AERFAI Summer School

Speech Production Models in ASR

Richard Rose
June, 2008

McGill University
Dept. of Electrical and Computer Engineering

OUTLINE

1. Speech Production Models

- Motivating Articulatory Based Models for ASR
- Review of Speech Production and Distinctive Features
- Sounds to Words – Problems with Pronunciation Dictionaries
- The Role of Speech Production Models in Speech Perception

2. Exploiting Speech Production Models in ASR

- Statistical methods for phonological distinctive feature detection
- Incorporating distinctive feature knowledge in ASR model structure
- Development of models of articulatory dynamics
- Integrating distinctive features in traditional ASR systems

3. Resources for Research

- Articulatory measurements and clinical tools
- Speech corpora
- Projects dedicated to speech production models in ASR

1. Speech Production Models

- Motivating Articulatory Based Models for ASR
- Review of Speech Production and Distinctive Features
- Sounds to Words – Problems with Ponemic Pronunciation Dictionaries
- The Role of Speech Production Models in Speech Perception

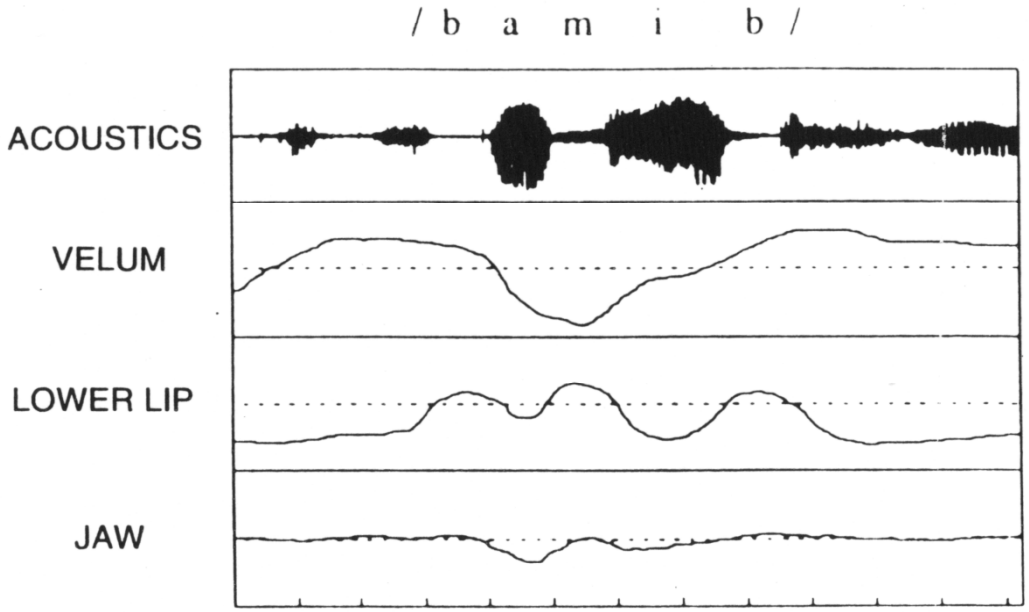
Motivating Articulatory-Based Models for ASR

- 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

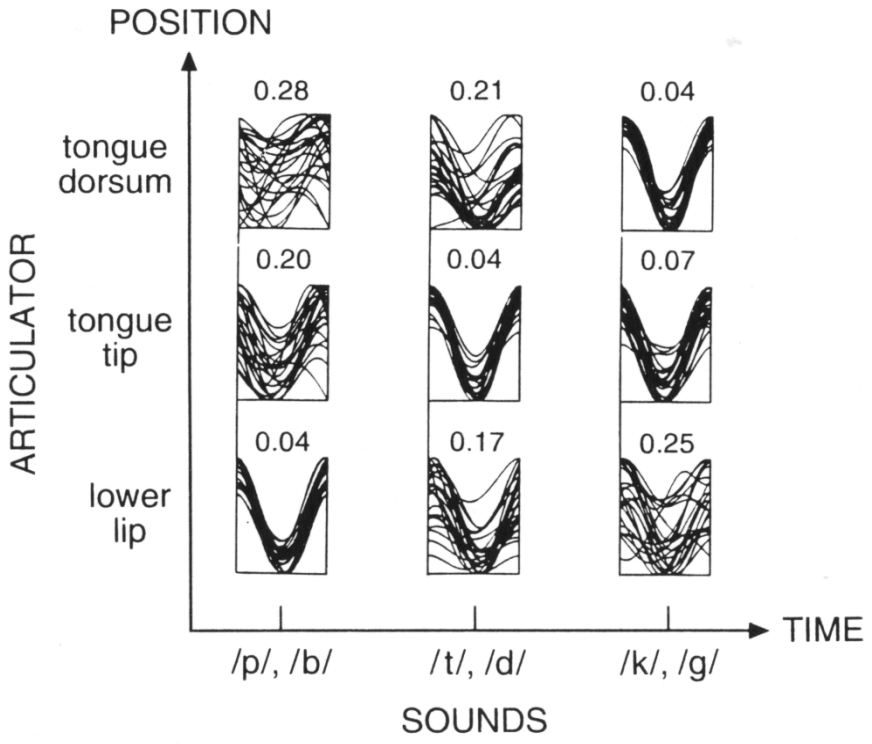
Acoustic waveform and measured articulatory trajectories for utterance of "It's a /bamib/ sid" (Krakow, 1987)



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

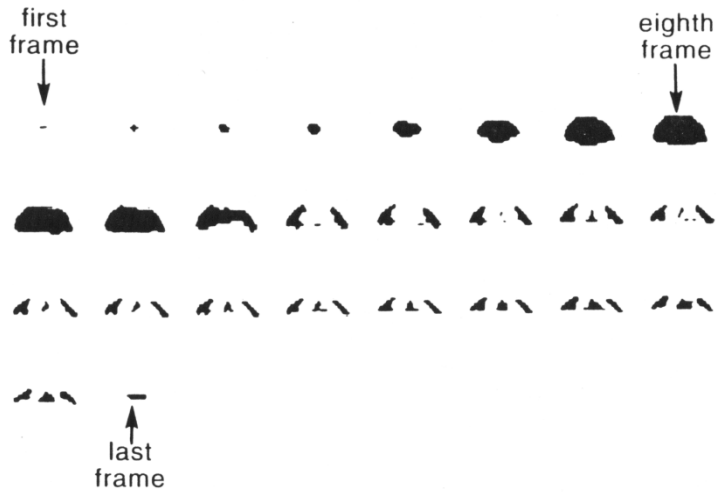


Papcun et al, 1992

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][Wrench, 2002]

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

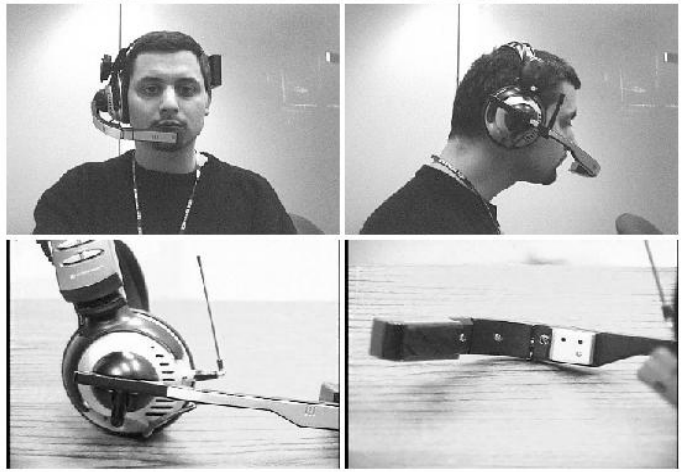


Sequence of binary lip/tongue images for word “one” (Petajan)

Placement of EMA coils (Zlokarnik)

“Partial” Direct Measurements - Visual Information

- Partial direct articulatory measurements fused with acoustic information in audio-visual ASR [Potamianos et al, 2004]



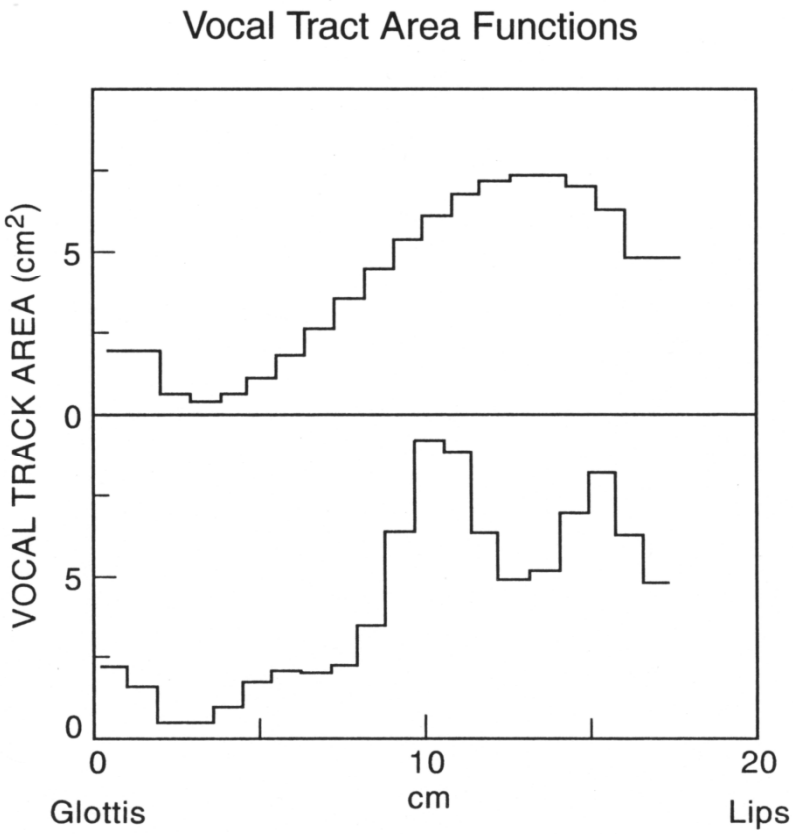
IBM Audio-Visual Headset
[Potamianos et al, 2004]

Motivating Articulatory-Based Models for ASR

- Challenges for Incorporating Articulatory Models
  - One-to-many acoustic to vocal tract area mapping
  - Non-linear relationship between production, acoustics, and perception
  - 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



Frequencies, Bandwidths

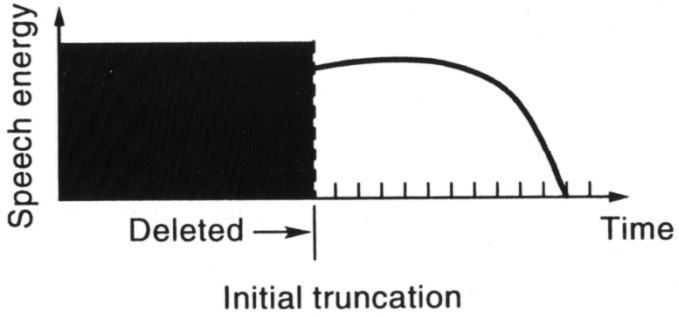
F	653	1188	2177	3321	HZ
B	54	132	119	188	HZ

F	653	1188	2177	2783	HZ
B	55	139	169	143	HZ

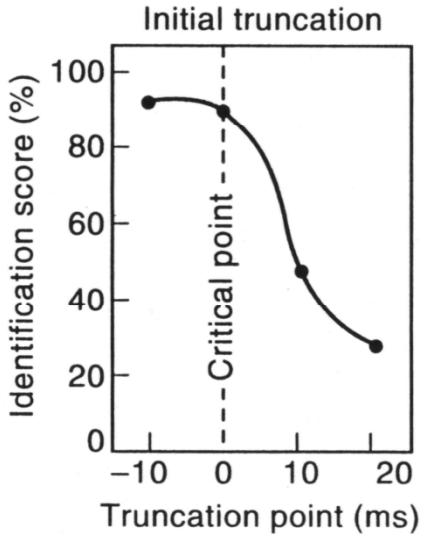
Different Vocal Tract Shapes for Producing Vowel /a/
(Sondhi attributed to Atal)

Acoustic Coding of Articulatory Information

- Perceptually salient information necessary for making phonemic distinctions can be contained in fast-varying, short duration acoustic intervals [Furui, 1986]
- 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

[From Furui, 1986]

1. Speech Production Models

- Motivating Articulatory Based Models for ASR
- Review of Speech Production and Distinctive Features
- Sounds to Words – Problems with Pronunciation Dictionaries
- The Role of Speech Production Models in Speech Perception

A Brief Review of Distinctive Features

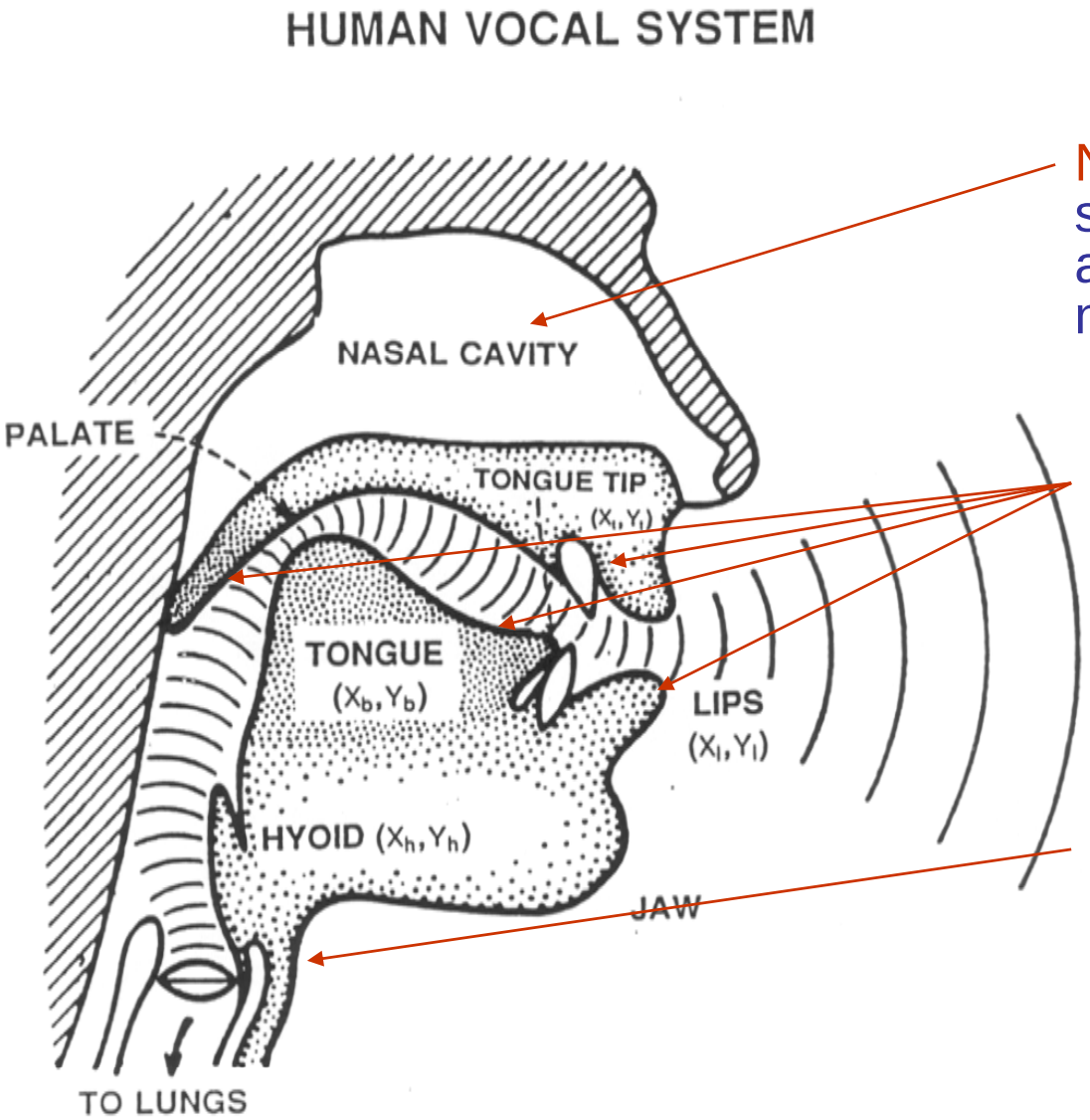
- We need a way to describe the sounds of speech in any language in terms of the underlying speech production system

- **Distinctive Features** – Serve to distinguish one phoneme from another by describing:

  1. **The Manner** in which the sound is produced
     - Voiced, Unvoiced, Vocalic, Consonantal, Nasal

  2. **The Place** where the sound is articulated
     - Labial, Dental, Alveolar, Palatal, Velar

