



NUANCE

Integrating High-Level Information for Robust Speaker Recognition

Larry P. Heck

**Director, Speech R&D
Nuance Communications**

**CLSP Workshop 2002
The Johns Hopkins University**

High-Level Information for Speaker Recognition Outline

- Introduction: Background & Motivations
- Problem Formulation: Component Factorization
- Speaker-Dependent Language Modeling
- Speaker-Dependent Prosodic Modeling
- Summary



High-Level Information for Speaker Recognition Outline

- Introduction: Background & Motivations
- Problem Formulation: Component Factorization
- Speaker-Dependent Language Modeling
- Speaker-Dependent Prosodic Modeling
- Summary



High-Level Information for Speaker Recognition

Motivations

- Humans use several levels of perceptual cues for speaker recognition

High-level cues
(learned traits)



Low-level cues
(physical traits)

??

Hierarchy of Perceptual Cues

Semantics, diction, pronunciations, idiosyncrasies	Socio-economic status, education, place of birth
Prosodics, rhythm, speed intonation, volume modulation	Personality type, parental influence
Acoustic aspect of speech, nasal, deep, breathy, rough	Anatomical structure of vocal apparatus

Difficult to automatically extract



Easy to automatically extract

- Current systems rely on low-level information in speech
 - Short time extent analysis windows (20-30 ms)
 - Spectral energy-based (MFCCs)



High-Level Information for Speaker Recognition

Motivations

- Examples of high-level information in speech signal:
 - Speaking rate
 - Timing Patterns
 - Pitch Patterns
 - Idiosyncratic word/phrase usage
 - Idiosyncratic pronunciations
 - Laughter

- Why use high-level information for speaker recognition?
 - Robustness:
 - Reliance on low-level features has exacerbated lack of robustness to noise and channel mismatch
 - Higher-level cues are not as affected by noise or channel mismatch
 - “Orthogonal” Knowledge: potentially significant *additive* wins



High-Level Information for Speaker Recognition Outline

- Introduction: Background & Motivations
- Problem Formulation: Component Factorization
- Speaker-Dependent Language Modeling
- Speaker-Dependent Prosodic Modeling
- Summary



High-Level Information for Speaker Recognition Problem Formulation

□ Speaker recognition problem:

$$\operatorname{argmax} P(S | O)$$

S = speaker

O = observations

□ Observations can take many forms:

$$O = \{X, W, F, C, \dots\}$$

- Low-level features (MFCC): $X = \{x_1, x_2, \dots, x_T\}$
- Words/phrases/phonemes: $W = \{W_1, W_2, \dots, W_n\}$
- Prosodic Events: $F = \{F_1, F_2, \dots, F_p\}$
- Channel information: $C = \{\text{handheld/handsfree landline \& wireless tel., PC microphones, conference room mics.}\}$

$$\operatorname{argmax}_S P(S | X, W, F, C)$$



High-Level Information for Speaker Recognition Problem Formulation

- Factorization into separate knowledge components:

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \frac{P(X|S, W, F, C)}{P(X|W, F, C)} \cdot \frac{P(W|S)}{P(W)} \cdot \frac{P(F|S, W)}{P(F|W)} \cdot \frac{P(C|S)}{P(C)} \cdot P(S)$$

Text-Dependent
Speaker Recognizer



Where W and C are assumed independent, with F and C also independent ([See derivation](#))

- Low-level features (MFCC): $X = \{x_1, x_2, \dots, x_T\}$
- Words/phrases/phonemes: $W = \{W_1, W_2, \dots, W_n\}$
- Prosodic Events: $F = \{F_1, F_2, \dots, F_p\}$
- Channel information: $C = \{\text{handheld/handsfree landline \& wireless tel., PC microphones, conference room mics.}\}$

Note: combination of knowledge components will typically require tunable weightings (e.g., frame-level independence assumptions) and log (dynamic range)



High-Level Information for Speaker Recognition Problem Formulation

- Factorization into separate knowledge components:

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \frac{P(X|S, W, F, C)}{P(X|W, F, C)} \cdot \frac{P(W|S)}{P(W)} \cdot \frac{P(F|S, W)}{P(F|W)} \cdot \frac{P(C|S)}{P(C)} \cdot P(S)$$

Speaker-Dependent
Language Model



Where W and C are assumed independent, with F and C also independent ([See derivation](#))

- Low-level features (MFCC): $X = \{x_1, x_2, \dots, x_T\}$
- Words/phrases/phonemes: $W = \{W_1, W_2, \dots, W_n\}$
- Prosodic Events: $F = \{F_1, F_2, \dots, F_p\}$
- Channel information: $C = \{\text{handheld/handsfree landline \& wireless tel., PC microphones, conference room mics.}\}$

Note: combination of knowledge components will typically require tunable weightings (e.g., frame-level independence assumptions) and log (dynamic range)



High-Level Information for Speaker Recognition Problem Formulation

- Factorization into separate knowledge components:

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \frac{P(X|S, W, F, C)}{P(X|W, F, C)} \cdot \frac{P(W|S)}{P(W)} \cdot \frac{P(F|S, W)}{P(F|W)} \cdot \frac{P(C|S)}{P(C)} \cdot P(S)$$

Speaker-Dependent
Prosodic Model



Where W and C are assumed independent, with F and C also independent ([See derivation](#))

- Low-level features (MFCC): $X = \{x_1, x_2, \dots, x_T\}$
- Words/phrases/phonemes: $W = \{W_1, W_2, \dots, W_n\}$
- Prosodic Events: $F = \{F_1, F_2, \dots, F_p\}$
- Channel information: $C = \{\text{handheld/handsfree landline \& wireless tel., PC microphones, conference room mics.}\}$

