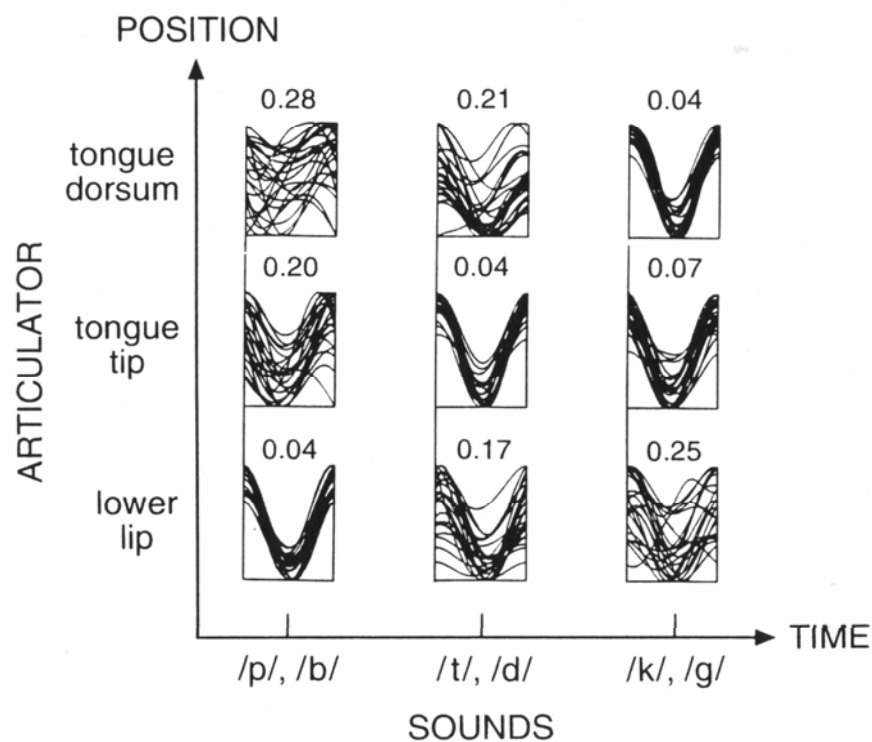
Acoustic waveform and measured articulatory trajectories for utterance of “It’s a /bamib/ sid” (Krakow, 1987)



Reduced Variability Through Critical Articulators

- ASR models with structure defined in an articulatory domain may exploit invariance properties associated with critical articulators
- Critical Articulator: “The articulator most crucially involved in a consonants production”
- Less susceptible to coarticulatory influences
- Less overall variability

Peak-to-Peak Xray microbeam Trajectories

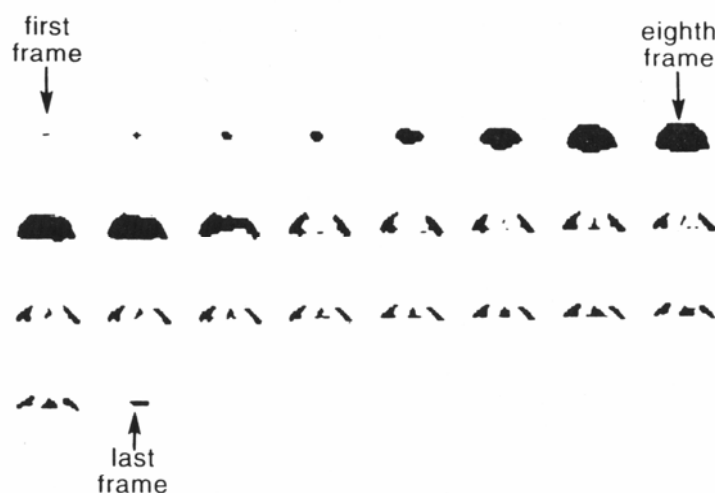


Papcun et al, 1992

Evidence for Usefulness of Articulatory Information

- **ASR Performance Improved using “direct measurements”**
 - **Audio-Visual ASR** [2002 Eurosip Journal on Applied Sig. Proc. Spec. Issue on Joint Audio-Visual Speech Proc.]
 - **Electromagnetic Articulography (EMA)** [Zlokarnik, 1993]

Acoustic ASR	65%	89.4%
Acoustic + Art	78%	94.4%



Sequence of binary lip/tongue images for word “one” (Petajan)



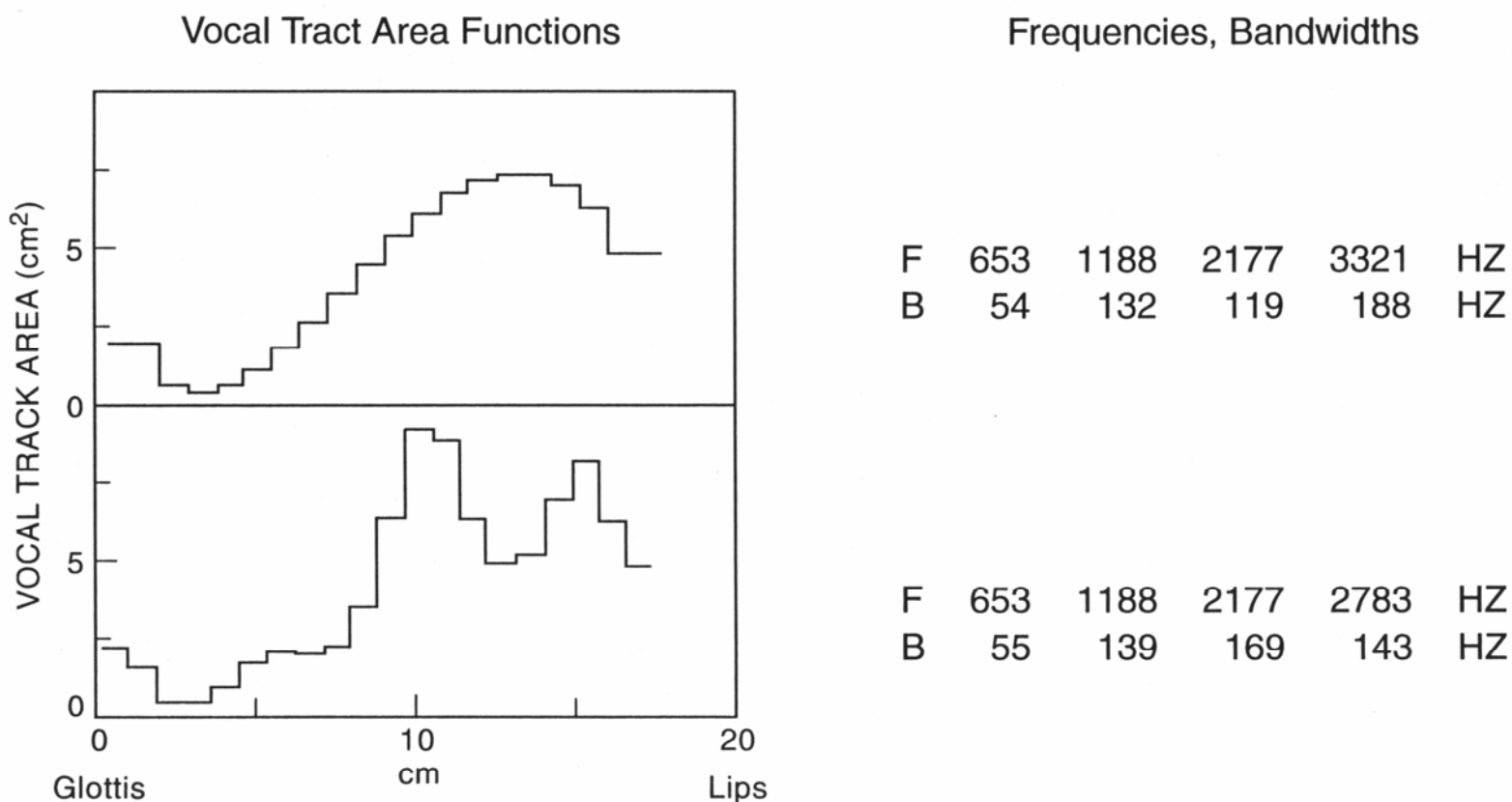
Placement of EMA coils (Zlokarnik)