Speech Production – Distinctive Features



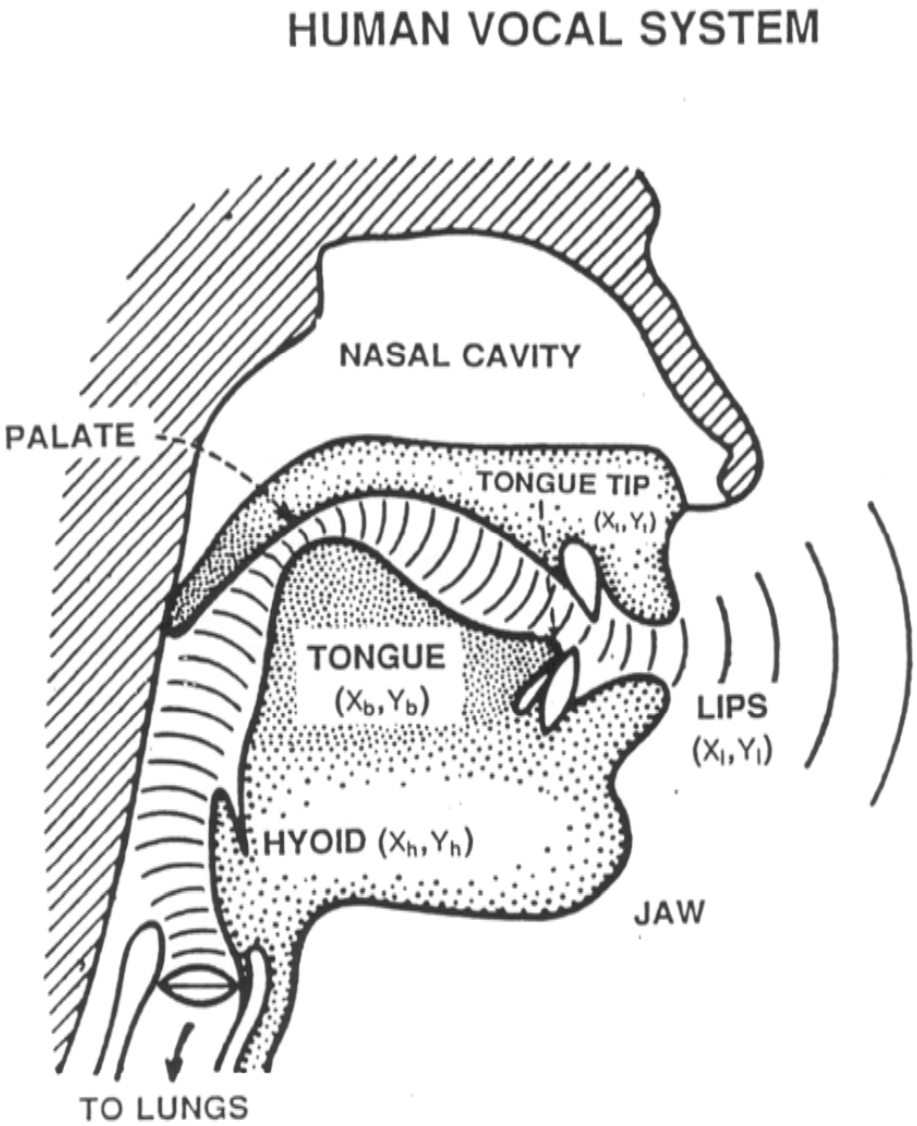
Nasal Cavity: Produces sound when velum is open and air flows through the nose

Articulators: Tongue, Lips, Teeth, Velum, and Hard Palate

Vocal Folds / Glottis: Vibrates in response to air pressure originating in the lungs

[from Rabiner and Juang, 1993]

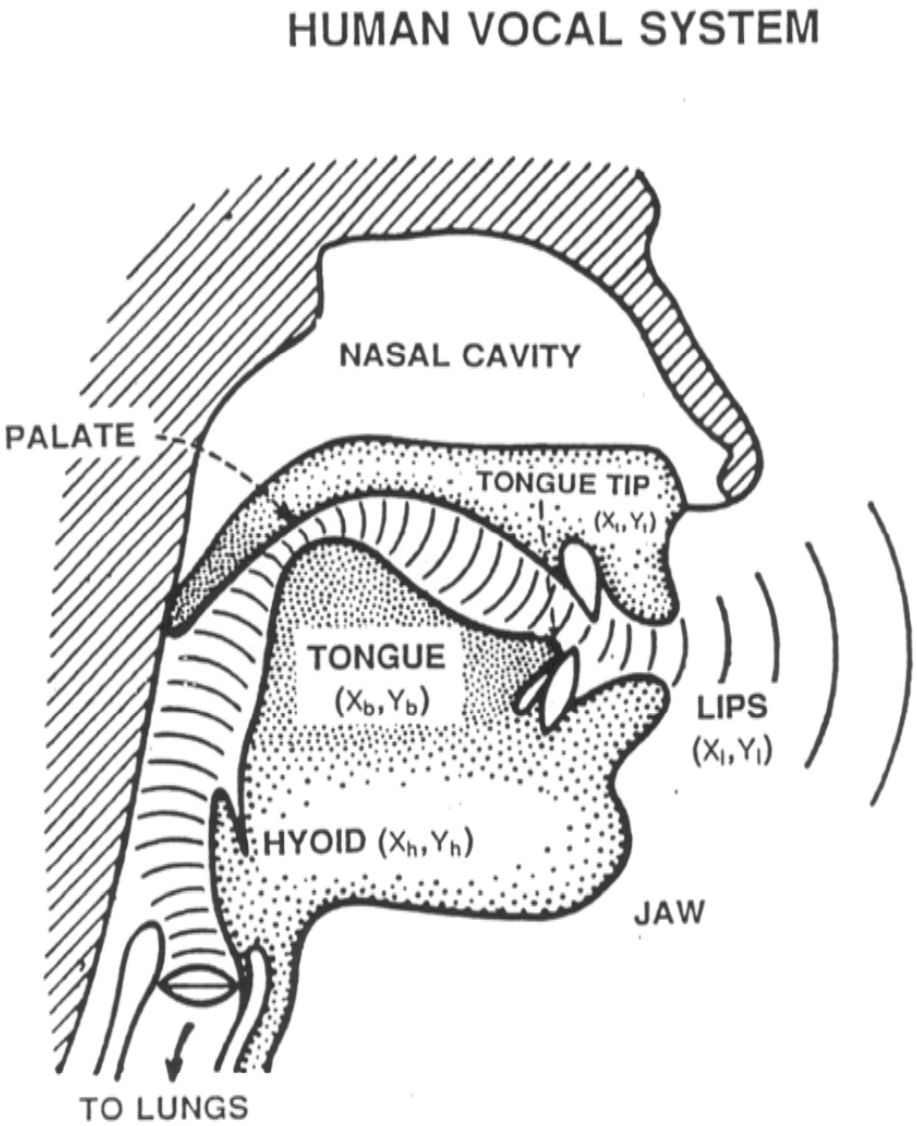
Speech Production – Distinctive Features



[from Rabiner and Juang, 1993]

- Manner of Production
  - Voiced: Glottis closed with glottal folds vibrating
  - Unvoiced: Glottis open
  - Sonorant: No major constriction in the vocal tract and vocal cords set for voicing
  - Consonantal: Major constriction in vocal tract
  - Nasal: Air travels through the nasal cavity

Speech Production – Distinctive Features



[from Rabiner and Juang, 1993]

- Place of Articulation
  - Bilabial - Lips - /P/,/B/,/M/
  - Dental - Tongue Tip and Front Teeth- /TH/,/DH/
  - Alveolar - Alveolar Ridge and Tip of Tongue - /T/,/D/,N/,/S/,/Z/,/L/
  - Palatal - Hard Palate and Tip of Tongue - /Y/,/ZH/
  - Velar - Soft Palate (Velum) and Back of Tongue - /K/,/G/,/NG/

Classes of Sounds: Vowels

- Distinctive Features that are common to all vowels:

  +Voiced, +Sonorant, -Consonantal

- Vowels are distinguished by Distinctive Features:
  - Tongue Position: Front, Mid, Back
  - Jaw Position: High, Mid, Low
  - Lip Rounding: Rounded, Not-Rounded
  - Tense / Lax: Widening of the cross-sectional area of the pharynx by moving the tongue root forward

Vowels of English

English vowels include monothongs, dipthongs, and reduced vowels

tense / lax pairs

JAW POSITION \ TONGUE BODY	Front	Mid	Back
High	/IY/ peat /IH/ pit	/ER/ pert	/UW/ boot /UH/ foot
Mid	/EY/ eight /EH/ pet	/AH/ putt	/OH/ open /AO/ all
Low	/AE/ pat		/AA/ father

REDUCED VOWELS: /AX/ about /IX/ roses /AXR/ butter

DIPTHONGS: /AY/ bite /OY/ Boyd /AW/ bout

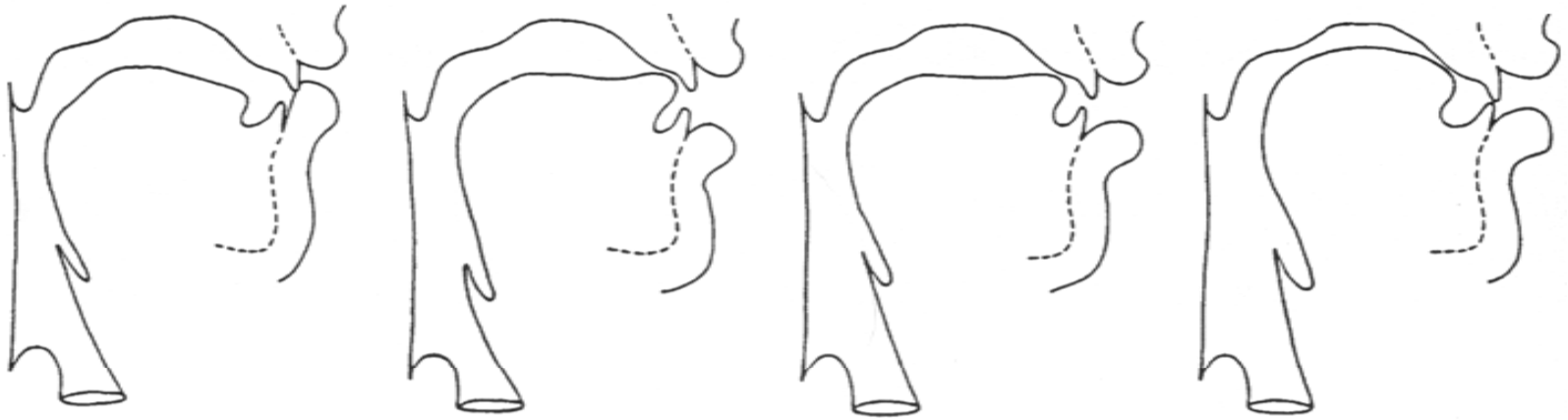
Classes of Sounds: Consonants

- Distinctive Features that are common to all consonants:

  -Sonorant, +Consonental

- Consonants are distinguished by distinctive features:
  - Place of Articulation
    - Labial, Dental, Aveolar, Palatal, Velar
  - Manner of Articulation
    - Stop: Complete Stoppage of airflow in the Vocal Tract followed by a release
    - Fricative: Noise from constriction in the vocal tract
    - Nasal: Velum open and air flows through nasal cavity

Classes of Sounds: Fricatives



/F/	/TH/	/S/	/SH/
find	the	say	show
Labial	Dental	Alveolar	Palatal-Alveolar

Classes of Sounds: Nasals and Affricatives

- Nasals:
  - Distinctive Feature Common to Nasals is +nasal (velum open)
  - Distinguished by places of articulation
    - /M/ mom – labial
    - /N/ none – alveolar
    - /NG/ sing - velular
- Affricatives:
  - Alveolar-stop palatal-fricative pair
  - Distinguished by voicing
    - /JH/ judge – voiced
    - /CH/ church – unvoiced
- Aspirant:
  - One aspirant in English produced by turbulant exication at the glottis
    - /H/ hat

Classes of Sounds: Semi-Vowels

- Transition Sounds:
  - Liquids: Some obstruction of the airstream of the mouth but not enough to cause frication

    /L/ - lack /R/ - red
  - Glides: Tongue moves rapidly in a gliding fashion either toward or away from neighboring vowel

    /W/ - way /Y/ - you

Example: Distinctive Features used to Define Phonological Rules for Morphologically Related Words

An example: The plural form of English nouns

- Orthographically: Plural is formed by adding “s” or “es”
- Phonemically: Plurals result in adding one of three endings to the word: /S/, /Z/, or /IH/ /Z/
- The actual ending depends on the last phoneme of the word.

Which plural ending would be associated with the following 3 groups of words?

What is the minimum feature set for the phonemes that proceed these plural endings?

1. breeze, fleece, fish, judge, witch

/IH//Z/: +consonental, +strident, -stop, +alveolar

2. mop, lot, puck, leaf, moth

/S/: +consonental -vocalic -voiced

3. tree, tray, bow, bag, mom, bun, bang, ball, bar

/Z/: +voiced

Phonology: From Phonemes to Spoken Language

- Phonology: Mapping from baseform phonemes to acoustic realizations (surface form phonemes)
- Allophones: Predictable phonetic variants of a phoneme
- Phonological Rules: Applied to phoneme strings to produce actual pronunciation of words in sentences
  - Assimilation: Spreading of phonetic features across phonemes
  - Flapping: Change alveolar stop to a “flap” when spoken between vowels
  - Nasalization: Impart nasal feature to vowels preceding nasals
  - Vowel Reduction: Change vowel to /AX/ when unstressed

Representations	Flapping Rule (CITY)	Vowel Reduction Rule (PHONOLOGY)
Phonemic	**/C/ /IH/ /T/ /IY/**	**/F/ /OH/ /N/ /AA/ /L/ /AX/ /J/ /IY/**
Phonetic	**/C/ /IH/ /D/ /IY/**	**/F/ /AX/ /N/ /AA/ /L/ /AX/ /J/ /IY/**

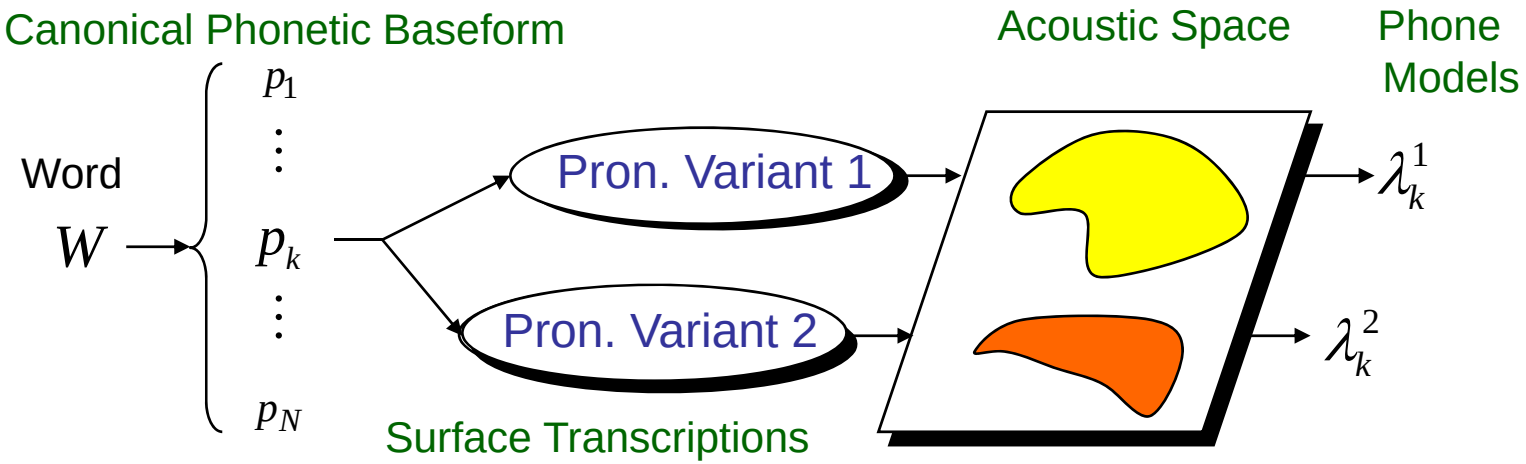
1. Speech Production Models

- Motivating Articulatory Based Models for ASR
- Review of Speech Production and Distinctive Features
- Sounds to Words – Problems with Phonemic Pronunciation Dictionaries
- The Role of Speech Production Models in Speech Perception

Sounds to Words – Problems with Dictionaries

Mismatch: Canonical baseforms vs. Surface Form Variant

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



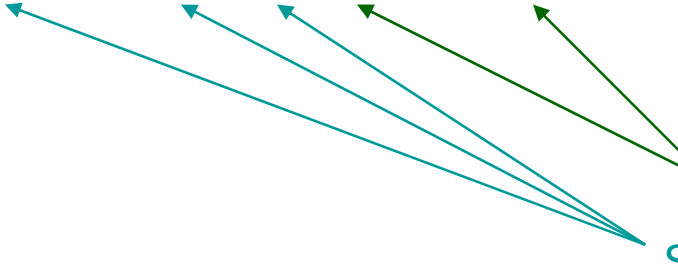
- The challenge is to predict pronunciation variants during recognition:

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

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



Deletion

Pronunciation variants

Surface acoustic information
(cl: closure / r: release)

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	-

  - Compare phone accuracy (PAC) and word accuracy (WAC) using S-F and B-F HMM models [Rose et al, 2008]

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) [Rose et al, 2008]

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

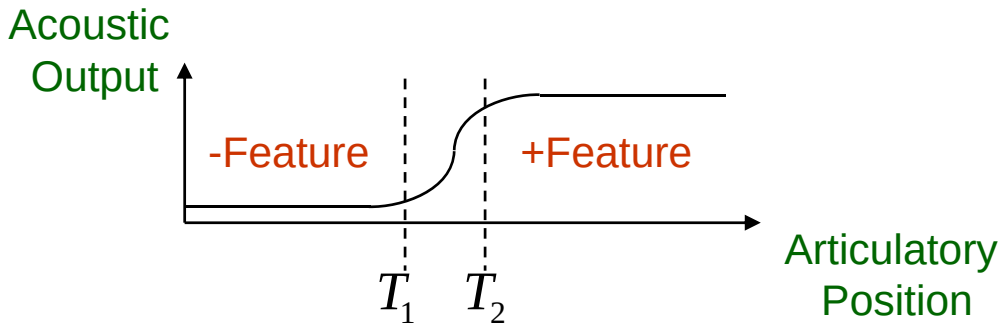
    … But this does not result in better ASR word accuracy