Note: combination of knowledge components will typically require tunable weightings (e.g., frame-level independence assumptions) and log (dynamic range)



High-Level Information for Speaker Recognition Problem Formulation

- Factorization into separate knowledge components:

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \frac{P(X|S, W, F, C)}{P(X|W, F, C)} \cdot \frac{P(W|S)}{P(W)} \cdot \frac{P(F|S, W)}{P(F|W)} \cdot \frac{P(C|S)}{P(C)} \cdot P(S)$$

“Channel Profile”
for Speaker



Where W and C are assumed independent, with F and C also independent ([See derivation](#))

- Low-level features (MFCC): $X = \{x_1, x_2, \dots, x_T\}$
- Words/phrases/phonemes: $W = \{W_1, W_2, \dots, W_n\}$
- Prosodic Events: $F = \{F_1, F_2, \dots, F_p\}$
- Channel information: $C = \{\text{handheld/handsfree landline \& wireless tel., PC microphones, conference room mics.}\}$

Note: combination of knowledge components will typically require tunable weightings (e.g., frame-level independence assumptions) and log (dynamic range)



High-Level Information for Speaker Recognition Problem Formulation

- Factorization into separate knowledge components:

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \frac{P(X|S, W, F, C)}{P(X|W, F, C)} \cdot \frac{P(W|S)}{P(W)} \cdot \frac{P(F|S, W)}{P(F|W)} \cdot \frac{P(C|S)}{P(C)} \cdot P(S)$$

Speaker Prior
(CID, calling patterns, etc.)

Where W and C are assumed independent, with F and C also independent ([See derivation](#))

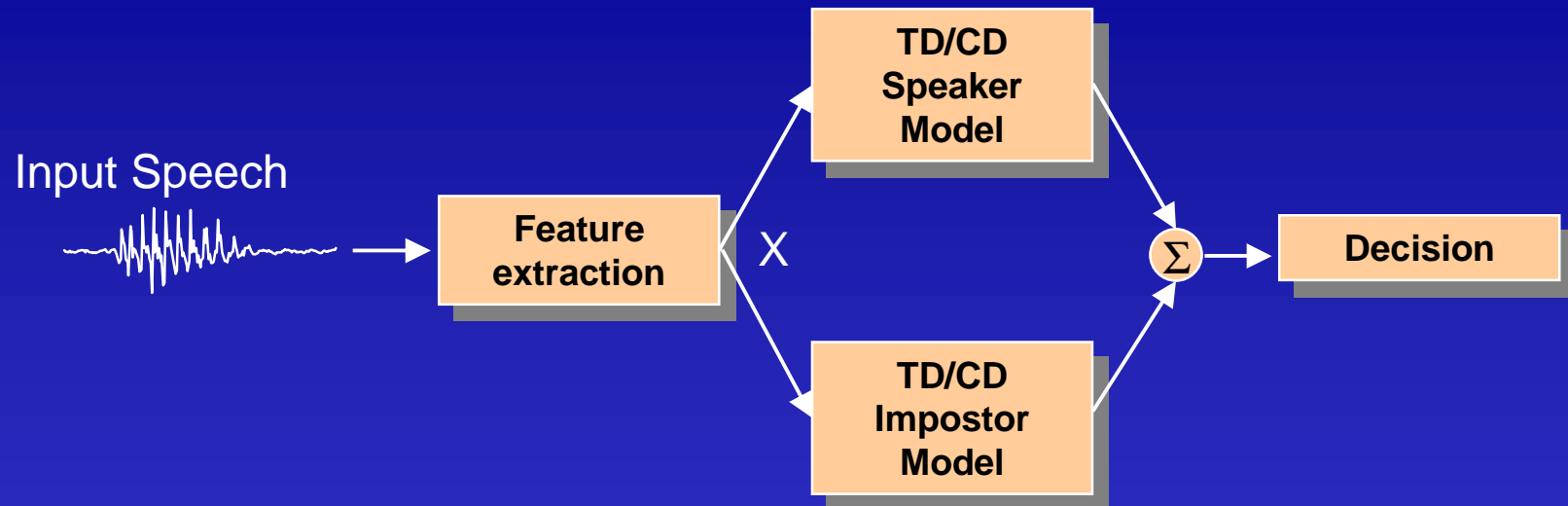
- Low-level features (MFCC): $X = \{x_1, x_2, \dots, x_T\}$
- Words/phrases/phonemes: $W = \{W_1, W_2, \dots, W_n\}$
- Prosodic Events: $F = \{F_1, F_2, \dots, F_p\}$
- Channel information: $C = \{\text{handheld/handsfree landline \& wireless tel., PC microphones, conference room mics.}\}$

Note: combination of knowledge components will typically require tunable weightings (e.g., frame-level independence assumptions) and log (dynamic range)



Text-Dependent Speaker Recognizer Represents Current State-of-the-Art

- Current state-of-art speaker recognizer implemented at LR detector
 - Model speaker with channel- and word/phrase/phoneme-dependent HMMs
 - Important to “normalize out” pronunciations, channel with speaker-independent BGM



Common Assumptions:

- Low-level features (X) not dependent on prosodics: $P(\mathbf{X}|S, X, W, F, C) \approx P(\mathbf{X}|S, X, W, C)$

- Observations are independent: $P(\mathbf{X}|S, X, W, C) \approx \prod_{i=1}^n P(x_i|S, X, W, C)$