Motivating Articulatory-Based Models for ASR

- A case for Articulatory Representations
 - Speech as an organization of articulatory movements
 - Reduced variability of critical articulators
 - Evidence for usefulness of articulatory knowledge
- Challenges for Incorporating Articulatory Models
 - One-to-many acoustic to vocal tract area mapping
 - Ambiguity of instantaneous acoustic spectra
 - Coding of perceptually salient articulatory information

Acoustic to Vocal Tract Area Mapping

- Mapping from transfer function to area function is not unique
- Inversion techniques affected by source excitation

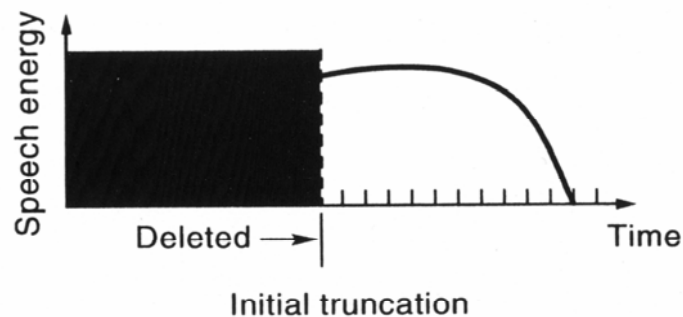


Different Vocal Tract Shapes for Producing Vowel /a/
(Sondhi attributed to Atal)

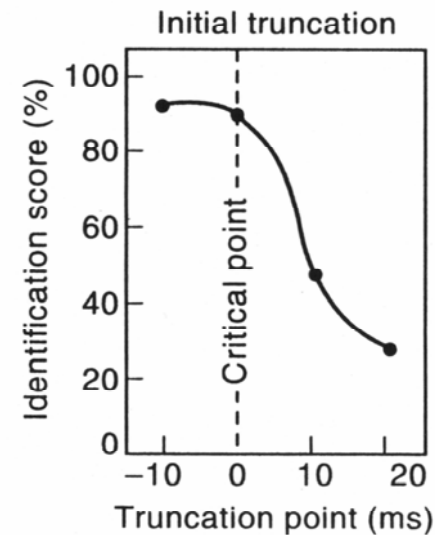


Acoustic Coding of Articulatory Information

- Perceptually salient information necessary for making phonemic distinctions can be contained in fast-varying, short duration acoustic intervals
- Difficult to exploit this information to predict motion of articulators
- Evidence: Japanese CV syllable identification tests [Furui, 1986]



Truncation of Initial
Portion of CV Syllable



Syllable Identification Performance
for Different Truncation Points

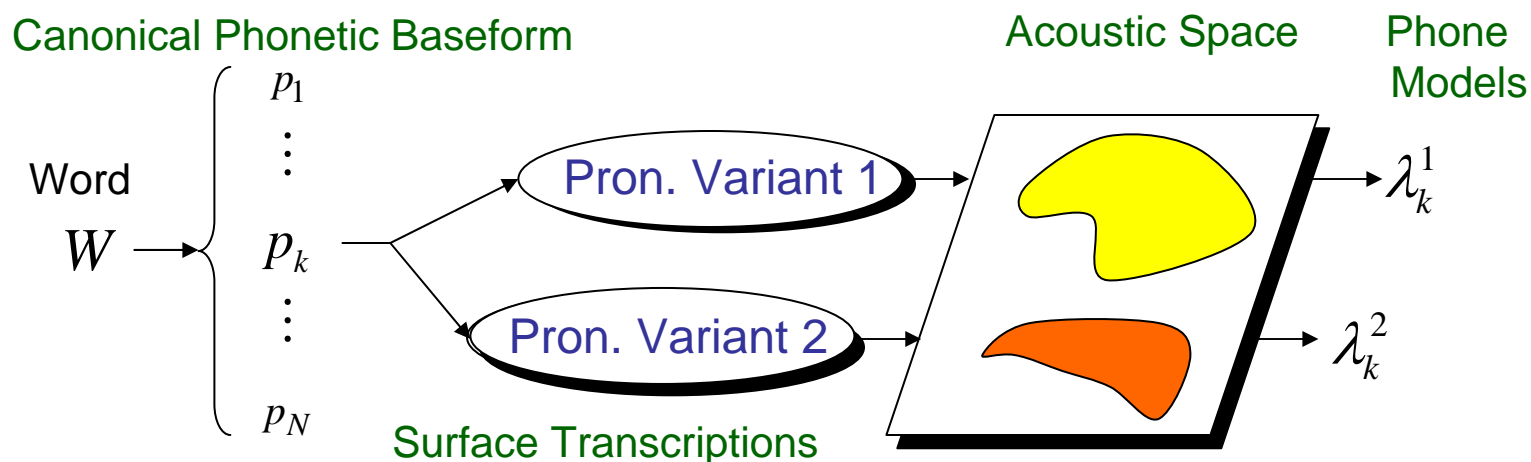
OUTLINE

- Motivating Articulatory-Based Models for ASR
- **Sounds to Words – Problems with Pronunciation Dictionaries**
- Phonological Distinctive Features (PDF) for ASR
- Integrating PDF's in ASR

Sounds to Words – Problems with Dictionaries

Mismatch: Canonical baseforms vs. Surface Form Variant

- Surface-form phone models can be trained using surface acoustic trans.:



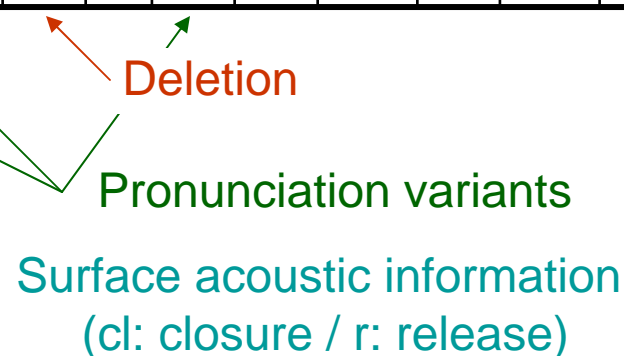
- The challenge is to predict pronunciation variants during recognition:

$$p_k \xrightarrow{?} \{\lambda_k^1, \lambda_k^2\}$$

Problems with Dictionaries

Base-form vs. surface-form pronunciations:

Word	purpose						and			respect							
Base-Form	p	er	p	-	ax	s	ae	n	d	r	ih	s	p	-	eh	k	t
Surface-Form	pr	er	pcl	pr	ix	s	eh	n	-	r	ix	s	pcl	pr	eh	kcl	tr



Canonical Pronunciation Dictionary Coverage vs. Ambiguity

- Adding pronunciation variants to increase coverage can introduce ambiguity among dictionary entries

Word	Canonical Baseform
an	/eh/ /n/
and	/ae/ /n/ /d/
had	/h/ /ae/ /d/
head	/h/ /eh/ /d/
purpose	/p/ /er/ /p/ /ax/ /s/
respect	/r/ /ih/ /s/ /p/ /eh/ /k/ /t/



Impact of Canonical Phonemic Baseforms

- **Speaking Style: Increased speaking rate** [Bernstein et al, 1996]
 - Number of words per second increases with speaking rate
 - Number of phones per second stays roughly the same
 - Phones are deleted, not just reduced
- **Speaking Style: Spontaneous Speech** [Fosler et al, 1996]
 - Switchboard Corpus: ~67% of labeled phones agree with canonical pronunciations
- **Inherent Ambiguity of the Phoneme** [Greenberg, 2000]
 - Inter-labeler agreement for labeling phonemes in spontaneous speech is only 75 to 80 percent

Potential: Huge WAC improvement possible

ASR with “Correct Pronunciations” can increase WAC by 40%

Impact of Canonical Phonemic Baseforms

- Better modeling of surface-form phones does not increase WAC
- Demonstration: TIMIT Corpus
 - Train context dependent HMM phone models from
 - Surface-form (S-F) acoustic transcriptions – manually labeled
 - Base-form (B-F) transcriptions – From canonical pronunciations

Word Trans.	purpose					and			respect						
Base-Form Trans.	p	er	p	ax	s	ae	n	d	r	ih	s	p	eh	k	t
Surface-Form Trans.	p	er	p	ix	s	ix	n	-	r	ix	s	p	eh	k	-

- Compared phone accuracy (PAC) and word accuracy (WAC) using S-F and B-F HMM models

Impact of Canonical Phonemic Baseforms

- Better modeling of surface-form phones does not increase WAC
- Demonstration: TIMIT Corpus
 - Train context dependent HMM phone models from
 - Surface-form (S-F) acoustic transcriptions – manually labeled
 - Base-form (B-F) transcriptions – From canonical pronunciations
 - Phone accuracy (PAC) and word accuracy (WAC)

HMM Training Transcriptions	Phone Acc. S-F Trans.	Phone Acc. B-F Trans.	WAC B-F Dict.
Surface-form	69.1%		92.0%
Base-form		63.3%	96.1%

- HMMs trained from S-F trans. provide best model of acoustic variants
 - ... But this does not result in better ASR word accuracy

OUTLINE

- Motivating Articulatory-Based Models for ASR
- Sounds to Words – Problems with Pronunciation Dictionaries
- **Phonological Distinctive Features (PDFs) for ASR**
- Integrating PDF's in ASR

Phonological Distinctive Features (PDFs) for ASR

- Few ASR systems exploit direct Articulatory Measurements
 - Exception is research in audio-visual ASR [2002 Eurosip Journal on Applied Sig. Proc. Spec. Issue on Joint Audio-Visual Speech Proc.]
 - Other examples - low power radar sensors (GEMS) [Fisher,2002]
- Many ASR systems exploit phonological distinctive features
 - Defined to describe articulatory phenomena
 - Form the basis of speech production rules [Chomsky and Halle, 1967]

Phonological Distinctive Features (PDFs) for ASR

- Example of multi-valued definition of PDFs [King et al, 2000]

Feature	Values
Manner of Articulation	Vowel, Fricative, Approximant, Nasal
Place of Articulation	Low, Mid, High, Palatal, Labial, Coronal-Dental, Labial-dental, Labial, Coronal, Velar, Glottal ...
Phonation	Voiced, Unvoiced
Centrality	Central, Full, Undefined
Continuant	Continuant, Non-continuant
Front-back	Back, Front
Roundness	Round, Not-Rounded
Tenseness	Lax, Tense

- Many other definitions of Features
 - Binary PDFs [Chomsky and Halle, 1967]
 - Articulatory Features [Deng and Sun, 1999] [Bridle et al, 1998]

The Case for PDFs in ASR

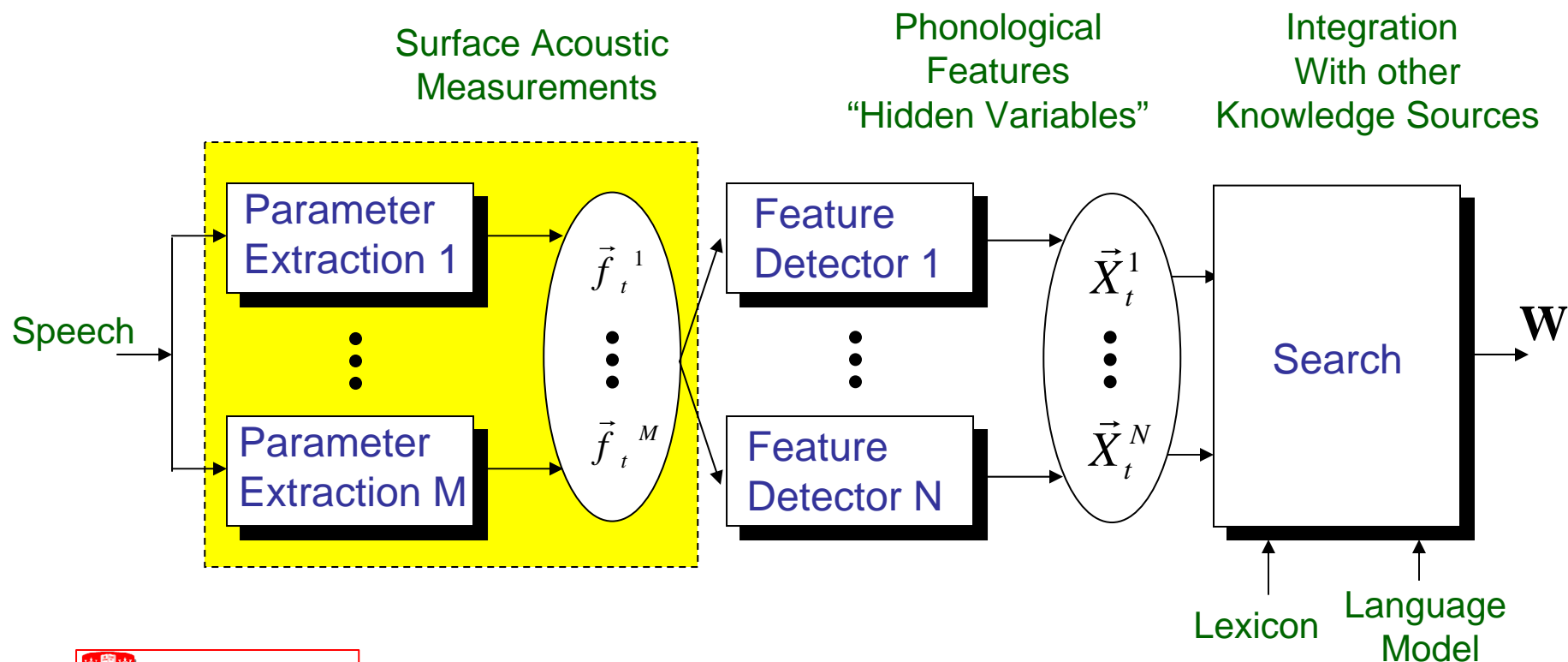
- PDFs used as a “hidden process”
 - Exploit advantages of **articulatory based representation**
 - **Overlapping**, as opposed to segmental, models of speech
 - **Invariance** properties associated with critical articulators
- Incorporating speech production based models into ASR search
 - **Pronunciation variation** – Changes to canonical baseforms modeled as time evolution of synchronized features [Livescu, et al, 2004]
 - **Articulatory HMM state space** – States described in terms of “feature spreading” [Deng and Sun, 2000]
 - **Adaptation/Normalization**: Can be made more efficient when performed in a “physiologically plausible” space (VTLN)