1. Speech Production Models

- Motivating Articulatory Based Models for ASR
- Review of Speech Production and Distinctive Features
- Sounds to Words – Problems with Pronunciation Dictionaries
- The Role of Speech Production Models in Speech Perception

Connection Between Distinctive Features and Speech Perception

- Quantal Theory of Speech Perception: Every distinctive feature in every language represents a nonlinear discontinuity in the relationship between articulatory position and acoustic output [Stevens, 1989]



- Example: Opening velum by $T_2 - T_1 = 2$ millimeters while uttering the phoneme /d/ causes increase in acoustic output energy of 20 – 30 dB
  - /d/ becomes /n/ and [–sonorant] becomes [+sonorant]
- Similar non-linear discontinuities exist in the relationship between acoustics and perceptual space

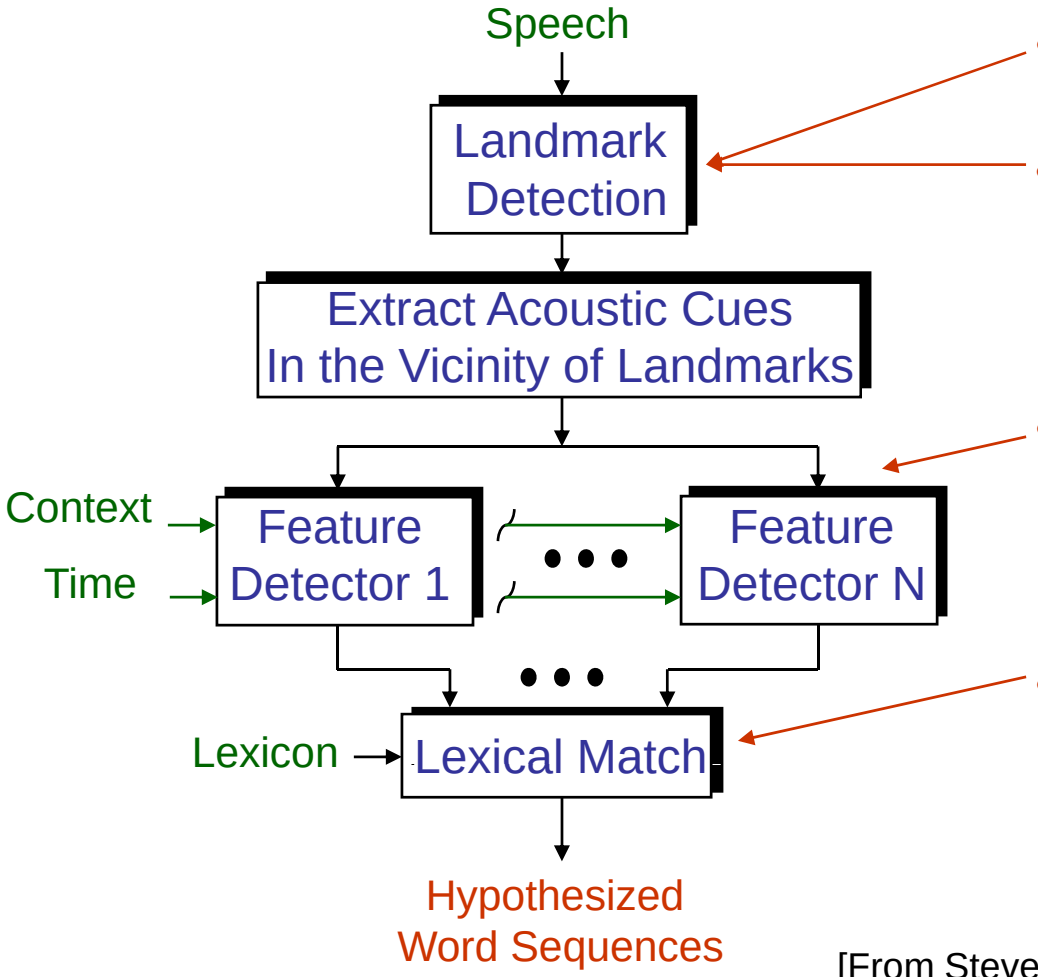
A Model of Human Speech Perception - Distinctive Features and Acoustic Landmarks

- Model speech perception process using a discrete lexical representation [Stevens, 2002]:
  - Words are a sequence of discrete segments
  - Segments are a discrete set of distinctive features

- Landmarks: Provide evidence for broad classes of consonant or vowel segments
- Articulatory Features: Associated with articulation event and acoustic pattern occurring near landmarks

Landmark / Feature Based Model of Human Perception

Model of Lexical Access in Human Speech Perception [Stevens,2002]

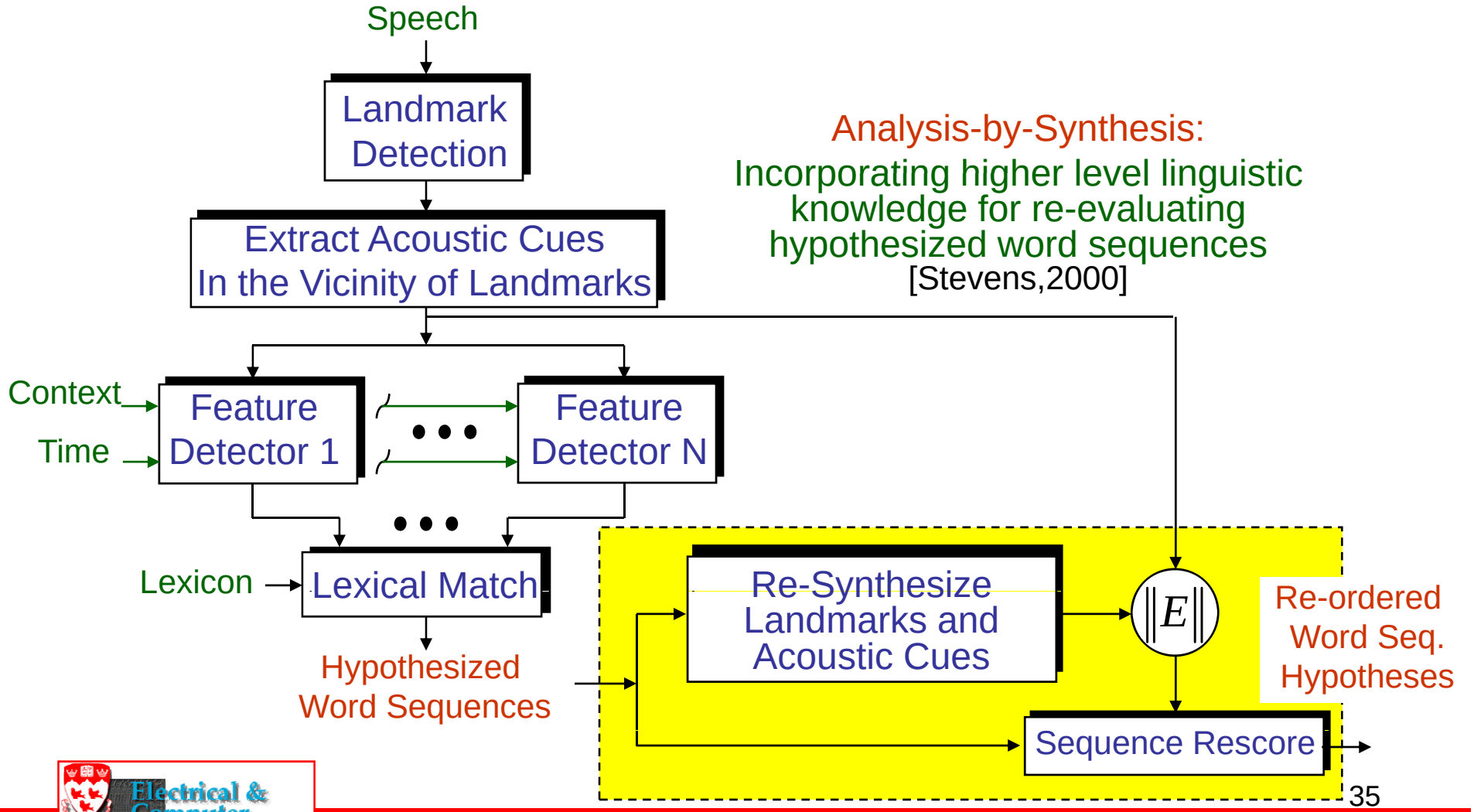


- Vowel Landmarks - Peaks in first formant
- Consonant Landmarks - Acoustic discontinuities
- Articulator Bound Features – Extracted from Acoustic Cues within tens of milliseconds of landmarks
- Words in Lexicon – Formed from segments made up of “bundles” of features

[From Stevens, 2002]

Landmark / Feature Based Model of Human Perception

Model of Lexical Access in Human Speech Perception [Stevens,2002]



[From Stevens, 2002]

2. Exploiting Speech Production Models in ASR

- Statistical methods for phonological distinctive feature (PDF) detection
- Incorporating distinctive feature knowledge in ASR model structure
- Articulatory models of vocal tract dynamics
- Integrating distinctive features in traditional ASR systems

Statistical methods for phonological distinctive feature (PDF) detection

- The definition of PDFs for ASR
- Obtaining acoustic parameters from surface acoustic measures
- Issues for incorporating PDFs and training PDF Detectors
- Statistical methods for PDF detection

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
- 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

Phonological Distinctive Features (PDFs) for ASR

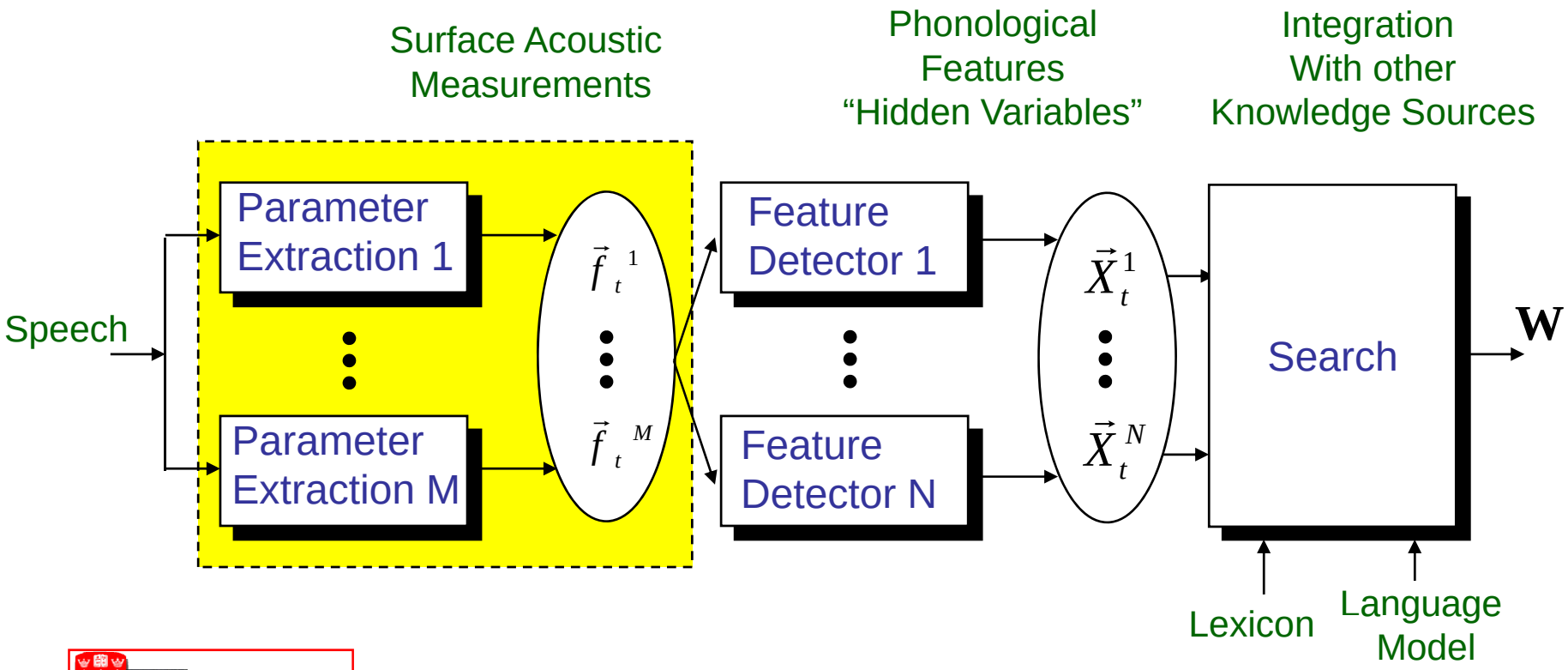
- Example of multi-valued definition of PDFs [King et al, 2000]

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]
  - Government Phonology [Haegeman, 1994][Ahern, 1999]
  - Articulatory Features [Deng and Sun, 1999] [Bridle et al, 1998]

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

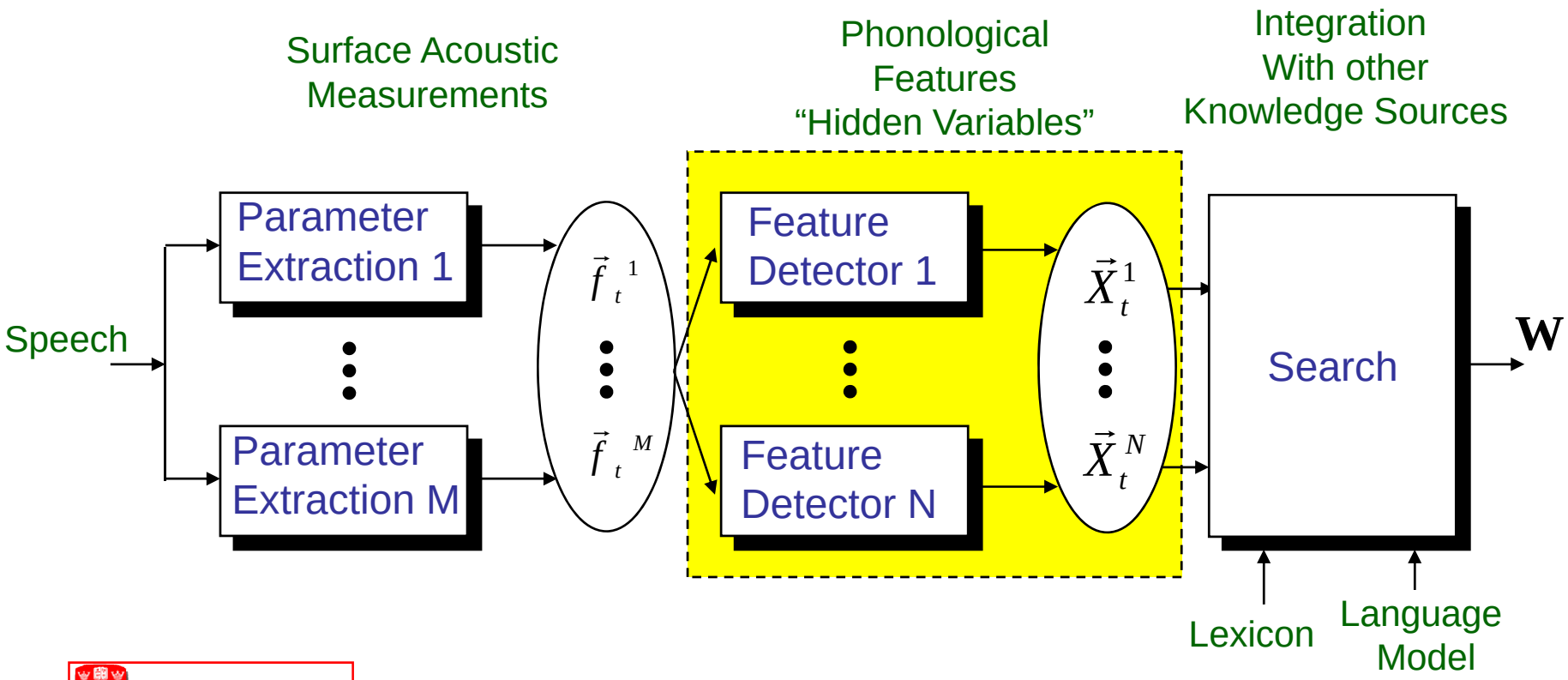
- Define acoustic correlates for a feature
- Determine acoustic parameters that characterize acoustic correlates
  - Example: acoustic parameters for stop consonants [Epsy-Wilson]