$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

High-Level Information for Speaker Recognition Outline

- Introduction: Background & Motivations
- Problem Formulation: Component Factorization
- Speaker-Dependent Language Modeling
- Speaker-Dependent Prosodic Modeling
- Summary



Speaker-Dependent Language Modeling

Introduction: Word/Phrase Usage

□ Motivation

- Humans distinguish between familiar speakers much better than unfamiliar speakers → cue on idiosyncrasies
- Word/phrase cues for familiar speakers likely at higher levels

□ Prior Work

- Idiosyncratic Word/Phrase Usage: N-grams over words.
 - Heck (1998): tri-grams (with backoff to bi-,uni-grams)
 - ◆ ASR-based transcriptions (high confidence words only)
 - ◆ Smoothed SD-LMs with SI-LMs
 - Doddington (2001): bigram (with count thresholding)
 - ◆ Manual transcriptions
 - ◆ NIST 2001 Extended Data Task

□ Related Work

- Determining Authorship of Written Text (Holmes, 1985)

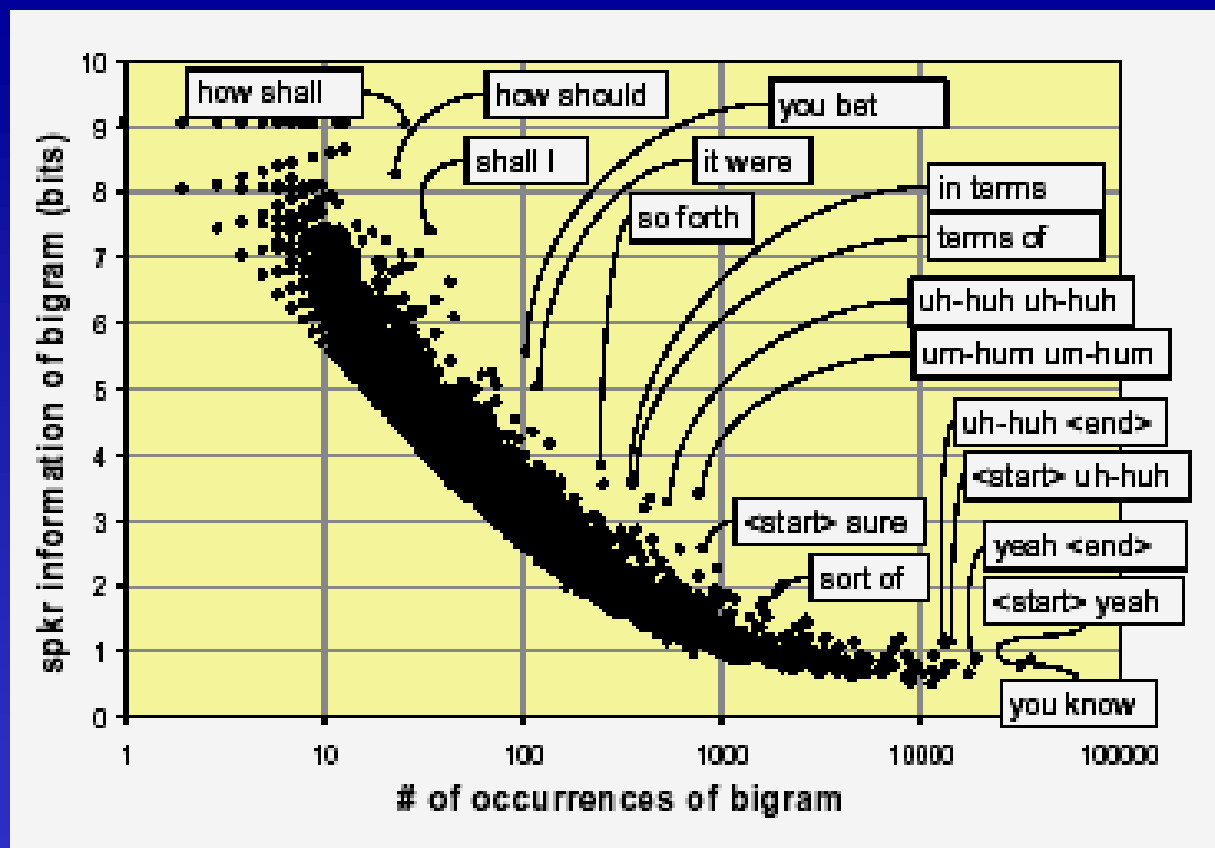
$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Language Modeling

Idiosyncratic Word/Phrase Usage

Analysis*

- Switchboard I
- Uniform prior of speakers
- 520 speakers
- Plotting N-gram entropy
- “how shall”:
 - Occurred 25 times
 - Only one speaker spoke this phrase
 - Occurred in ½ of 26 conversations for this speaker