Issues for Incorporating PDFs in ASR

- Extracting acoustic correlates of PDFs from the surface acoustic waveform
- Reliable detection of PDFs from acoustic correlates
- Practical advantages for PDF-based ASR systems over MFCC-based ASR

Phonological Distinctive Features (PDF) for ASR

- Obtaining Acoustics Correlates of PDFs from Surface Acoustic Waveforms
 - Acoustic Correlates: Relationship between S-A parameters and PDFs



Obtaining PDF's from Surface Acoustic Measures

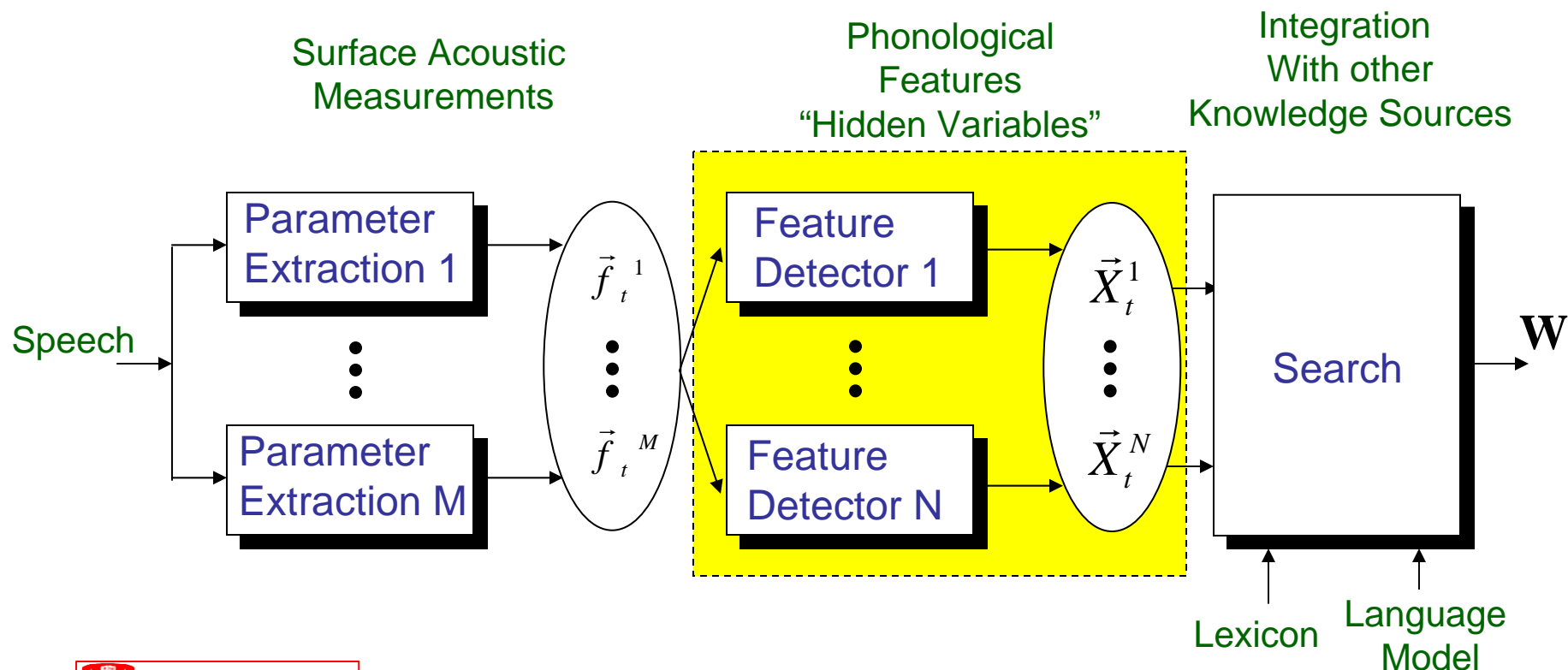
- Define acoustic correlates for a feature
- Determine acoustic parameters that characterize acoustic correlates
 - Example: acoustic parameters for stop consonants [Epsy-Wilson]

Feature	Acoustic Correlates	Acoustic Parameters
Stop consonant (non-continuant)	Closure followed by abrupt spectral change	Closure: Energy: 0.2-3KHz Energy: 3-6KHz ACorr: $R(1)/R(0)$
		Burst: Spectral Flatness

- Acoustic parameters and feature detectors
 - Feature space transformations (LDA) and feature selection algorithms allow acoustic parameters to be identified from candidate params.