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

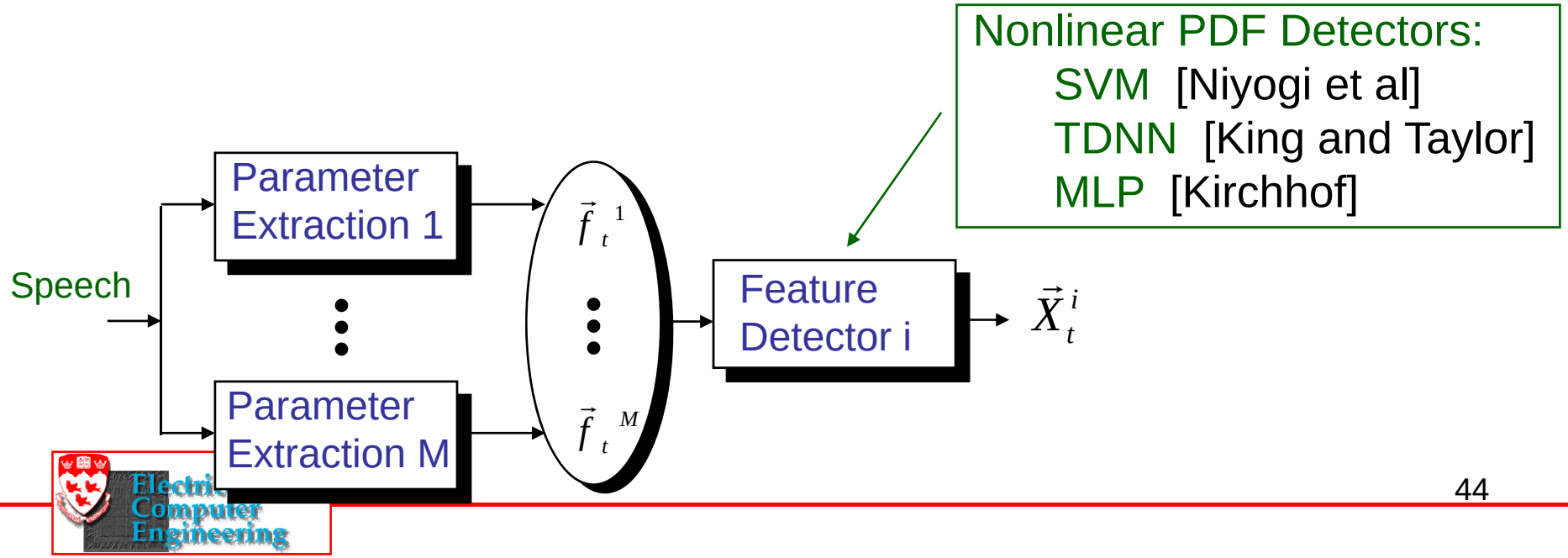


Issues for Training Statistical PDF Detectors

- Supervised Training – Defining "True" Feature Labels in Training
  - Mapping from phone to feature transcriptions [King et al , 2000]
    - Actual feature values may differ from canonical values
  - Using direct physical measurements [Wrench et al, 2000]
    - Difficult to convert physical measurements to feature values
  - Manual labeling of distinctive features [Livescu et al, 2007]
    - Defining labeling methodology, Time consuming (~1000 times RT)
  - Embedded Training – Allow feature boundaries to vary [Frankel et al, 2007]
    - Provides re-alignment of features, but no measure of quality

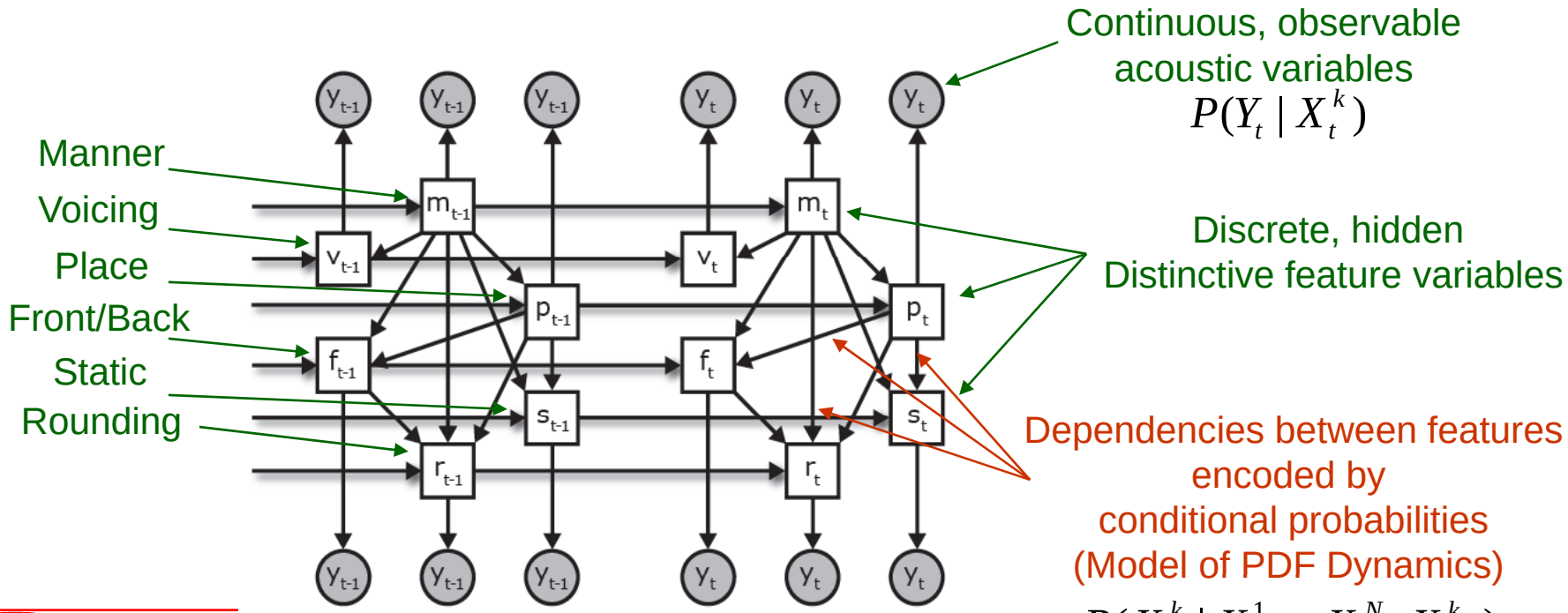
Detecting PDFs From Surface Acoustic Parameters

- 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)



Detecting PDFs From Surface Acoustics – Dynamic Bayesian Networks

- Modeling Asynchrony Among Distinctive Features
  - Models of Vocal Tract Dynamics [Bridle et al, 1999][Deng et al, 1998]
  - Dynamic Bayes networks (DBN) [Frankel et al, 2007][Livescu et al, 2004]



[From Frankel et al, 2007]

Detecting PDFs Using Dynamic Bayesian Networks

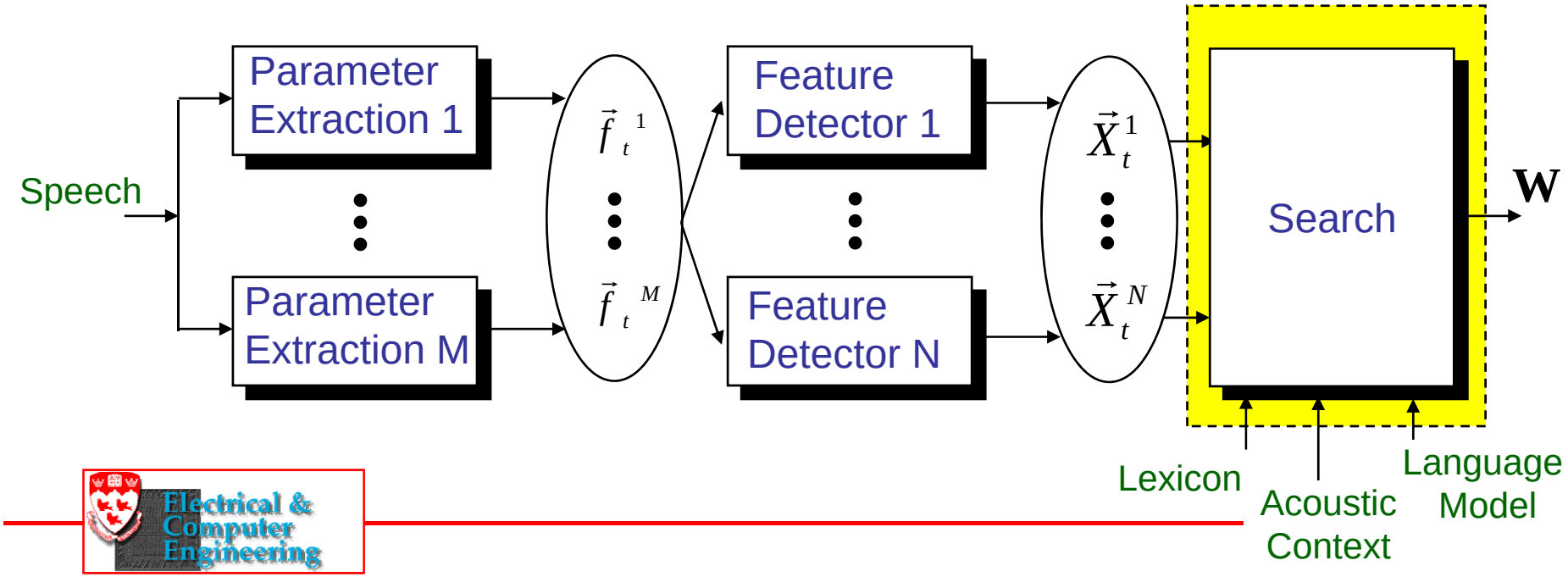
- Modeling Acoustic Observations $P(Y_t \mid X_t^k)$: Gaussian mixtures or artificial neural networks
- Modeling PDF State Process $P(X_t^k \mid X_t^1,..,X_t^N, X_{t-1}^1,..,X_{t-1}^N)$: Hierarchical conditional probability tables – Allows for asynchrony among feature values
- Embedded Training:
  - Initial training performed using phone alignments converted to feature values
  - Generate new PDF alignments and retrain with re-aligned transcriptions
- Effects on Phone Recognition Accuracy:
  - Frankel et al found that embedded training had very little effect on phone accuracy [Frankel, 2007]
  - Observed feature asynchrony was representative of speech production

2. Exploiting Speech Production Models in ASR

- Statistical methods for phonological distinctive feature (PDF) detection
- Incorporating distinctive feature knowledge in ASR model structure
- Development of models of articulatory dynamics
- Integrating distinctive features in traditional ASR systems

ASR Model Structure Based on PDFs

- A Case for Model Structure Based on PDFs
  - HMM State Space: Model topology defined by feature spreading
  - Pronunciation: Feature based description of pronunciation variation
  - A Complete Model: Implementation of landmark based / distinctive feature approach to ASR

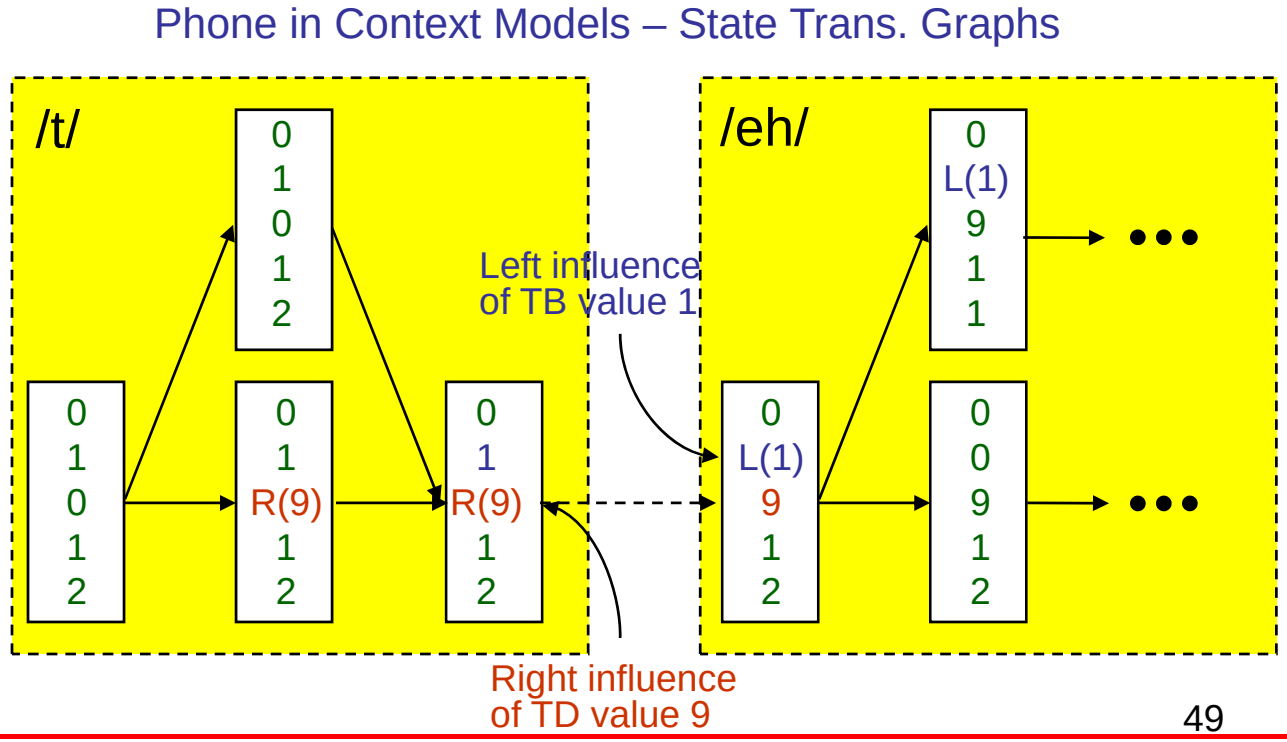


Modeling Structure Based on PDF's

- 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

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

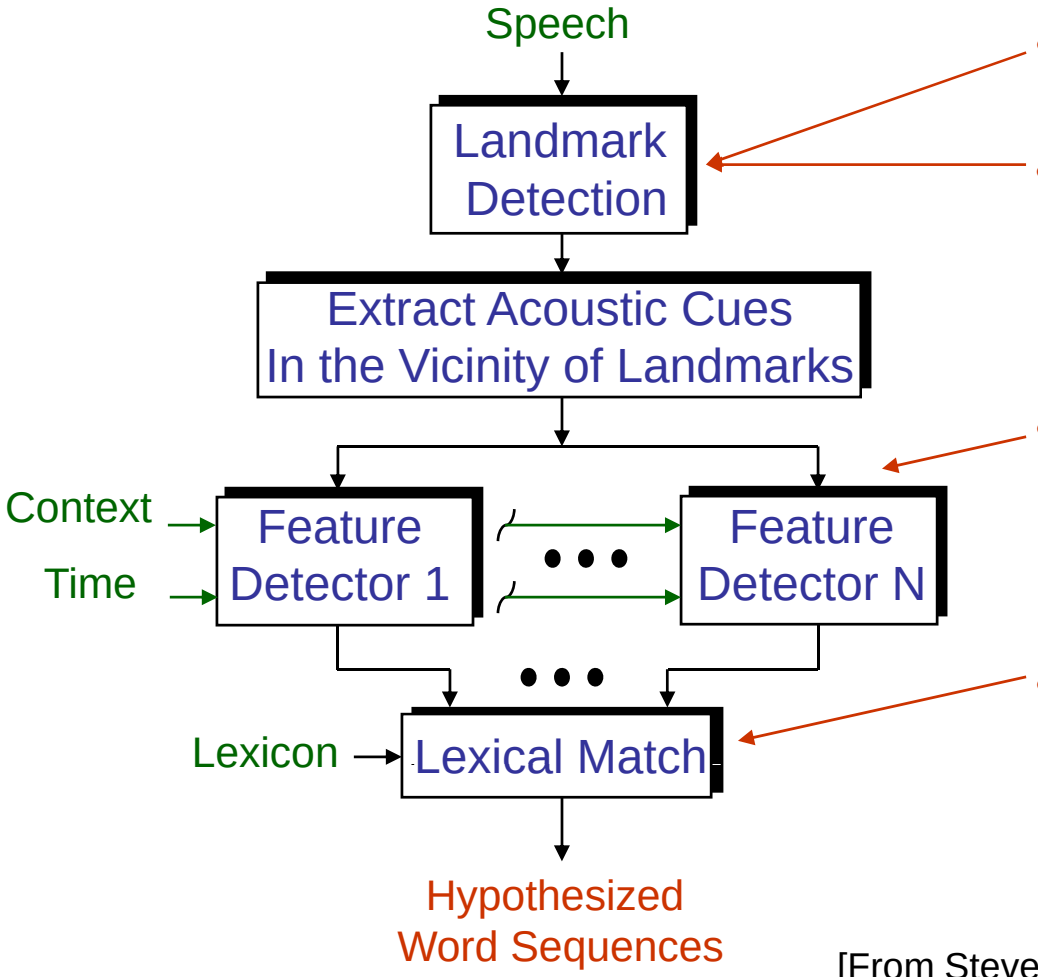
Substitution Model:

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

Feature Frames (t)