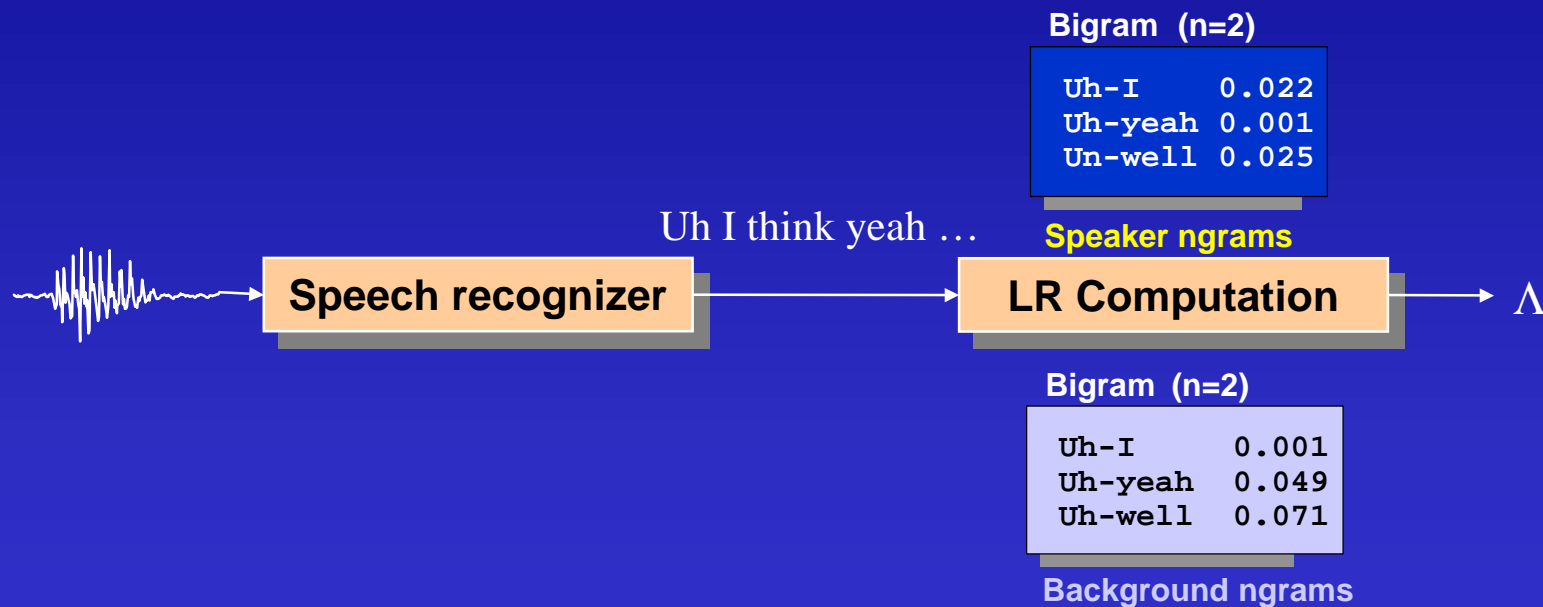
* Doddington (2001)

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Language Modeling

Idiosyncratic Word/Phrase Usage

- Training: estimate parameters of language models
 - $P(W|S)$: collect counts of n-grams for individual speaker
 - $P(W)$: collect counts of n-grams across many speakers
- Verification: compute LR score between speaker and background n-gram models



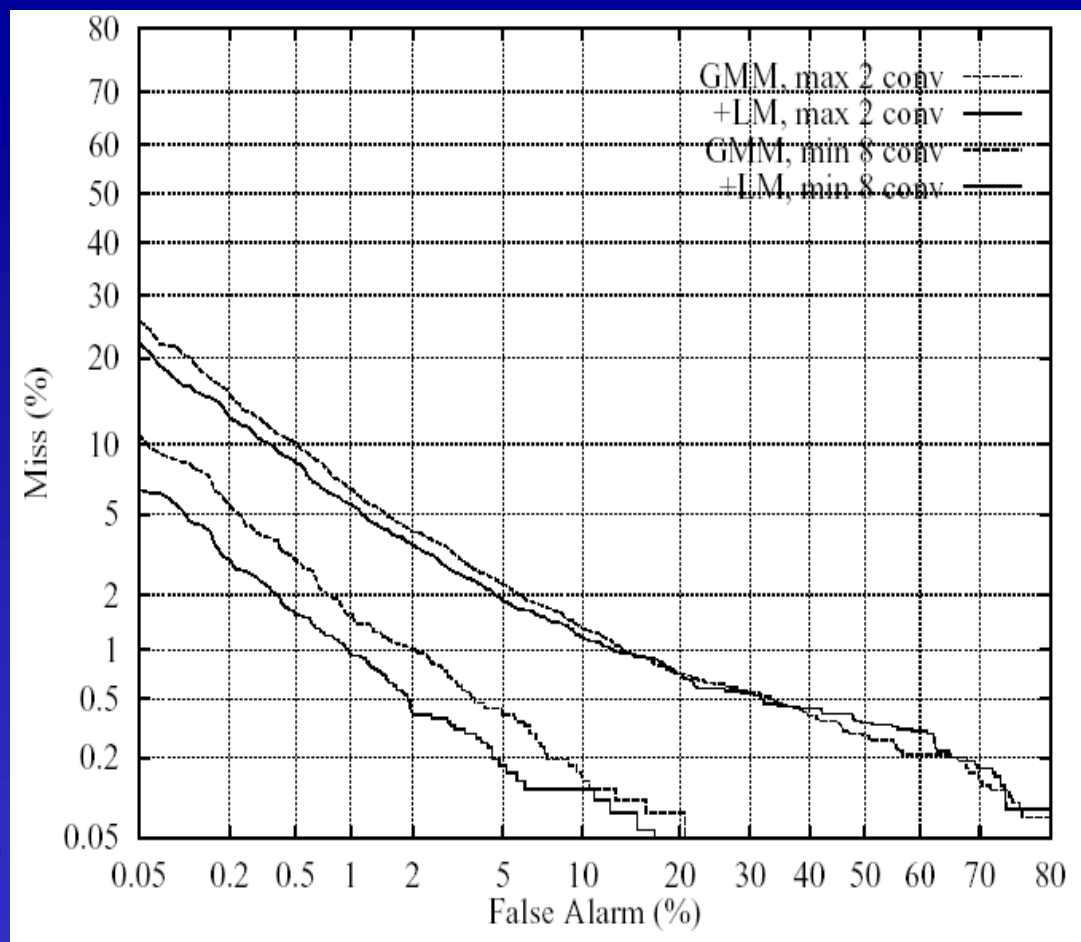
$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Language Modeling

Idiosyncratic Word/Phrase Usage

Performance*

- NIST 2001 Speaker Recognition Evaluation (Extended Data Task)
- All results in combination with GMM
- 1-2 conversation sides (3-6 min.)
 - ~10% improvement (FRR@2% FAR)
- 8-16 conversation sides (24-60 min.)
 - ~50% improvement (FRR@2% FAR)
 - ~20-25% EER w/ LM alone.