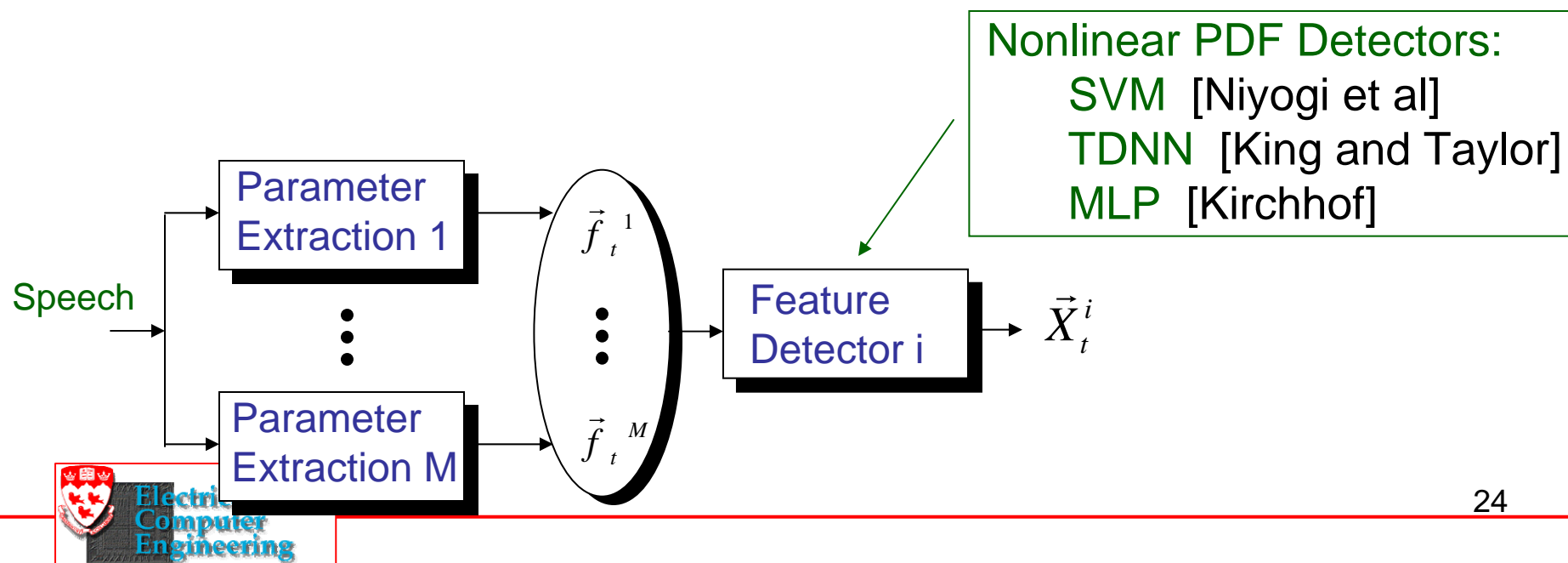
Phonological Distinctive Features (PDF) for ASR

- Detecting PDFs from Acoustic Parameters
 - Non-linear relationship between acoustic and articulatory distances



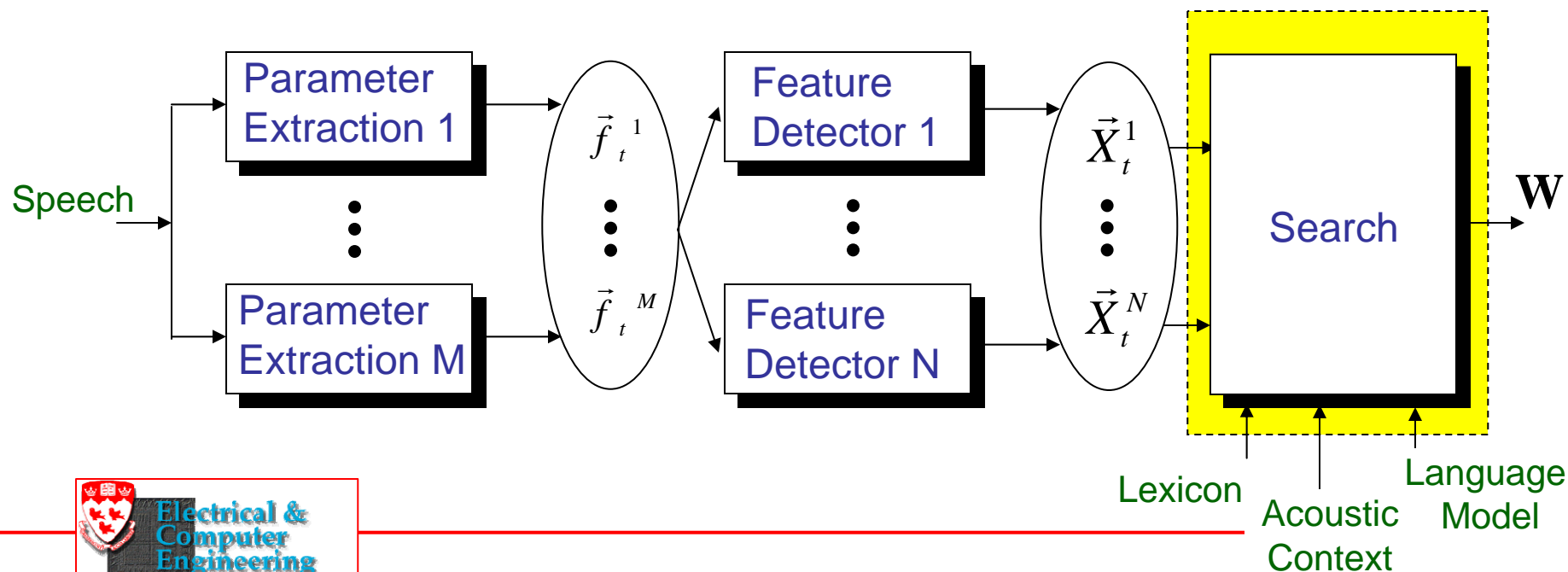
Detecting PDFs From Surface Acoustic Parameters

- Relationship between articulatory distances and acoustic distances can be highly nonlinear [Niyogi et al, Stevens et al]
- Only small regions of acoustic space correspond to regions of high articulatory discriminability
- Fits nicely as a problem for support vector machines (SVM)



Phonological Distinctive Features (PDF) for ASR

- Practical Advantages of Model Structure Based on PDFs
 - HMM State Space: Model topology defined by feature spreading
 - Pronunciation: Feature based description of pronunciation variation
 - Adaptation/Normalization: More efficient when performed in a “physiologically plausible” space