Landmark / Feature Based Model of Human Perception

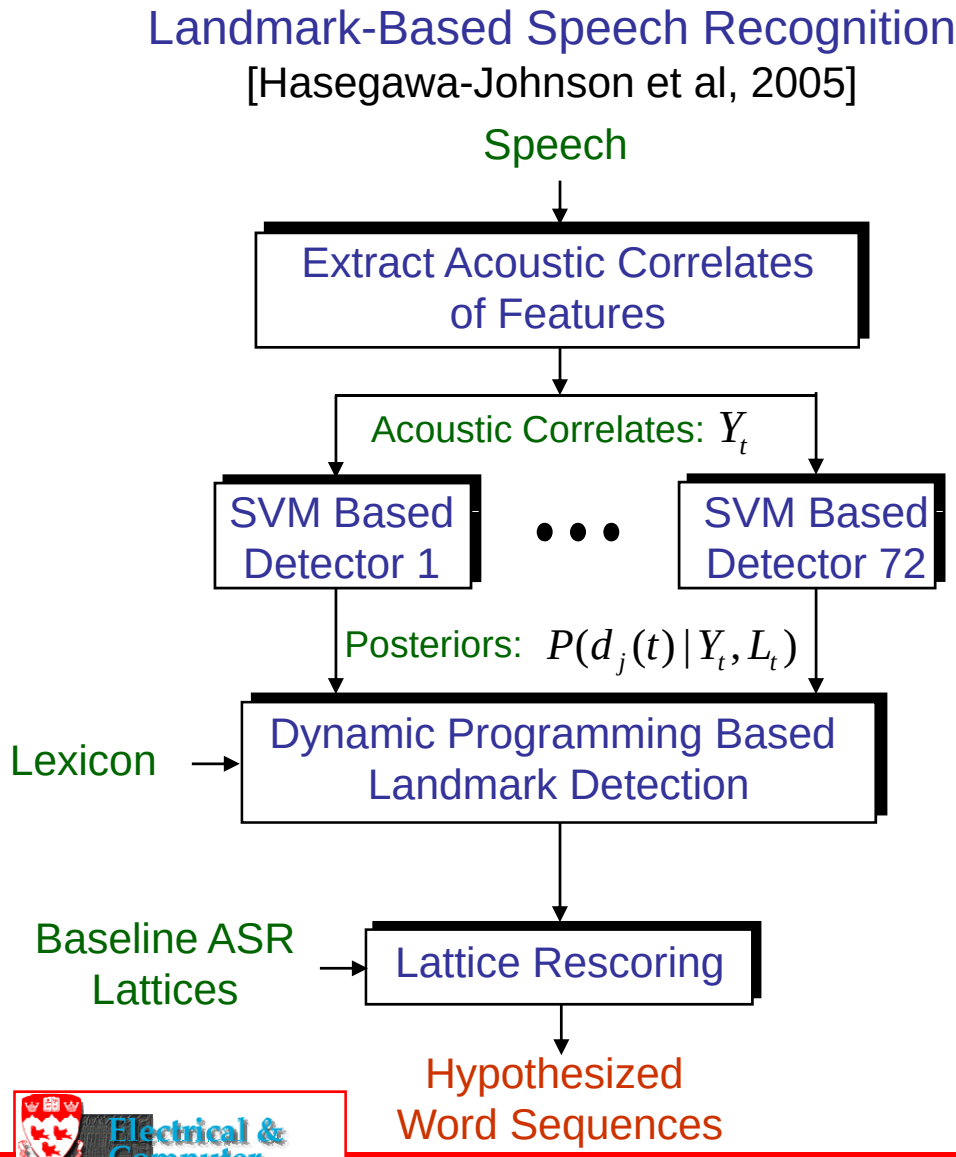
Model of Lexical Access in Human Speech Perception [Stevens,2002]



- Vowel Landmarks - Peaks in first formant
- Consonant Landmarks - Acoustic discontinuities
- Articulator Bound Features – Extracted from Acoustic Cues within tens of milliseconds of landmarks
- Words in Lexicon – Formed from segments made up of "bundles" of features

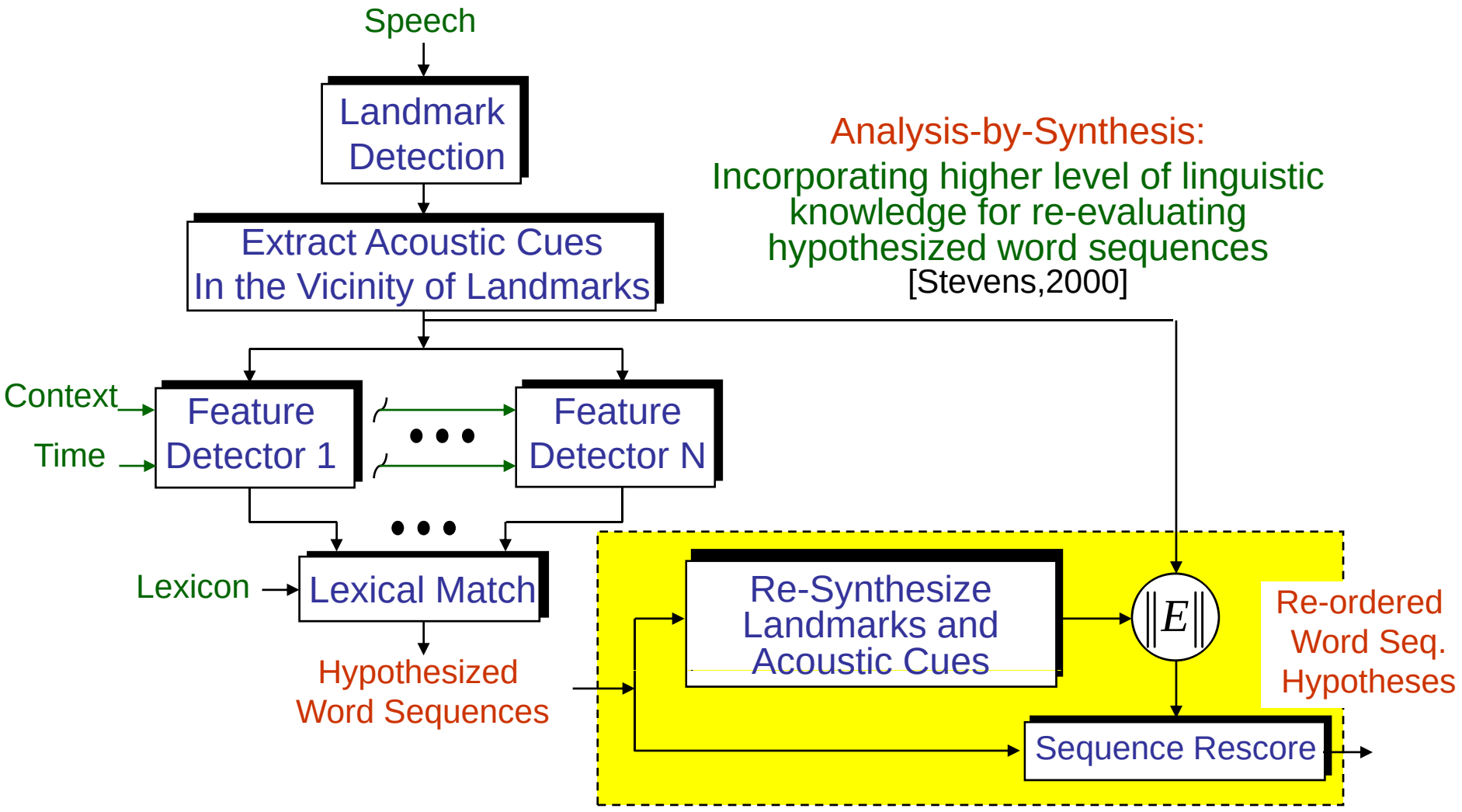
[From Stevens, 2002]

Landmark / Distinctive Feature Based Approach to ASR



- Acoustic Parameters:
  - Energy, spectral tilt, MFCC, formants, …, auditory cortical features [Mesgarni et al, 2004]
- Distinctive Feature Detectors:
  - Support Vector Machines (SVMs)
  - Produce posterior probabilities of ***distinctive feature values*** $d_j(t)$ for ***landmark type*** L_t at time t.
- Landmark Detection
  - Maximizes posterior probability of distinctive feature "bundles" w.r.t. canonical bundles in lexicon
- Lattice Rescoring
  - Rescore Switchboard ASR lattices generated by SRI

Landmark / Feature Based Model of Human Perception

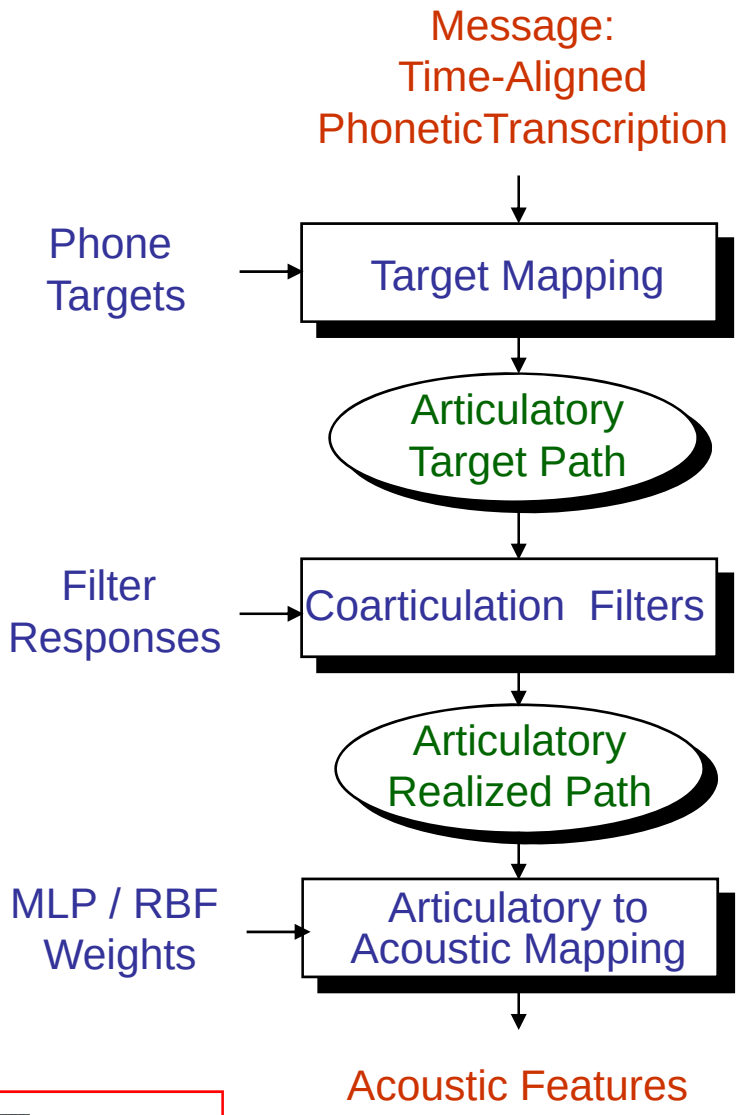


[From Stevens, 2002]

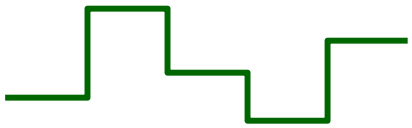
2. Exploiting Speech Production Models in ASR

- Statistical methods for phonological distinctive feature (PDF) detection
- Incorporating distinctive feature knowledge in ASR model structure
- Articulatory models of vocal tract dynamics
- Integrating distinctive features in traditional ASR systems

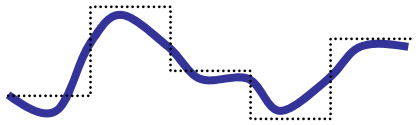
Articulatory Models of Vocal Tract Dynamics



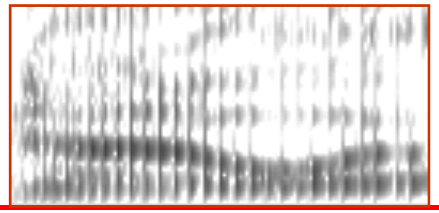
Phone segmentation



Target Path



Articulatory Trajectory



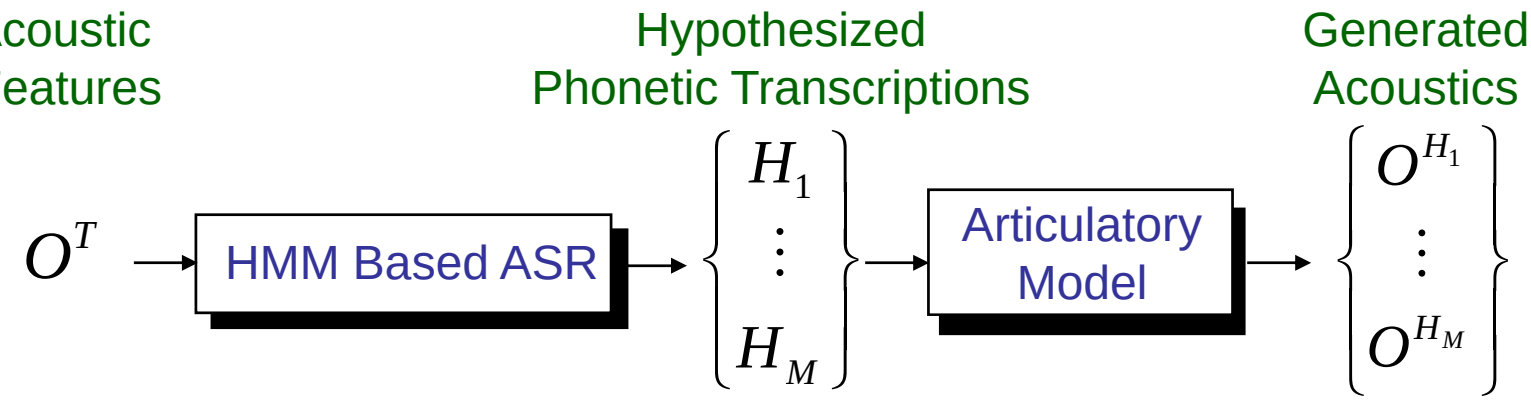
Acoustic Features (formants)

Electrical & Computer Engineering

[Bakis,1993]

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:

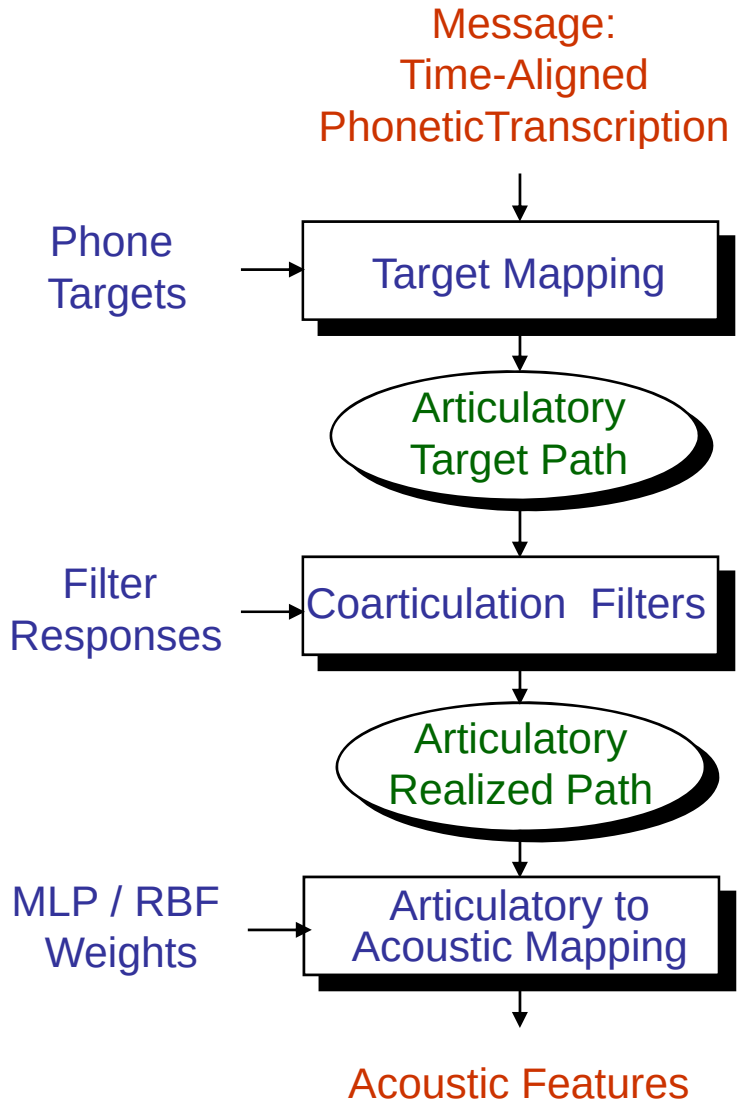


- Choose the phonetic transcription that is the most "plausible" according to the articulatory model

$$\hat{H} = \arg\max_{H} D\left(O^{H}, O^{T}\right)$$

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]



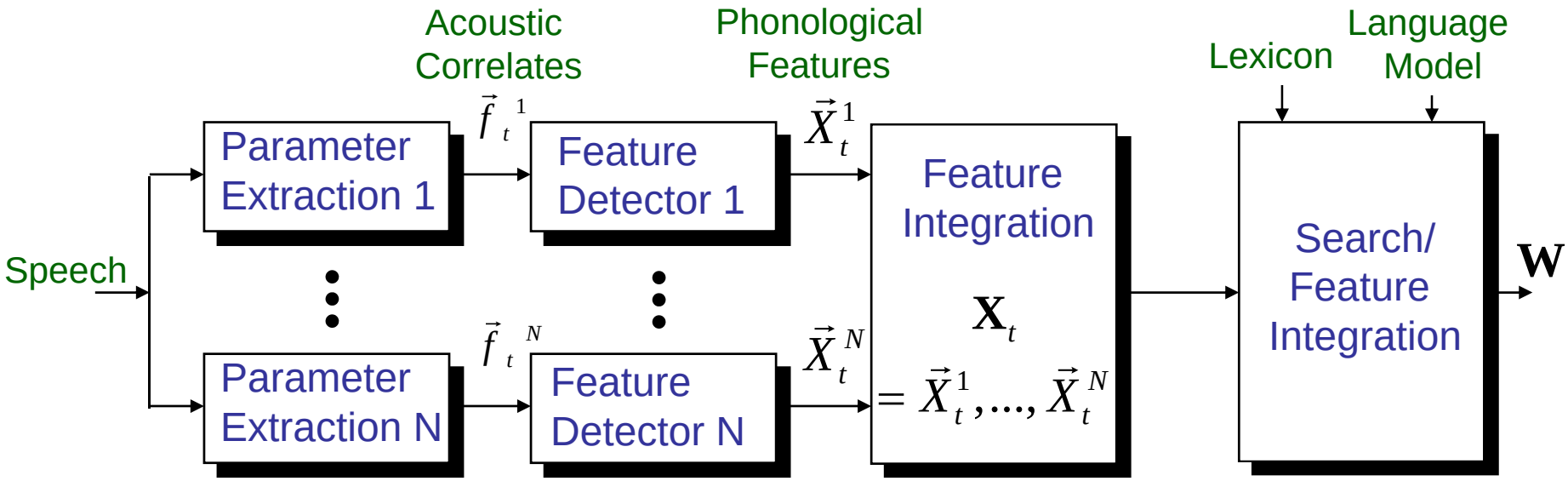
2. Exploiting Speech Production Models in ASR