* Weber, Manganaro, Peskin, Shriberg (2002)

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

- **Average Word Length (number of letters)**
 - Mendenhall (1901): Shakespeare or Bacon?
 - Briengar (1963): Frequency of 2-4 letter words:
Was Quintus Curtius Snodgrass written by Mark Twain?

- **Sentence Length (number of words)**
 - C.B. Williams in 1940: log of sentence length has normal distribution
 - Morton in 1965: analyzed ancient Greek writings

- **Vocabulary Richness**
 - Thisted & Efron in 1987: Newly discovered poem by Shakespeare?

- **Hierarchical Cluster Analysis**
 - Holmes in 1992: Detected changes in authorship in Mormon scripture

$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

- The Best Features → Frequency of Function Words
(conjunctions, prepositions, & pronouns)
 - F. Mosteller & D. Wallace (1963) Authorship of *The Federalist* papers: Hamilton or Madison?
 - Practically “twins” w/ respect to average sentence length
 - Used functions words → successfully assigned authorship
 - Peng & Hengartner, 2002
 - Canonical Discriminant Analysis: determined small set of useful words

□ Why are Function Word-Based Features the Best ?

- Occur frequently and are easy to count and identify
- Highly variable between authors & consistent within author

• Indicates “*involuntary writing style*”

→ No thought about use of words: no real intentional meaning

→ Topic independent

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

- R.D. Peng and N.W. Hengartner, “Quantitative Analysis of Literary Styles”, *The American Statistician*, 56 (3), 175—185
 - Extension of work by Mosteller & Wallace on *The Federalist Papers*
 - Authorship discrimination between 9 authors
 - Used 69 function words from Miller-Newman-Friedman list

a	been	had	its	one	the	were
all	but	has	may	only	their	what
also	by	have	more	or	then	when
an	can	her	must	our	there	which
and	do	his	my	should	things	who
any	down	if	no	so	this	will
are	even	in	not	some	to	with
as	every	into	now	such	up	would
at	for	is	of	than	upon	your
be	from	it	on	that	was	

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

- Documents divided into blocks of 1700 words
 - Large enough to get stable statistic
 - Small enough to for “authorship-change” resolution
- In each block, counts of each word on word list are tallied

Author	Dates Lived	# of Blocks
Marlowe	1564-1593	56
Shakespeare	1564-1616	179
Milton	1608-1674	56
Austen	1775-1817	437
Dickens	1812-1870	598
Doyle	1859-1930	552
Kipling	1865-1936	157
Cather	1873-1947	237
London	1876-1916	299

$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

- Goal: Determine linear discriminants with LDA to separate authors
- Data structures:
 - G: a "1" in (i,j) position indicates block i was written by author j.
 - X: (block number by function word count)

$$G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \Rightarrow \begin{bmatrix} \text{a} & \text{all} & \text{also} & \text{an} & \text{and} & \dots \\ 40 & 34 & 54 & 20 & 78 & \dots \\ 35 & 10 & 0 & 7 & 44 & \dots \\ 46 & 2 & 0 & 3 & 4 & \dots \\ 40 & 12 & 31 & 6 & 14 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} = X$$

$$\text{Between Groups Covariance Matrix} = \frac{1}{n} X^T G (G^T G)^{-1} G^T X$$

$$\text{Total Covariance Matrix} = \frac{1}{n} X^T X$$

Eigenvalue Problem \rightarrow

$$X^T G (G^T G)^{-1} G^T X \beta_i = \lambda_i X^T X \beta_i$$

$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

- If $\beta_1, \beta_2, \dots, \beta_L$ are the discriminant functions, then:
 - $Y_i = X\beta_i$ are the “canonical vectors”
- Plots illustrate effectiveness of function words in detecting authors