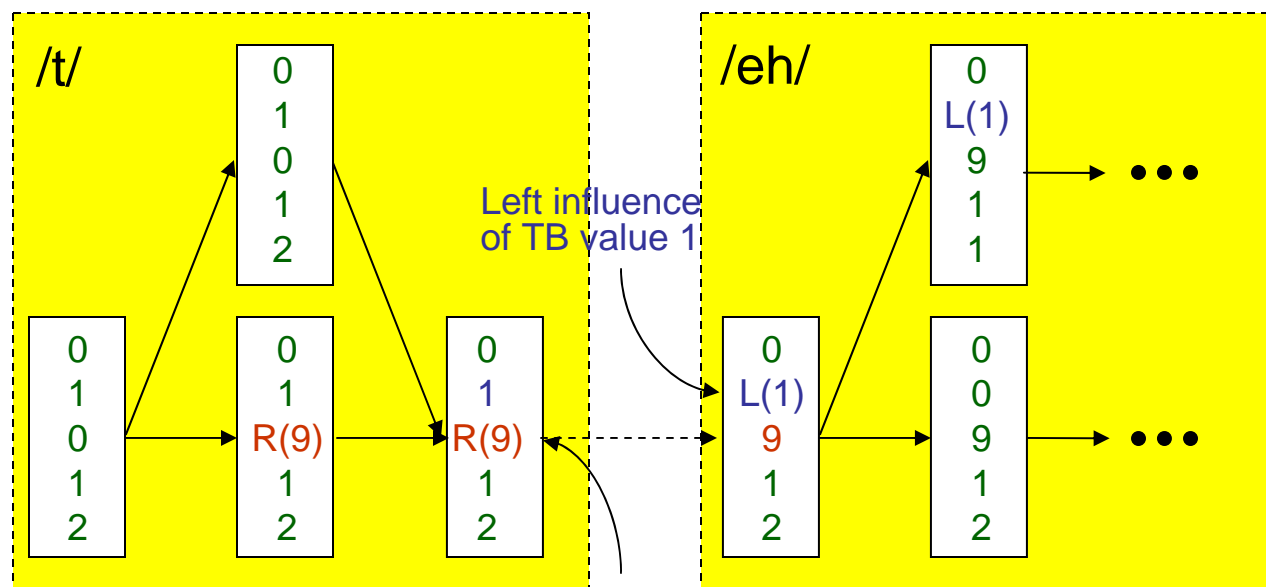
Modeling Structure Based on PDF's

- PDF Based HMM state space [Deng and Sun, 1999]
 - Phones in context defined in terms of articulatory features
 - Context specific nodes formed by spreading features
 - PDF based nodes permit defining context in articulatory space

Phone in Context Models – State Trans. Graphs

HMM States defined as Multi-valued Articulatory Features

- | |
|---------------|
| Lips |
| Tongue Body |
| Tongue Dorsum |
| Velum |
| Larynx |



Modeling Structure Based on PDF's

- PDF Based models of pronunciation variation [Livescu et al, 2004]
 - PDFs model asynchrony of articulators and articulatory dynamics
 - Model structure based on dynamic Bayesian networks (DBNs)
- Canonical Dictionary Expanded as PDFs [Livescu et al, 2004]

PDF
Baseform
Dictionary

Word	and		
Phones	ae	n	d
Index	0	1	2
Phonation	Voiced	Voiced	Voiced
Manner	Vowel	Nasal	Occlusive
Place	Low	Coronal	Coronal
Continuant	Continuant	Non-Continuant	Non-Continuant

Canonical Articulatory Baseforms

- Canonical Dictionary Expanded as PDFs [Livescu et al, 2004]

PDF
Baseform
Dictionary

Word	and		
Phones	ae	n	d
Index	0	1	2
Phonation	Voiced	Voiced	Voiced
Manner	Vowel	Nasal	Occlusive
Place	Low	Coronal	Coronal
Continuant	Continuant	Non-Continuant	Non-Continuant

- Probabilistic Models of Feature Asynchrony and Feature Substitution

Articulatory
Asynchrony

Manner Index	0	0	1	1	1	2	2	2
Place Index	0	0	0	0	1	1	1	2

Asynchrony Model:

$$P(| Index(X_t^i) - Index(X_t^j) |)$$

Articulatory
Dynamics
(Feature
Substitution)

Underlying U_t^i	Vow	Vow	Vow	Nas	Nas	Occ
Observed X_t^i	Vow	Vow	Nas	Nas	Nas	Nas

Substitution Model:

$$P(X_t^i = x | U_t^i = y)$$



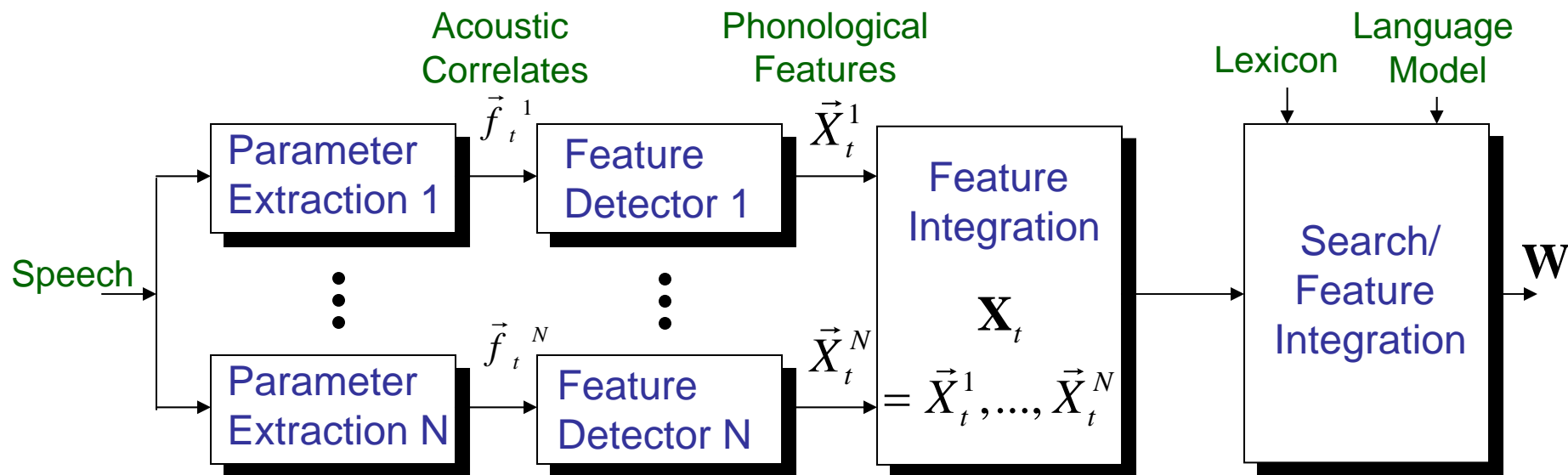
OUTLINE

- Motivating Articulatory-Based Models for ASR
- Sounds to Words – Problems with Pronunciation Dictionaries
- Phonological Distinctive Features (PDF) for ASR
- Integrating PDF's in ASR

Integrating Phonological Distinctive Features in ASR

- Incorporating PDFs in HMM based ASR
- Disambiguating ASR lattice hypotheses through PDF re-scoring
- Articulatory models of vocal tract dynamics
- Acoustic Landmarks