- Statistical methods for phonological distinctive feature (PDF) detection
- Incorporating distinctive feature knowledge in ASR model structure
- Articulatory models of vocal tract dynamics
- Integrating distinctive features in traditional ASR systems

Integrating Speech Production Models in Traditional ASR Systems

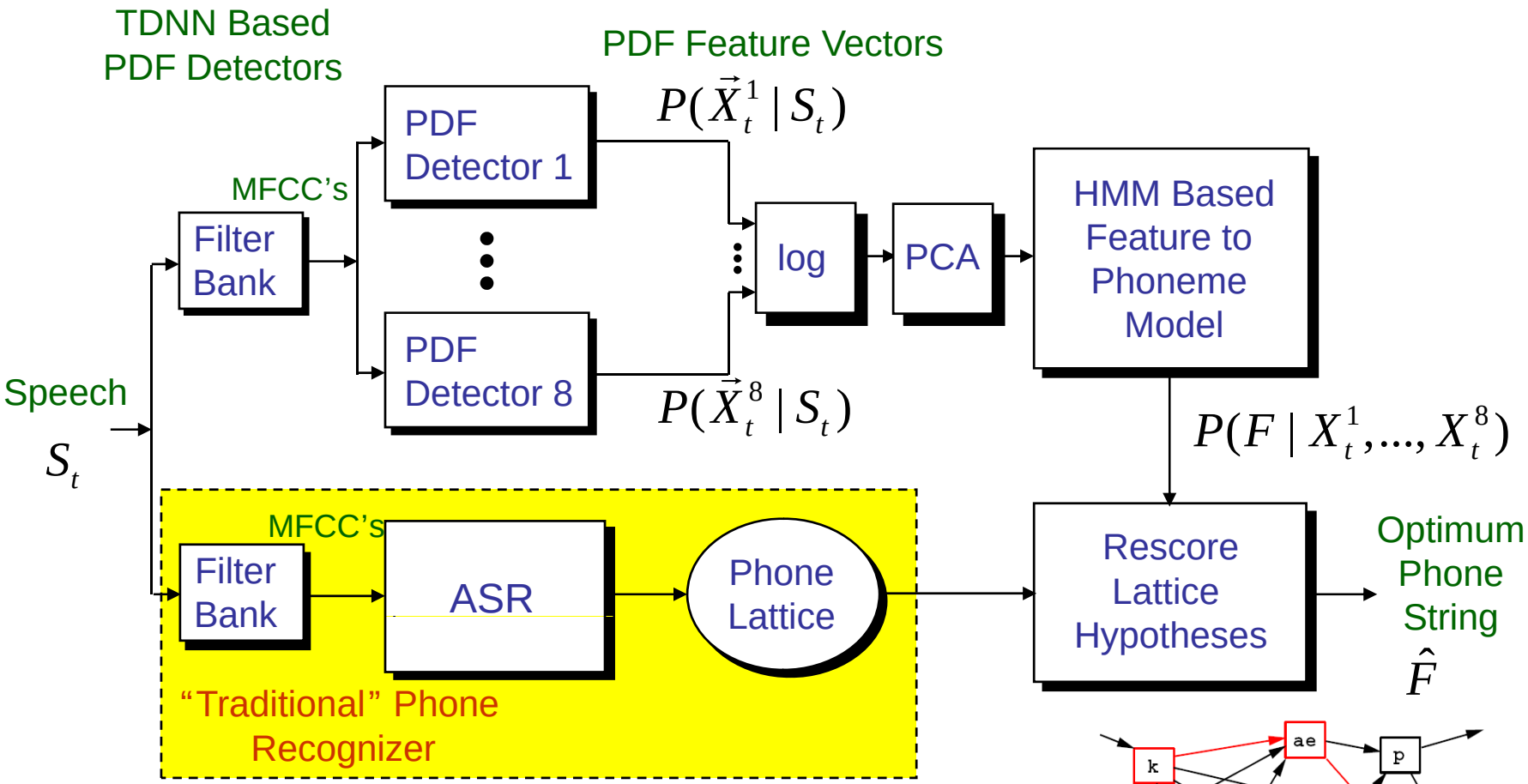
- PDF's as features in hidden Markov model ASR
- Disambiguating HMM based ASR lattice hypotheses through PDF re-scoring
- Review of the relationship between vocal tract shape and acoustic models
- Articulatory based model normalization / adaptation

PDFs as Features in HMM-Based ASR



- PDF Integration / Synchronization [Kirchhoff et al, 2000] [Stuker et al, 2003][Metz et al, 2003]
  - Coupled Features – Single observation stream: $P(s_k \mid \mathbf{X})$
  - Independent Features – Separate streams of PDFs integrated at the state level: $\prod_{i=1}^{N} P(s_t \mid X_t^i)$
  - Unsynchronized Features – Use of syllable rather than phone-based acoustic units
    - Articulatory synchronization believed to occur at syllable boundaries

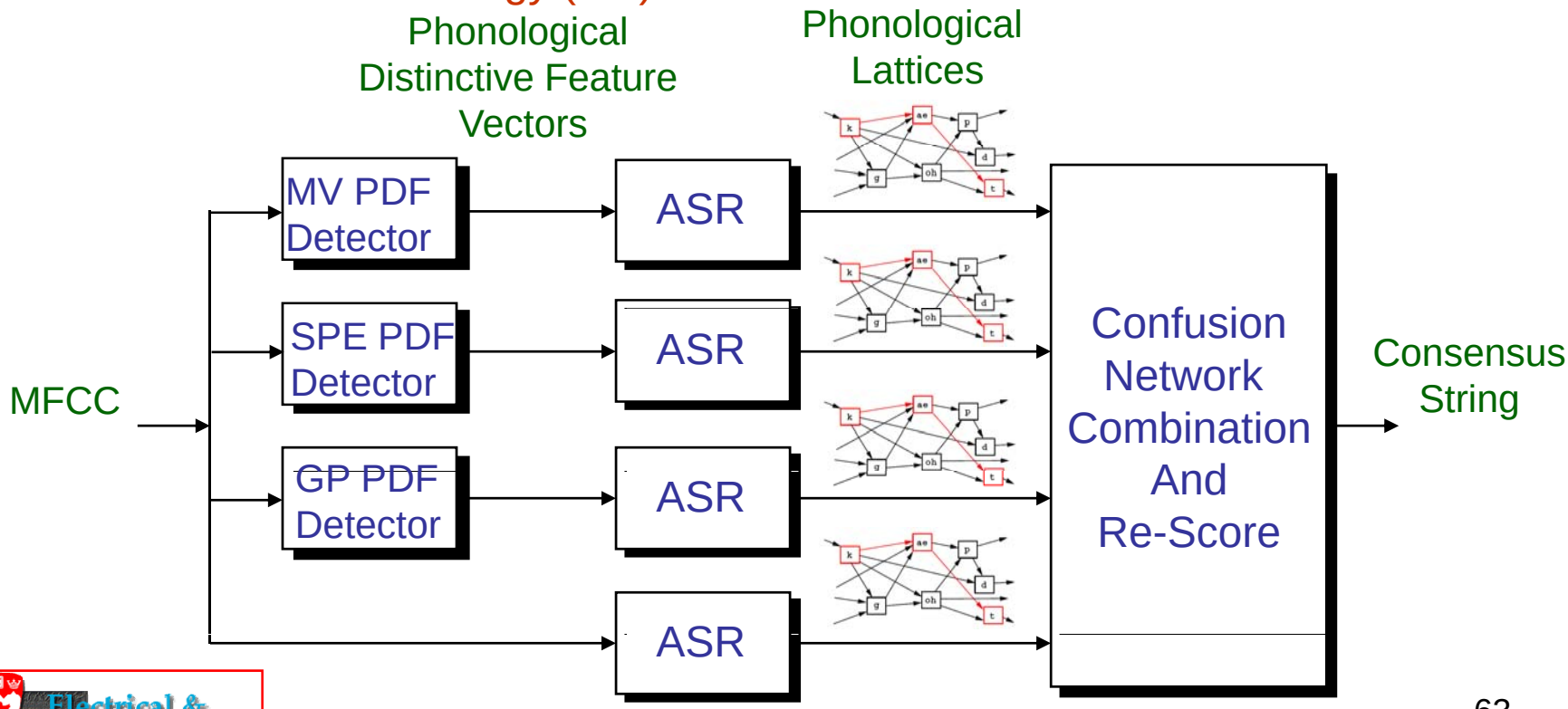
Disambiguating ASR Hypotheses by PDF Rescoring



- Used for re-scoring TIMIT phone lattices [Rose et al, 2006]
  - PAC increase from 69.1% to 72.5% with PDF re-scoring

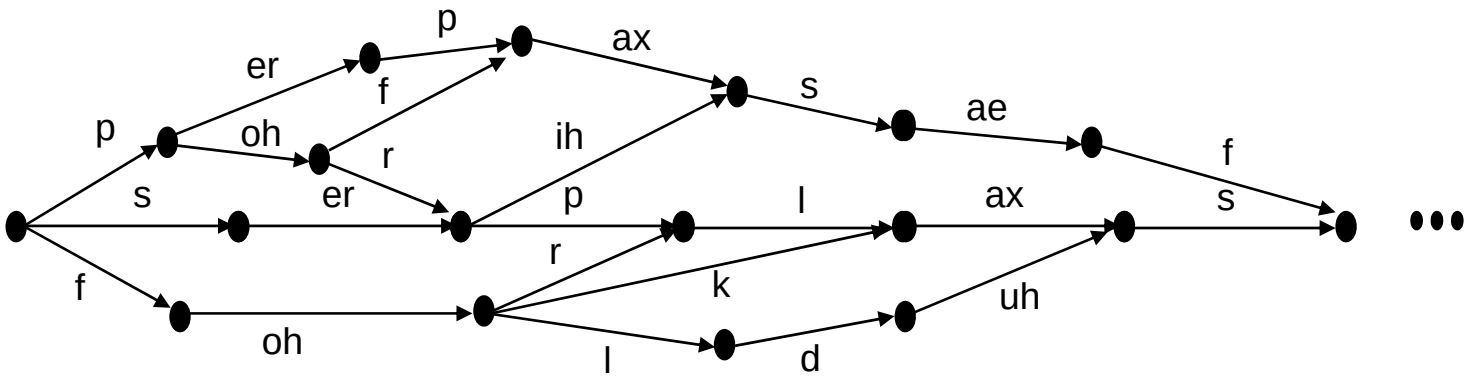
Confusion Network Combination

- Are different Phonological Distinctive Feature systems complementary?
- Combine phone lattices from features obtained from 3 different systems:
  - Multi-valued features (MV)
  - "Sound Patterns of English" features (SPE)
  - Government Phonology (GP)

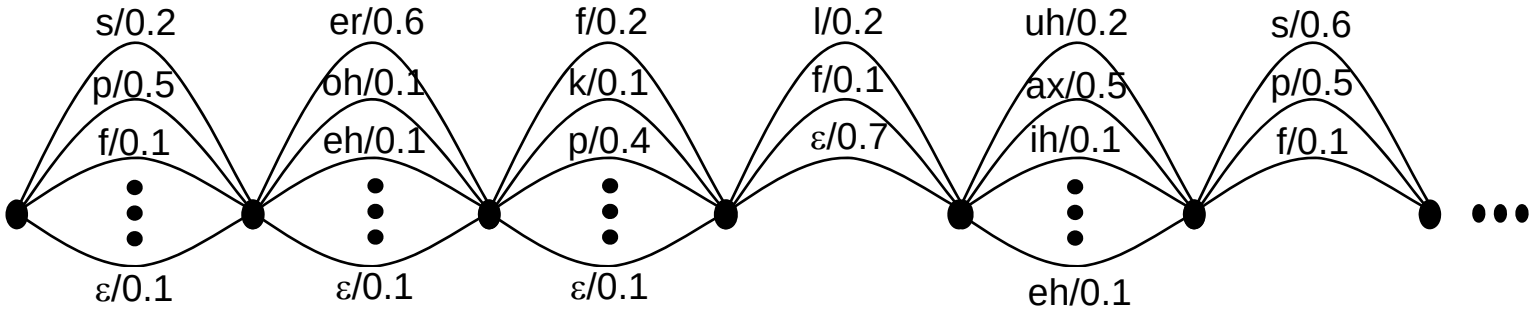


Confusion Network Combination

- Combine phone lattices produced from multiple DFDs …



… Into a confusion network …



… and re-score

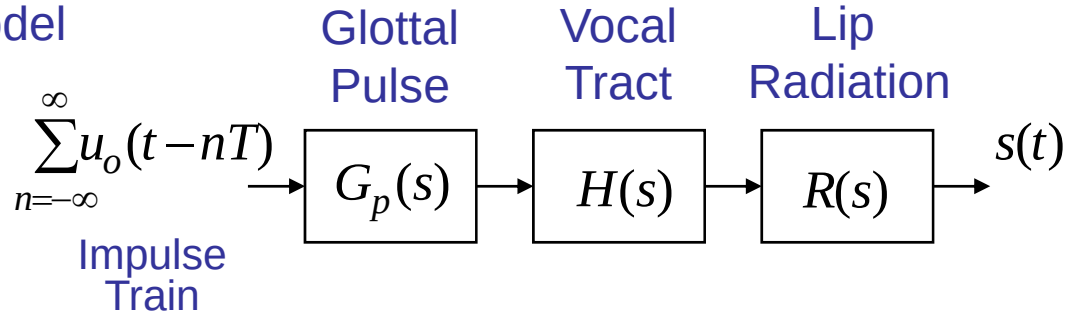
TIMIT Phone Recognition Accuracy	
MFCC	69.1%
MFCC+GP+MV+SPE	74.3%

Integrating Speech Production Models in Traditional ASR Systems

- PDF's as features in hidden Markov model ASR
- Disambiguating HMM based ASR lattice hypotheses through PDF re-scoring
- Review of the relationship between vocal tract shape and acoustic models
- Articulatory based model normalization / adaptation

Review: From Vocal Tract Shape to Acoustics - Theory of Speech Production

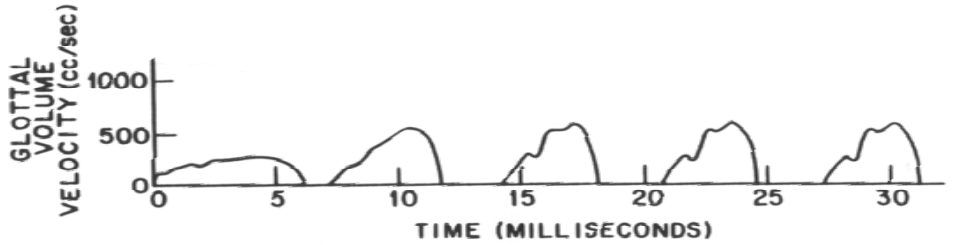
Speech Production Model for Voiced Sounds



Relate sound pressure level at the mouth, s(t), to the volume velocity at the glottis, u(t)

$$V(s) = G_p(s)H(s)R(s)$$

Glottal Pulses: Input Volume Velocity

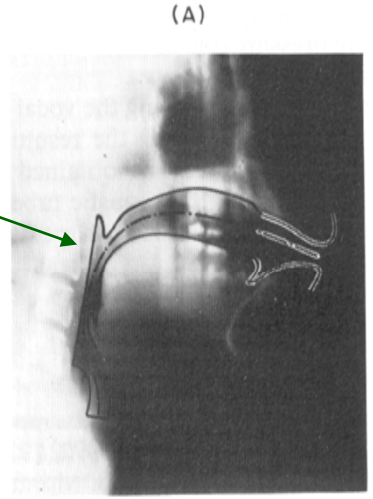


Sound Pressure Level at the Mouth

Vocal Tract Model

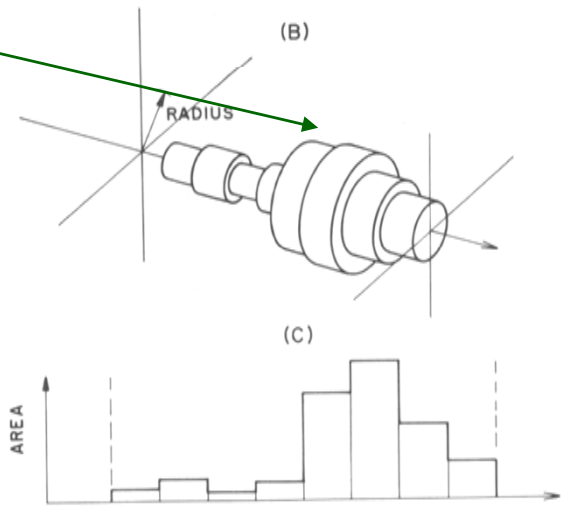
Model assumptions:

- Quasi-steady flow from pulsating jet in the larynx (more on this latter)
- Plane wave propagation through a series of concatenated acoustic tubes (cross sectional area << wave length)



Typical Wavelength: $\lambda = c / f \approx \frac{331 \text{ m/s}}{100 \text{ Hz}} = 3.3 \text{ meters}$