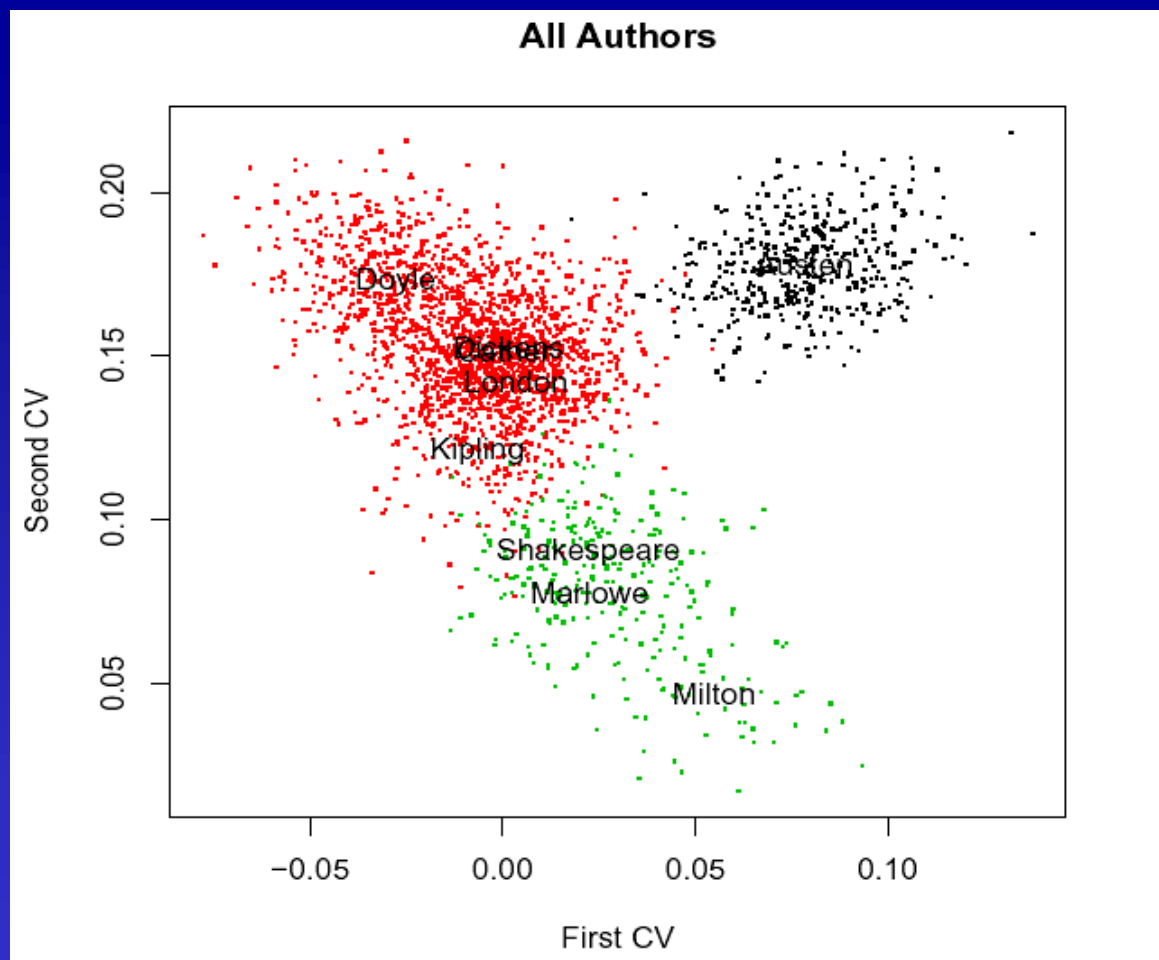


$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

- If $\beta_1, \beta_2, \dots, \beta_L$ are the discriminant functions, then:
 - $Y_i = X\beta_i$ are the “canonical vectors”
- Plots illustrate effectiveness of function words in detecting authors



$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

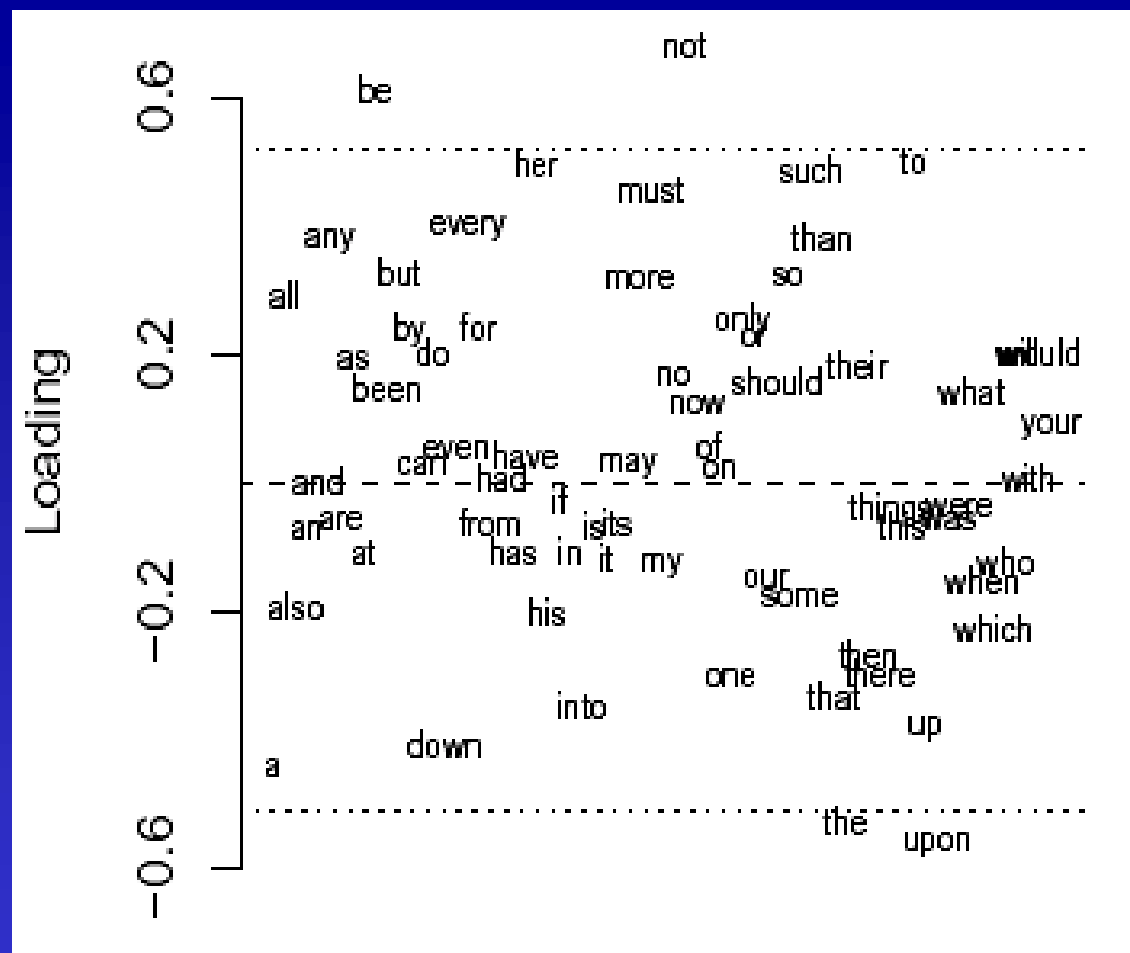
Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