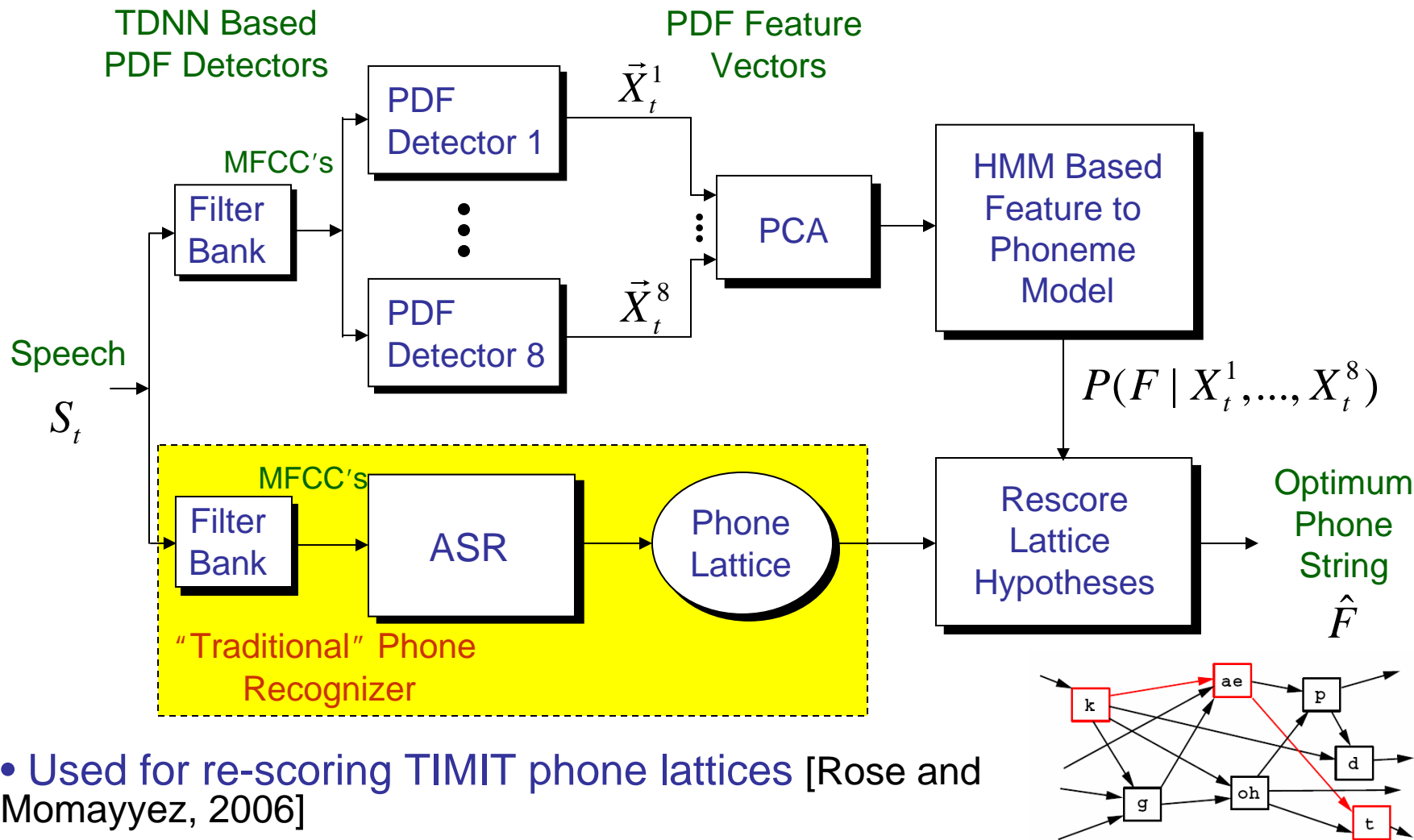
Incorporating PDFs in HMM-Based ASR



- **PDF Integration / Synchronization** [Kirchhoff et al, 2000] [Stuker et al, 2003]
 - **Coupled Features** – Single observation stream: $P(s_k | \mathbf{X})$
 - **Independent Features** – Separate streams of PDFs integrated at the state level:

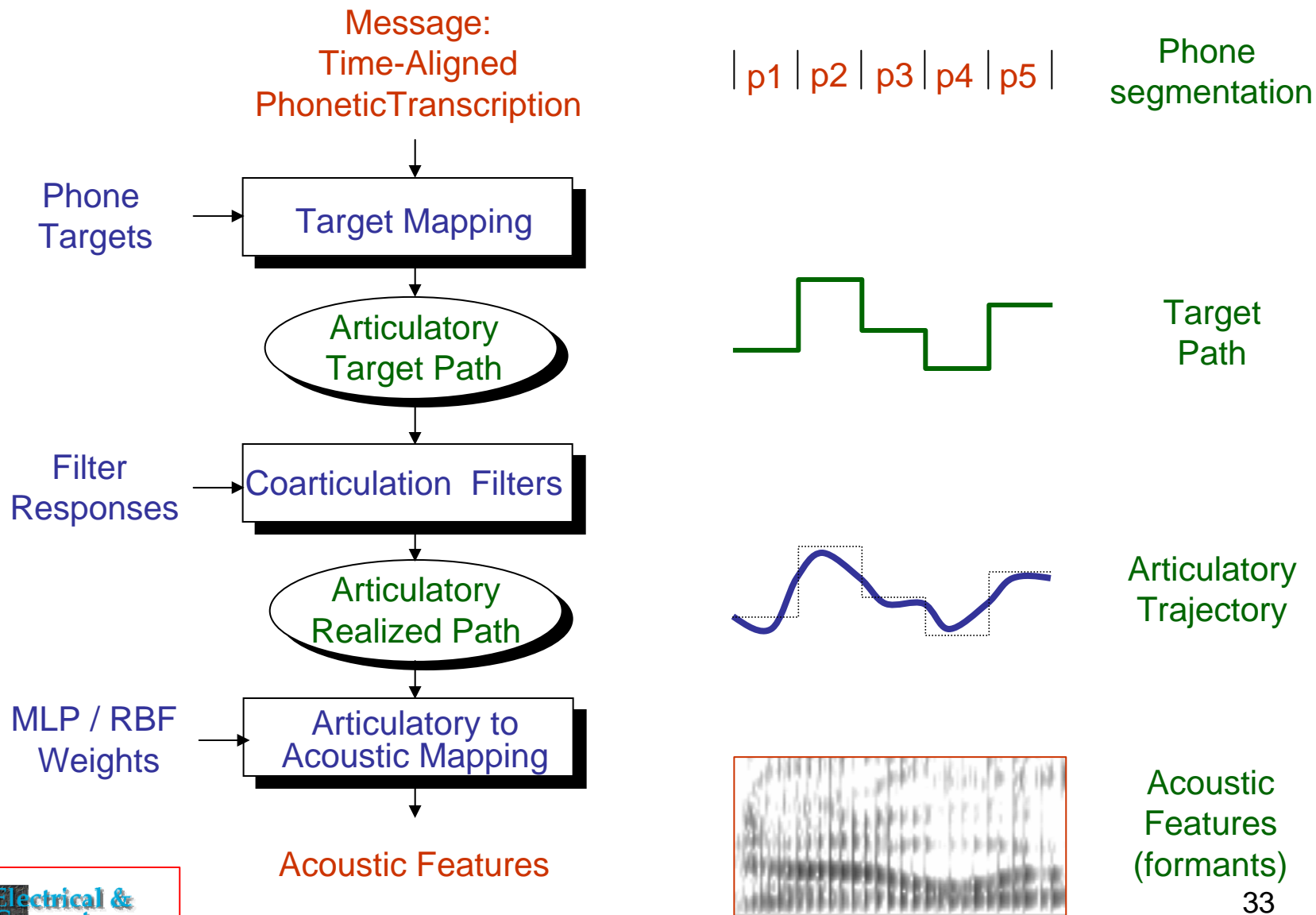
$$\prod_{i=1}^N P(s_t | X_t^i)$$
 - **Unsynchronized Features** – Use of syllable rather than phone-based acoustic units
 - Articulatory synchronization believed to occur at syllable boundaries