Typical Cross Sectional Area: $Area \approx 3 \text{ cm}$



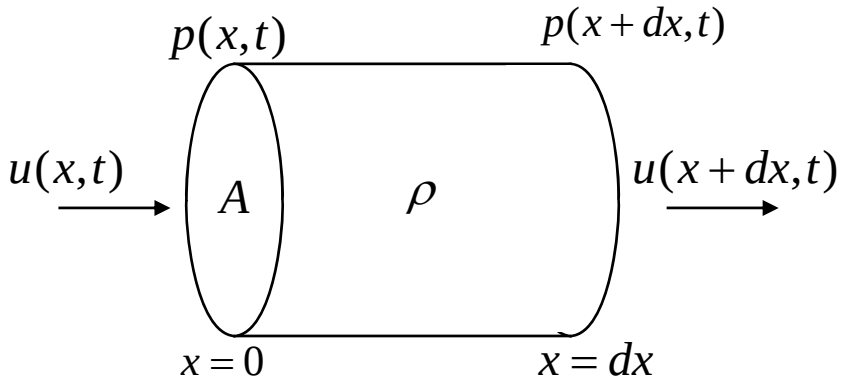
Vocal Tract Shape → Formants

1. Wave equation for acoustic tube
2. Acoustic tube transfer function
3. Tube formants

From Vocal Tract Shape to Formants – Acoustic Tube Model

[From Flanagan, "Analysis, Synthesis, and Perception", 1972]

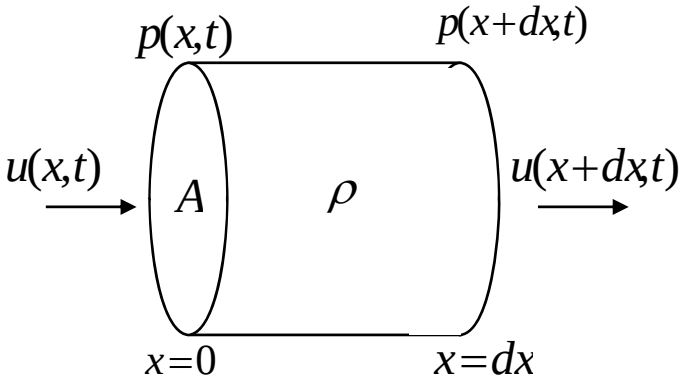
Cylindrical Tube of Length dx:



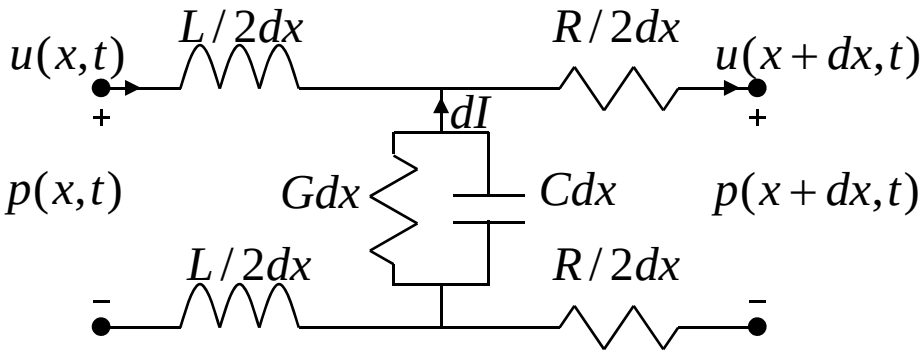
- Motion of Air through tube is characterized entirely by
  - Volume velocity: $u(x,t) = U(x)e^{st}$
  - Pressure: $p(x,t) = P(x)e^{st}$

A	Cross sectional area
ρ	Density of air
$\rho A dx$	Mass of air in tube
P_o	Atmospheric pressure
$P_o + p(x,t)$	Total pressure in tube

Electrical Analog of Acoustic Tube



Acoustic Tube

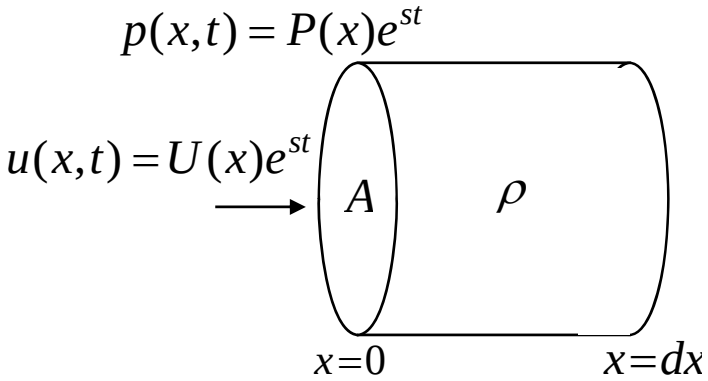


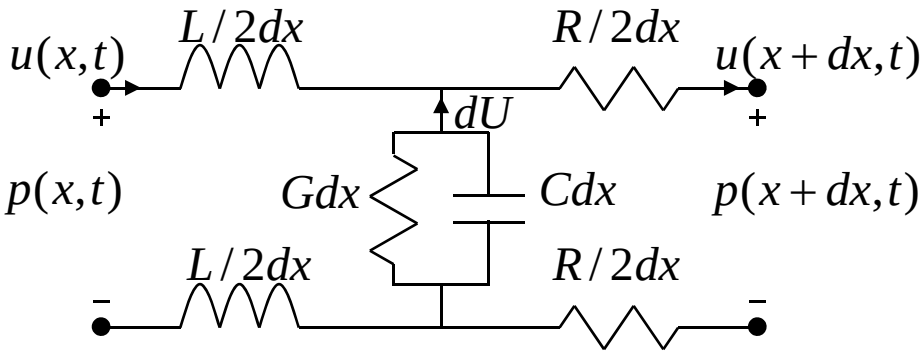
Electrical Analog

The relationship between current and voltage in the electrical circuit is equivalent to the relationship between volume velocity and pressure in the acoustic tube

Quantity	Acoustic	Electrical
$p(x,t)$	Pressure	Voltage
$u(x,t)$	Volume Velocity	Current
$L = \rho / A$	Inertance	Inductance
$C = A / \rho c^2$	Compliance	Capacitance
R	Viscous Friction	Series Resistance
G	Heat Loss	Shunt Conductance

Electrical Analog of Acoustic Tube





Apply Kirchoff's Laws to get:

1. Coupled Wave Equations:

$$\frac{dP(x)}{dx} = -zU(x)$$

$$\frac{dU(x)}{dx} = -yP(x)$$

where: $z = Ls + R \quad y = Cs + G$

2. Time Independent Wave Equations:

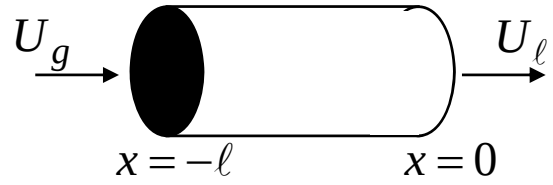
$$\frac{d^2P(x)}{dx^2} = zyP(x)$$

$$\frac{d^2U(x)}{dx^2} = zyU(x)$$

Find Transfer Function of a Single Acoustic Tube

Glottis:
Acoustic: closed ended
Electrical: open circuit

Lips:
Acoustic: open ended
Electrical: short circuit



Transfer Function

$$H(s) = \frac{U_\ell}{U_g} = \frac{U(0)}{U(-\ell)}$$

Estimate transfer function by applying boundary conditions to:

Solution to Coupled Wave Equations:

$$U(x) = U_+ e^{\gamma x} + U_- e^{-\gamma x}$$

$$P(x) = P_+ e^{\gamma x} + P_- e^{-\gamma x}$$

$$P_+ = -\sqrt{\frac{z}{y}} U_+$$

$$P_- = \sqrt{\frac{z}{y}} U_-$$

where propagation constant is: $\gamma = \pm\sqrt{zy}$

Transfer Function: $H(s) = \frac{U(0)}{U(-\ell)} = \frac{1}{\cosh \gamma \ell}$

Acoustic Tube Resonant Frequencies

Poles of Transfer Function: $H(s) = \frac{1}{\cosh \gamma \ell}$

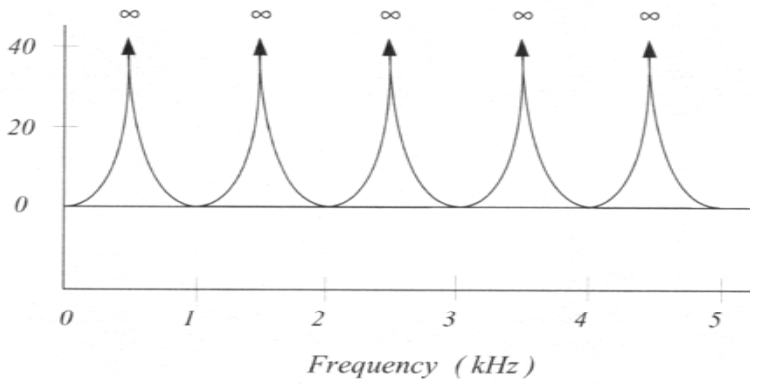
for the lossless case (R=G=0): $\gamma = [(sL+R)(sC+G)]^{\frac{1}{2}} = j\omega\sqrt{LC}$

occur when: $\omega_n \sqrt{LC}\ell = \frac{(2n-1)}{2}\pi$

$$\Rightarrow f_n = \frac{1}{4\sqrt{LC}\ell}(2n-1)$$

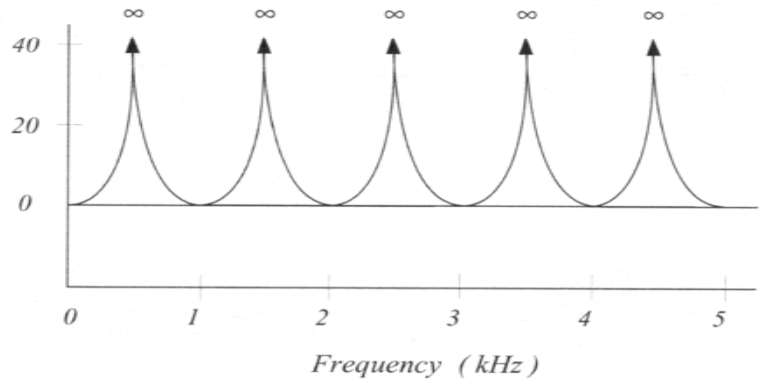
Typical Values: $\ell = 17.5cm \quad \sqrt{LC} = \sqrt{\frac{\rho}{A}\frac{A}{\rho c^2}} = \frac{1}{c} \approx 0.003 \quad \Rightarrow f_1 \approx 500Hz$

Transfer function for lossless acoustic tube contains equally space, zero bandwidth spectral resonances (formants):



Frequency Warping Based Speaker Normalization

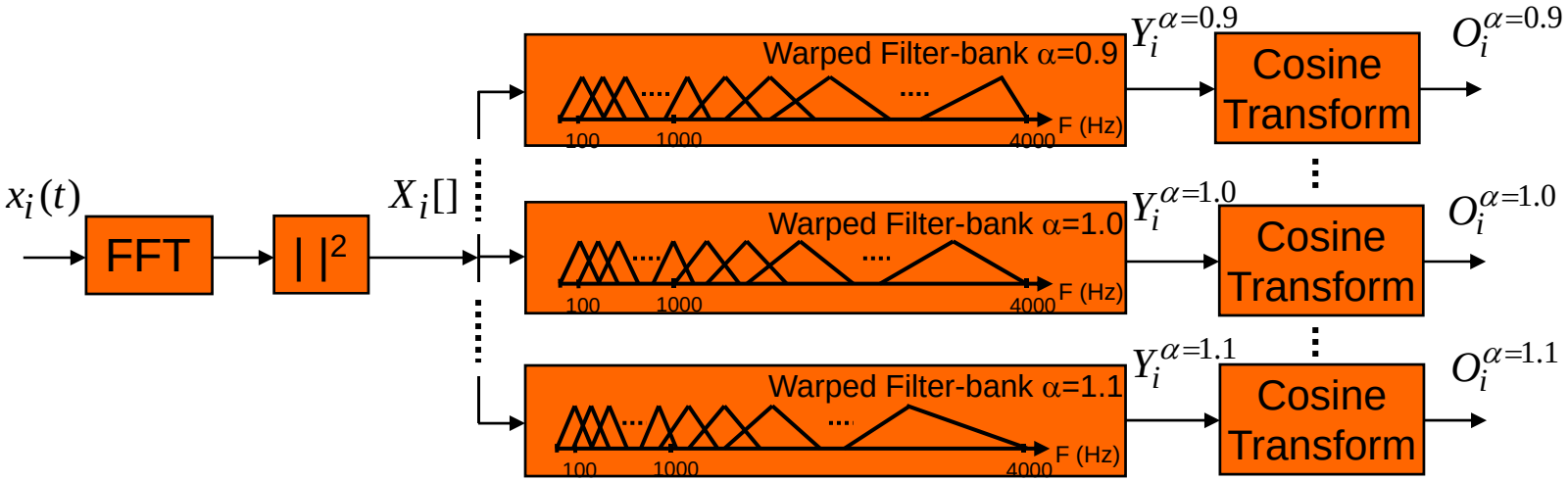
- Single tube model of reduced shwa vowel with length 17.5 cm will have formant frequecies 500 Hz, 1500 Hz, 2500 Hz, …



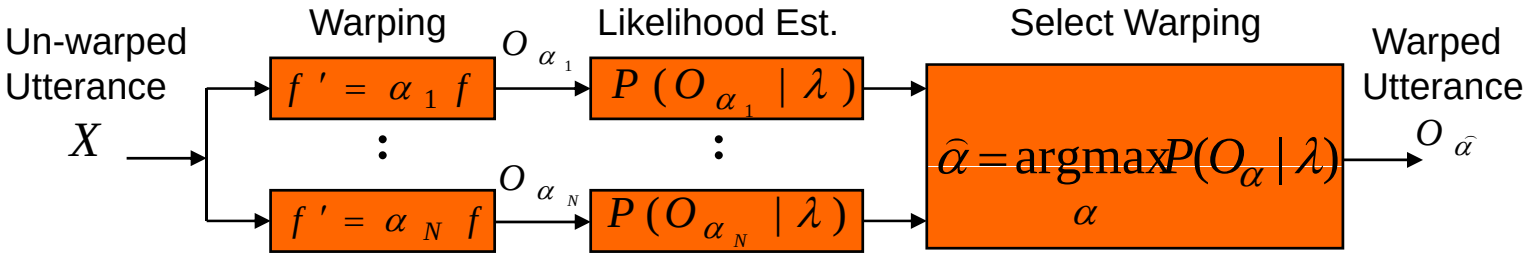
- Tube length ℓ and formant frequencies will vary among speakers according to $f_n \approx (2n-1)/4\ell c$
- ***Implies that the effects of speaker dependent variability can be reduced by frequency normalization***

Frequency Warping Based Speaker Normalization

- Normalize for speaker specific variability by linearly warping frequency axis, $f' = \alpha f$
- Warping can be performed by warping the mel-scale filter-bank [Lee and Rose, 1998]



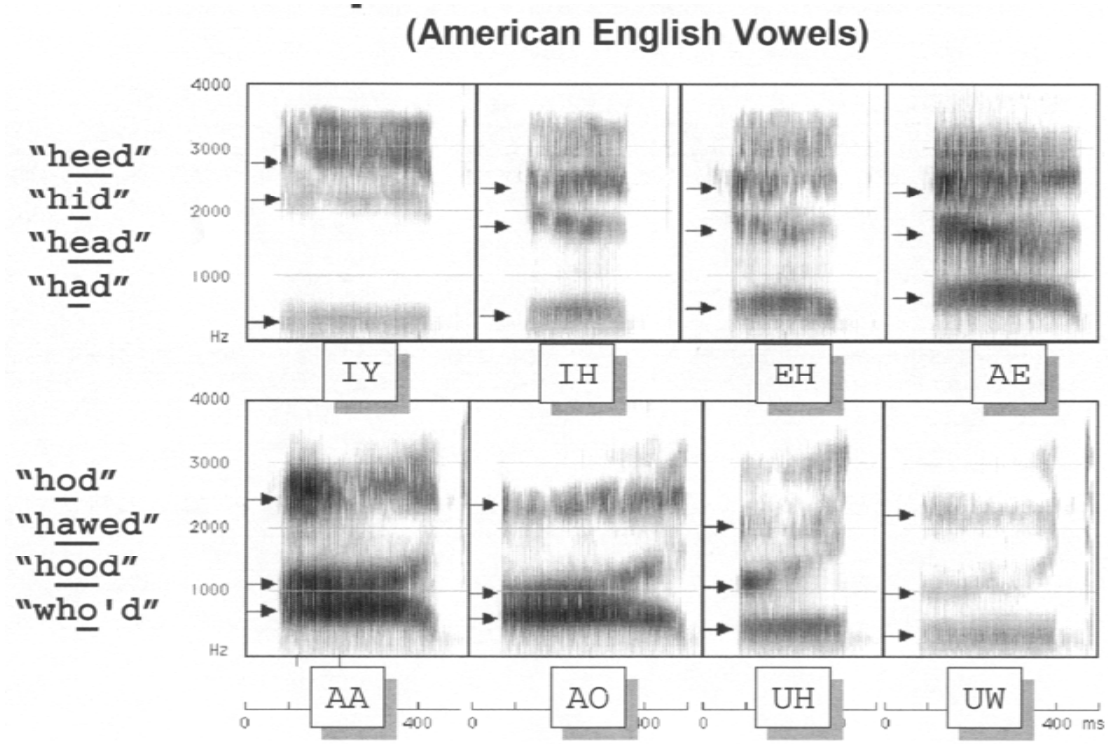
- Optimum warping factor found by performing ensemble search to maximize $P(O_\alpha \mid \lambda)$