“Loadings”: Correlations between columns of X and columns of XB

$$XB = X \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \beta_1 & \beta_2 & \cdots & \beta_L \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix}$$

1st CV



$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

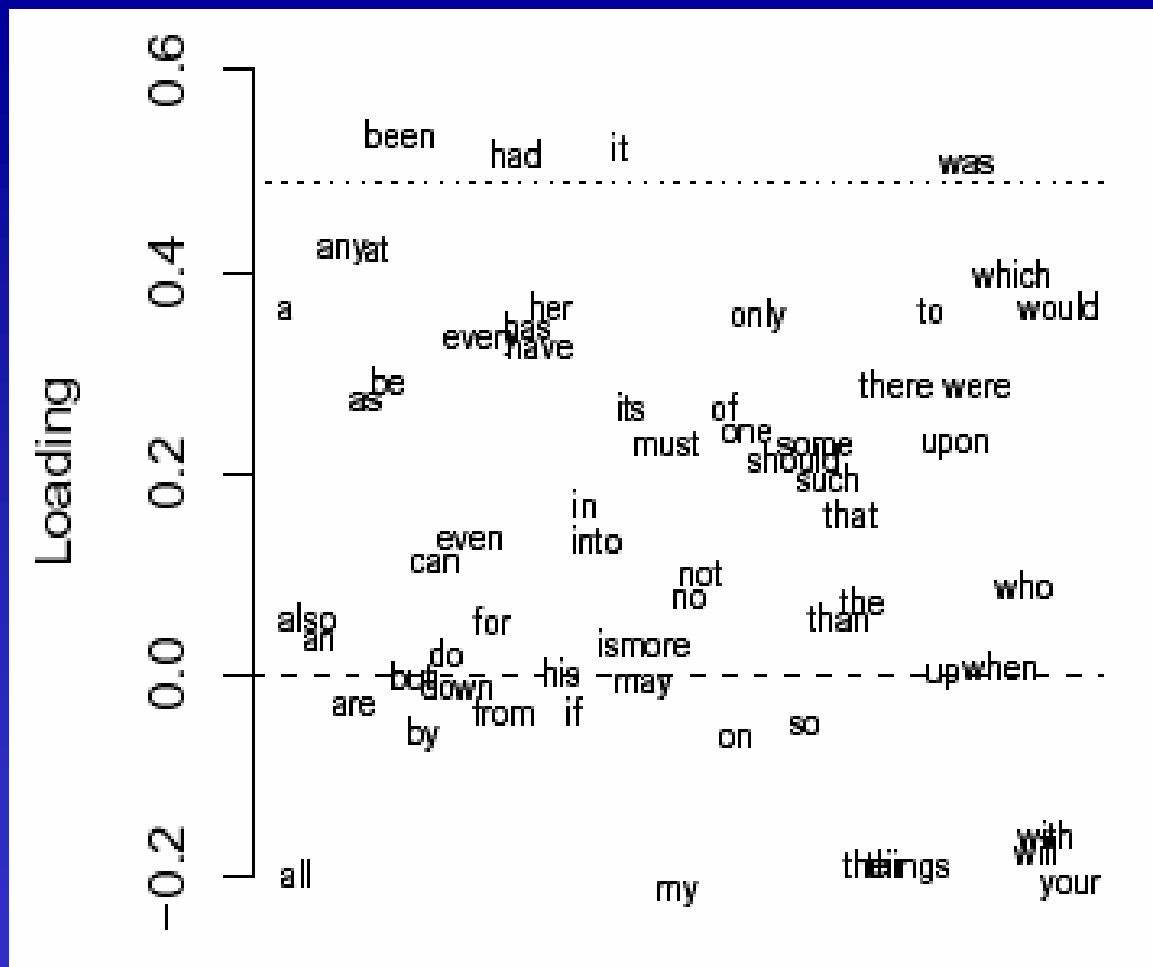
Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

“Loadings”: Correlations between columns of X and columns of XB

$$XB = X \begin{bmatrix} \vdots & \vdots & \vdots \\ \beta_1 & \beta_2 & \cdots & \beta_L \\ \vdots & \vdots & \vdots \end{bmatrix}$$

2nd CV



$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage in Written Text

Lessons Learned from Authorship Work

Identification Results

	Au	Ca	Di	Do	Ki	Lo	Ma	Mi	Sh	Error
Austen	99.1	0.2	0.5	0.2						0.9
Cather		89.7	2.1		2.1	6.0				10.3
Dickens	0.7	2.5	87.1	5.2	1.0	3.2			0.3	12.9
Doyle	0.2	0.2	7.0	90.5	0.9	0.7	0.2		0.4	9.5
Kipling		11.0	0.6		66.2	20.8			1.3	33.8
London	0.3	10.1	5.7	0.7	24.0	58.8	0.3			41.2
Marlowe							72.3		27.7	27.7
Milton							3.6	96.4		3.6
Shakespeare			0.6				9.2		90.2	9.8