Disambiguating ASR Hypotheses by PDF Rescoring



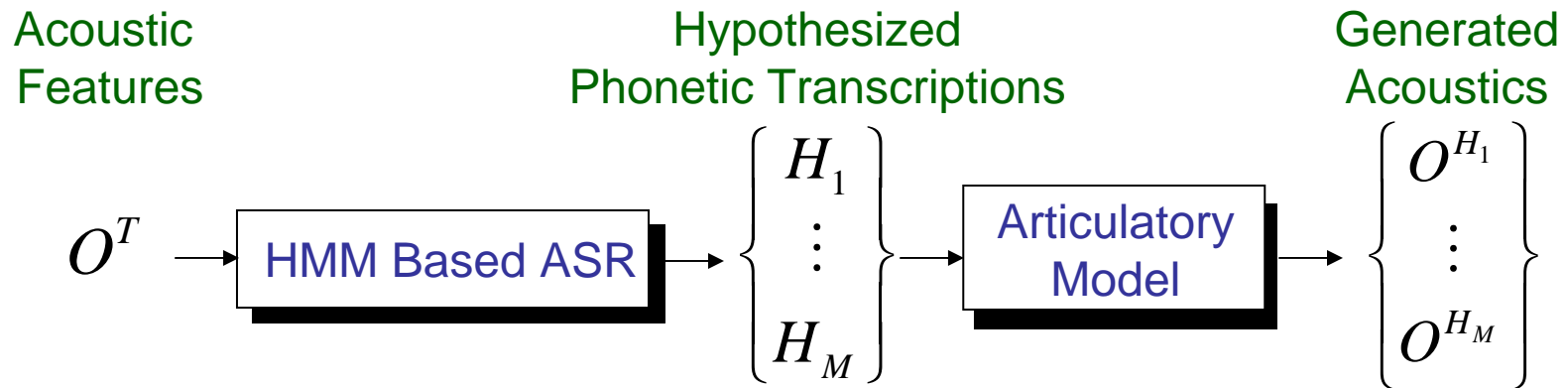
- Used for re-scoring TIMIT phone lattices [Rose and Momayyez, 2006]
- PAC increase from 70% to 73% with PDF re-scoring

Articulatory Models of Vocal Tract Dynamics



Articulatory Models of Vocal Tract Dynamics

- Multi-dimensional articulatory models obtained as the Cartesian product models for each articulator dimension result in enormous computational complexity during search
- Use traditional ASR to generate hypothesized phonetic transcriptions:

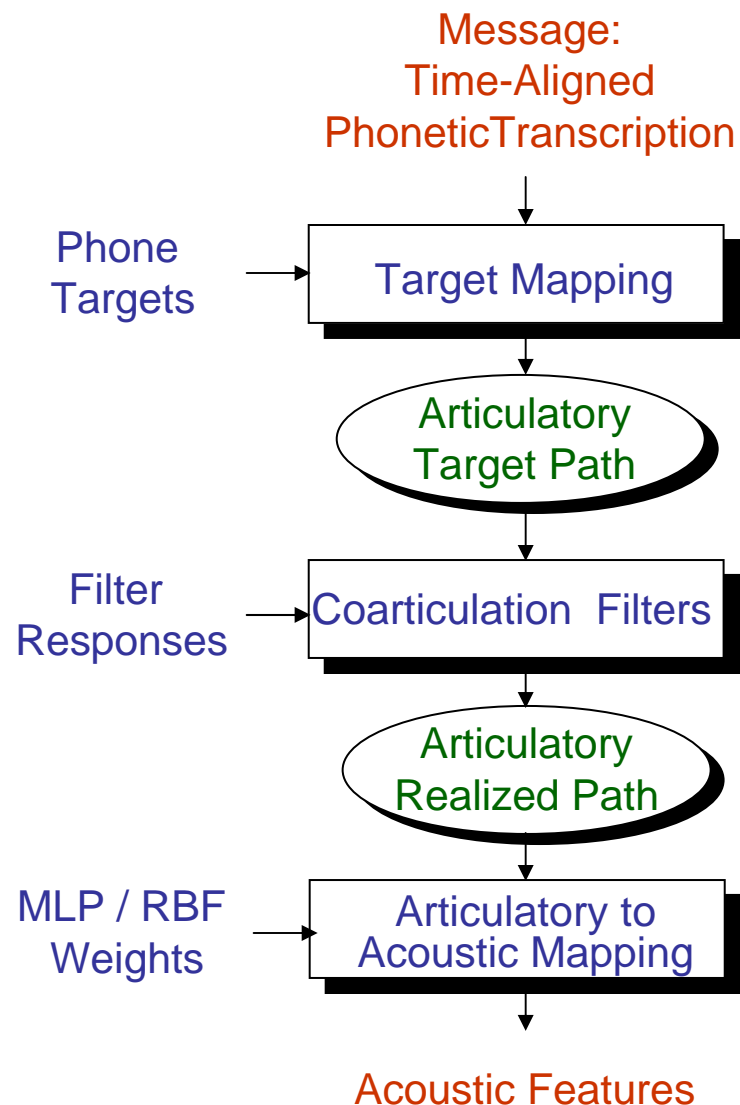


- Choose the phonetic transcription that is the most “plausible” according to the articulatory model

$$\hat{H} = \arg \max_H D(O^H, O^T)$$

Articulatory Models of Vocal Tract Dynamics

- **Coarticulation**
 - Empirically designed FIR filters [Bakis]
 - Deterministic hidden dynamic model (HDM) [Bridle et al, 1999]
 - Vocal tract resonance dynamics (VTR) [Deng et al, 1998]
- **Articulatory-to-Acoustic Mapping**
 - Radial basis functions [Bakis]
 - MLPs [Bridle et al, 1999]



Resources

- U.S. Government Sponsored JHU Workshops
 - 1997 – Doddington et al – Syllable-based speech processing
 - 1998 – Bridle et al – Segmental hidden dynamical models for ASR
 - 2004 – Hasagawa-Johnson et al – Landmark based speech recognition
 - 2006 – Livescu et al – Articulatory feature based speech recognition
- Corpora
 - Phonetically labeled
 - TIMIT
 - ICSI Switchboard transcription project [Greenberg, 2000]
 - Buckeye Corpus (Ohio State),
 - Switchboard (King et al, 2006)
 - Direct Articulatory Measurements
 - Wisconsin microbeam corpus
 - Audio-Visual TIMIT corpus (AVTIMIT) [MIT]

Summary

- Huge Potential for Incorporating Speech Production Models in ASR
 - Pronunciation Variations
 - Acoustic Invariance
 - Efficient Adaptation
- Still Debating:
 - Advantages of Phonological Distinctive Features for ASR
 - Our Ability to Pose or Learn Articulatory Based Models
- How Should we be Attacking the Problem?
 - Resources (Annotated Data)
 - Interim Milestones