- HMM model is trained from warped utterances to obtain a more “compact” model

Relationship Between Vocal Tract Shape and Formants

- In general, formant frequencies for different phonemes are a more complicated function of vocal tract shape:



[Jurafsky and Martin, 2008]

- ***Suggests that frequency warping based speaker normalization should be phoneme or PDF dependent …***

Time Dependent Frequency Warping Based Speaker Normalization

- Localized estimates of frequency warping based speaker normalization transformations can be obtained by optimizing a global criterion

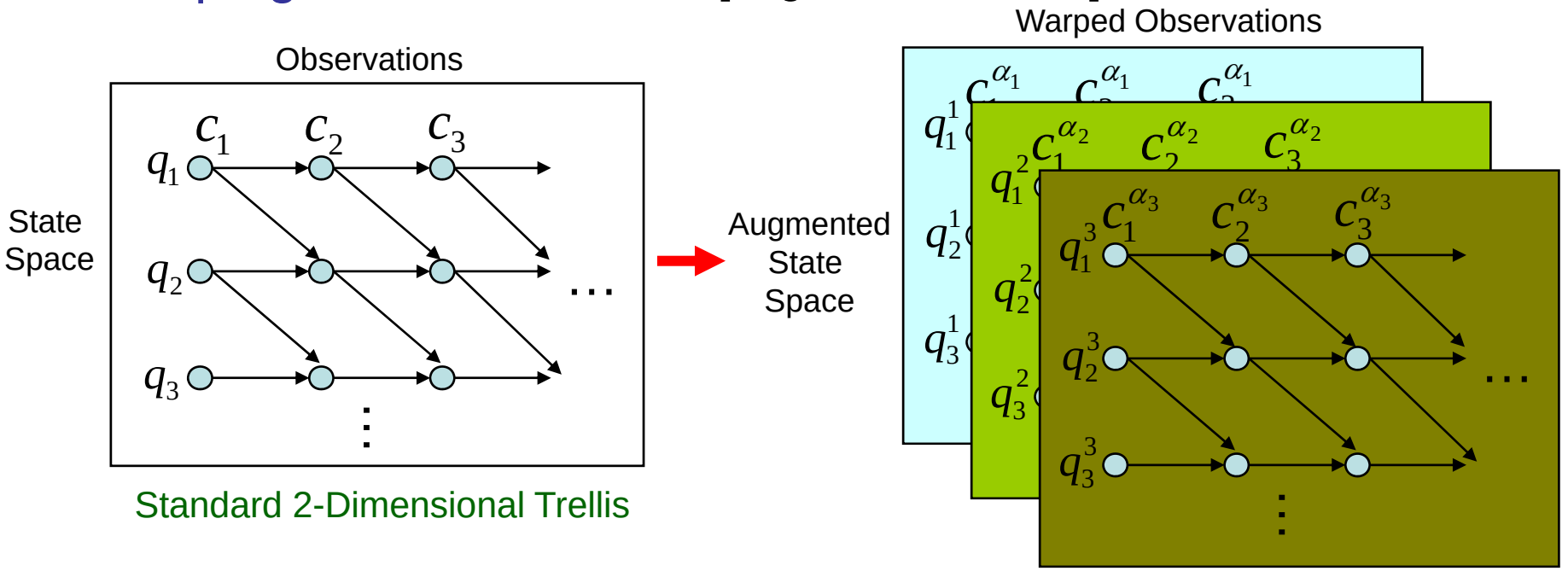
- Implement a decoder that simultaneously optimizes frame based acoustic likelihood and warping likelihood

- ***Augment the state space of the Viterbi decoder in ASR***
  [Miguel et al, 2005]

- There must be other speech production oriented adaptation normalization approaches!

Augmented State Space Acoustic Decoder

- "3D" Trellis: Augment HMM state space to incorporate warping factor ensemble [Miguel et al, 2008]



- Modified Viterbi Algorithm:

$$\phi_{j,n}(t) = \max_{i \in I, \alpha_m \in A} \{\phi_{j,m}(t-1) a_{i,j}^{m,n}\} b_j(c_t^{\alpha_m})\}$$

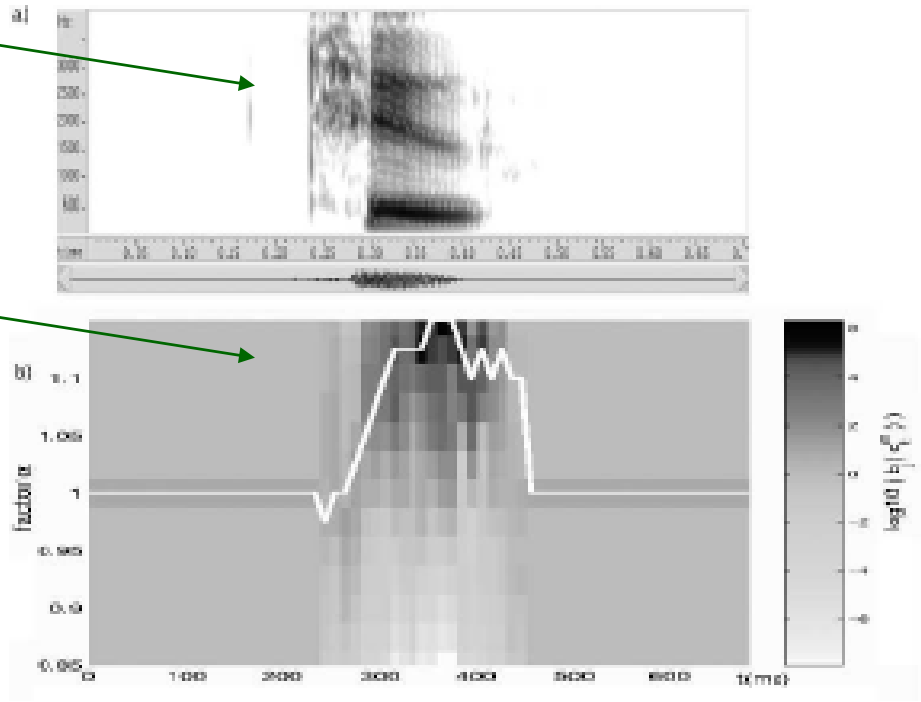
Decoding and Normalization Performed in a Single Pass

Frequency Warping Based Speaker Normalization

- Modify frequency warping based normalization to facilitate global optimization of ***frame based*** frequency warping

Utterance of the word “two”

Frame based Warping function likelihoods



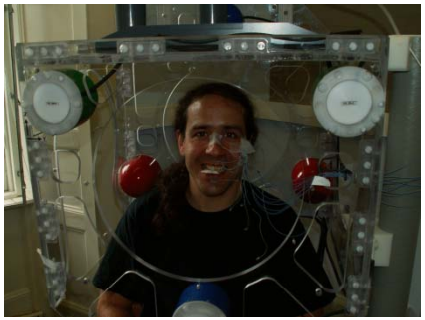
[Miguel et al, 2005]

- Augmented state space decoder – ML procedure to select from a discrete ensemble of warping functions ***for each frame***

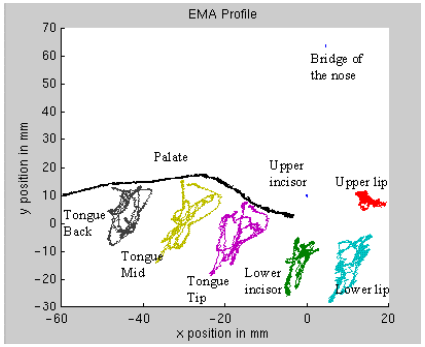
3. Resources

- Articulatory Measurement and Clinical Tools
- Corpora
- Workshops

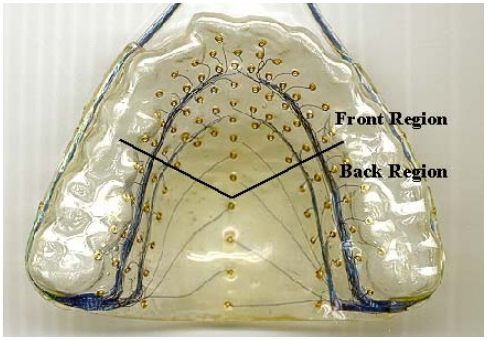
Direct Articulatory Measurements



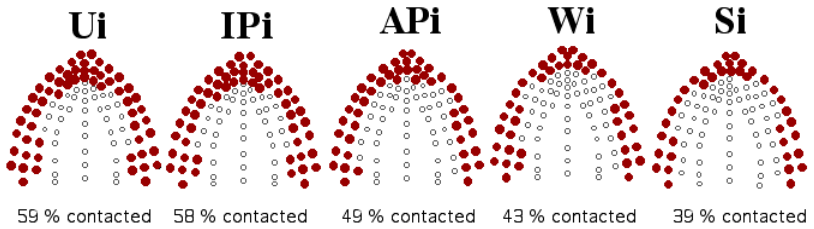
3D Articulagraph in Edinburough Speech Production Facility



2D EMA Trajectories from Oxford University Phonetics Lab



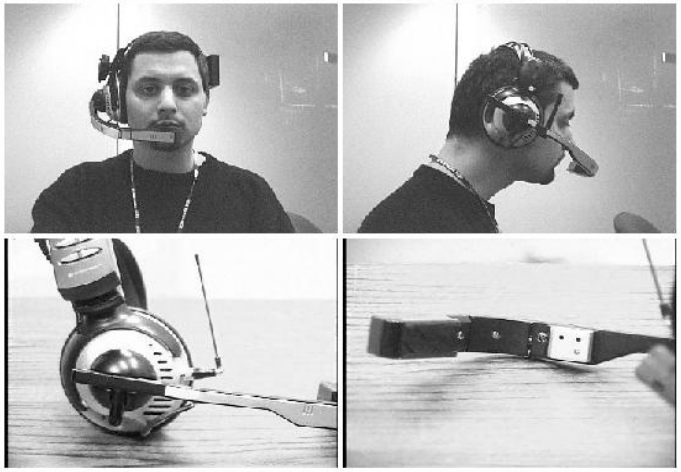
Electropalatograph (EPG) from UCLA Phonetics Lab



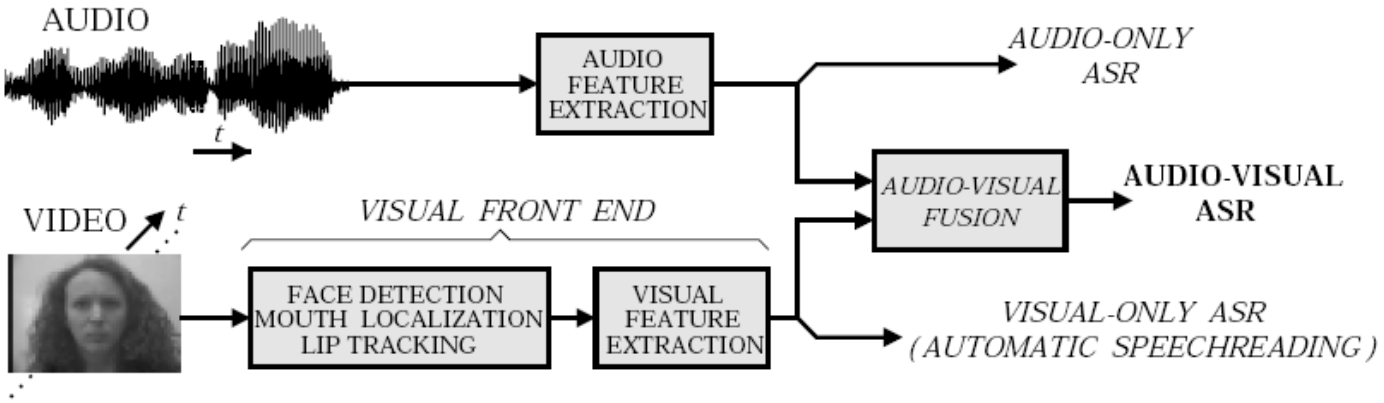
Linguopalatal contact measurements for different prosodic positions

"Partial" Direct Measurements - Visual Information

- Partial direct articulatory measurements fused with acoustic information in audio-visual ASR [Potamianos et al, 2004]



IBM Audio-Visual Headset
[Potamianos et al, 2004]



Fusing visual and acoustic measurements
[Potamianos et al, 2004]

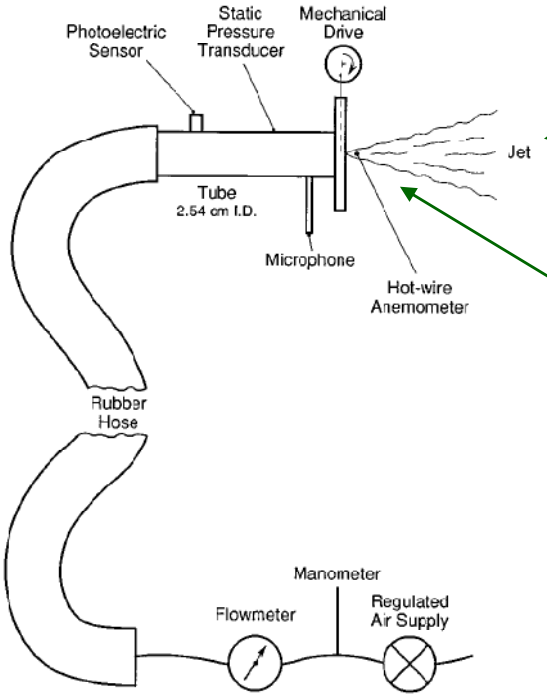
“Partial” Direct Measurements – Glottal Information

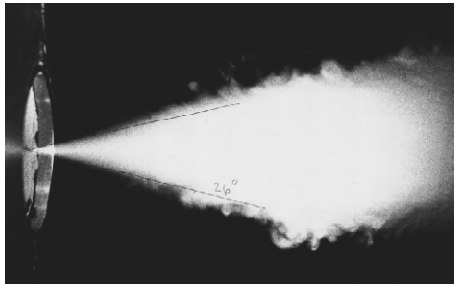
- Glottal Electro-Magnetic Sensors (GEMS):
  - Very low power radar-like sensors [Burnett et al, 1999]
  - Positioned Near Glottis: Measures motion of rear tracheal wall
  - Developed at Laurence Livermore and Commercialized by Aliph
- Research programs have investigated their use in very high noise environments

Hot-Wire Anenometer and Vocal Tract Aerodynamics

- Hot-Wire Anenometers have been used for verifying aeroacoustic models of phonation [Mongeau, 1997]

Apparatus for simulating the excitation of plane waves in tubes by small pulsating jets through time varying orifices [Mongeau, 1997]



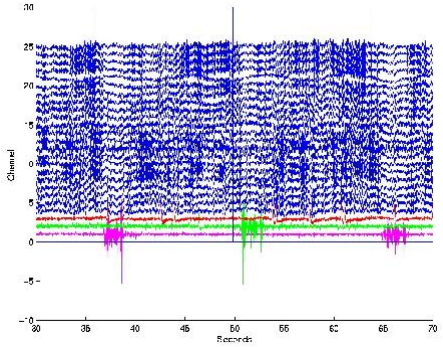


Pulsating jet
[Mongeau, 1997]

Hot Wire Anenometer

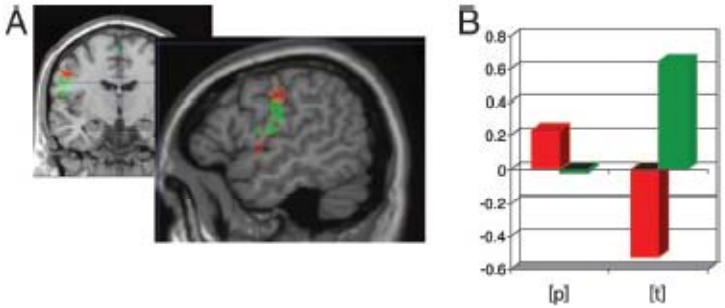
Clinical Tools - MRI and EEG

EEG Sensors in McGill Speech Motor Control Lab



Averaging of signals to separate evoked responses to various stimuli from background activity

Magnetic Resonance Imaging in McGill Speech Motor Control Lab



MRI images – Relationship between perception and articulatory motor control [Pulvermuller, 2006]

Resources – Corpora

- Phonetically labeled speech corpora
  - TIMIT
  - ICSI Switchboard transcription project [Greenberg, 2000]
  - Buckeye Corpus (Ohio State)
  - Svitchboard [King et al, 2006]
- Direct Articulatory Measurements
  - Wisconsin x-ray microbeam articulatory corpus
  - MOCHA – Parallel acoustic articulatory recordings (EMA, EPG, EGG measurements) of a handful of speakers reading ~450 sentences (Edinburgh) [Wrench et al, 2000]
  - Audio-Visual TIMIT corpus (AVTIMIT) [MIT]
  - CUAVE – Audio-visual corpus [Patterson, 2002]

Resources – Workshops

- 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

Speech Production Topics Not Covered

- Manifold Based Approaches
  - Assume that speech itself is constrained to lie in some subspace but we don not know the dimensionality of the subspace
  - Laplacian Eigenmaps, Locality Preserving Projections, ISOMAP
  - Consider practical gains from mapping data onto a space of intrinsic dimension associated with a non-linear manifold [He and Niyogi][Nilson and Kleijn][Tang and Rose]

- Speech modeling based on nonlinear vocal tract air-flow dynamics [Maragos et al]