$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Idiosyncratic Word/Phrase Usage

Idea: Adapt Text Authorship Approach to Spoken Words

- Primary Limitation of current N-gram approaches:
 - Requires significant amount of data before useful (20-60 min. of speech)
- Candidate Solution:
 - Determine most useful initial set of features (words/phrase-types) for spoken language (see below**)
 - Further reduce dimensionality to reduce amount of required training data:
 - LDA-based approach (ala text authorship identity approach)
 - LSI (Latent Semantic Analysis)-style approach
 - “Eigenwords” approach (eigenvalue decomposition of matrix of LM parameters)
- What are the Best Features for Spoken Language?**
 - Features that represent “*involuntary speaking style*”
 - Fillers/disfluencies (“um”, “uh”, “so”): supported by Weber (2002)
 - Backchannel words/phases (“yep”, “yeah”, “right”, “uh-huh”)
 - Functional words
 - Features that are easy to recognize with ASR
 - Use confidence measures to filter out falsely identified keywords

$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Language Modeling

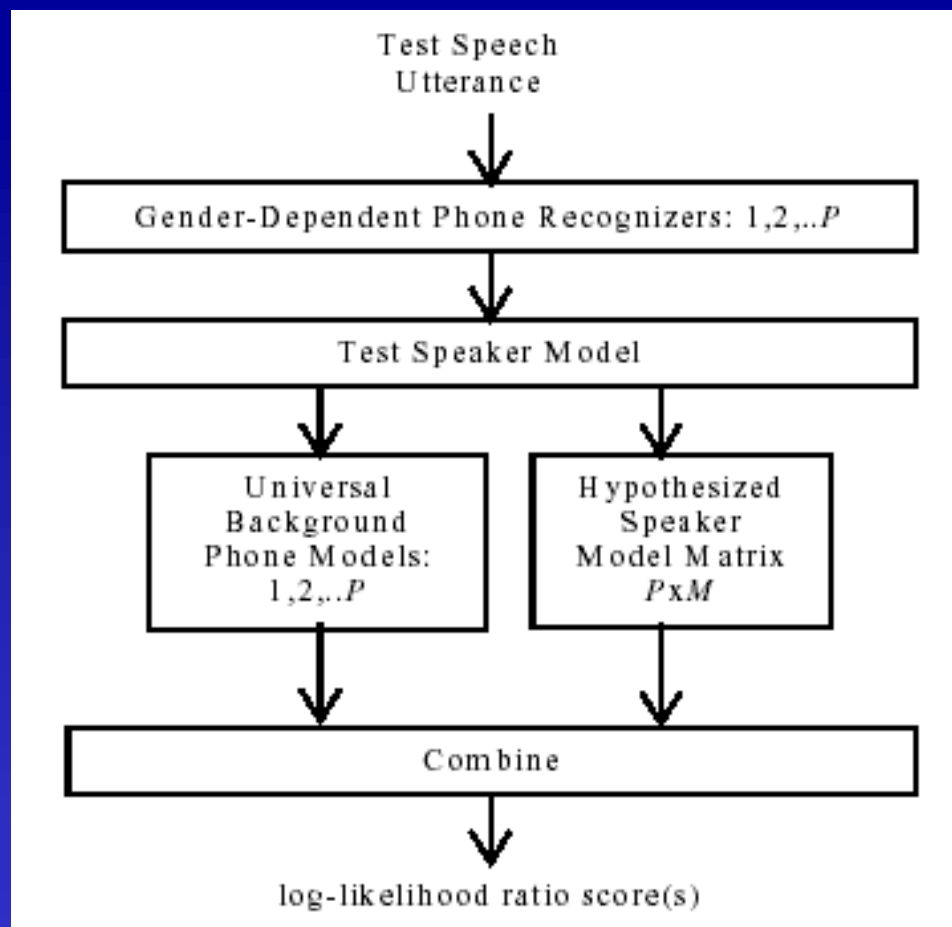
Idiosyncratic Pronunciations

- Motivation: humans recognize speakers by their idiosyncratic pronunciations
- Prior/Current Work
 - Andrews et al. (2001-2002):
 - phonetic “refraction” for speaker recognition.
 - Qin Jin et al. (2002):
 - robustness to near/far-field microphone mismatch with phoneme n-grams.
- Advantages of Phoneme-based Language Modeling
 - Does not require full ASR system (simple phone loop)
 - Easier to train
 - Computationally less expensive than ASR
 - Requires less training data than word/phrase-based LMs

$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Language Modeling Idiosyncratic Pronunciations

- Phone-based language modeling algorithm**
- Phone recognizers & models for 6 languages
- Gender-dependent (~20% rel. improvement)
- LR detector:
 $P(W|S)/P(W)$
- Results:
 - Training: 8 conversations (~20-25 minutes)
 - Testing: 1 conversation (~2-3 minutes)
 - 3.58% EER (significantly outperforms word-based LMs)



**Andrews, Kohler, Campbell, Godfrey, and Jaime Hernandez-Cordero (2001-2002)

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

High-Level Information for Speaker Recognition Outline

- Introduction: Background & Motivations
- Problem Formulation: Component Factorization
- Speaker-Dependent Language Modeling
- Speaker-Dependent Prosodic Modeling
- Summary

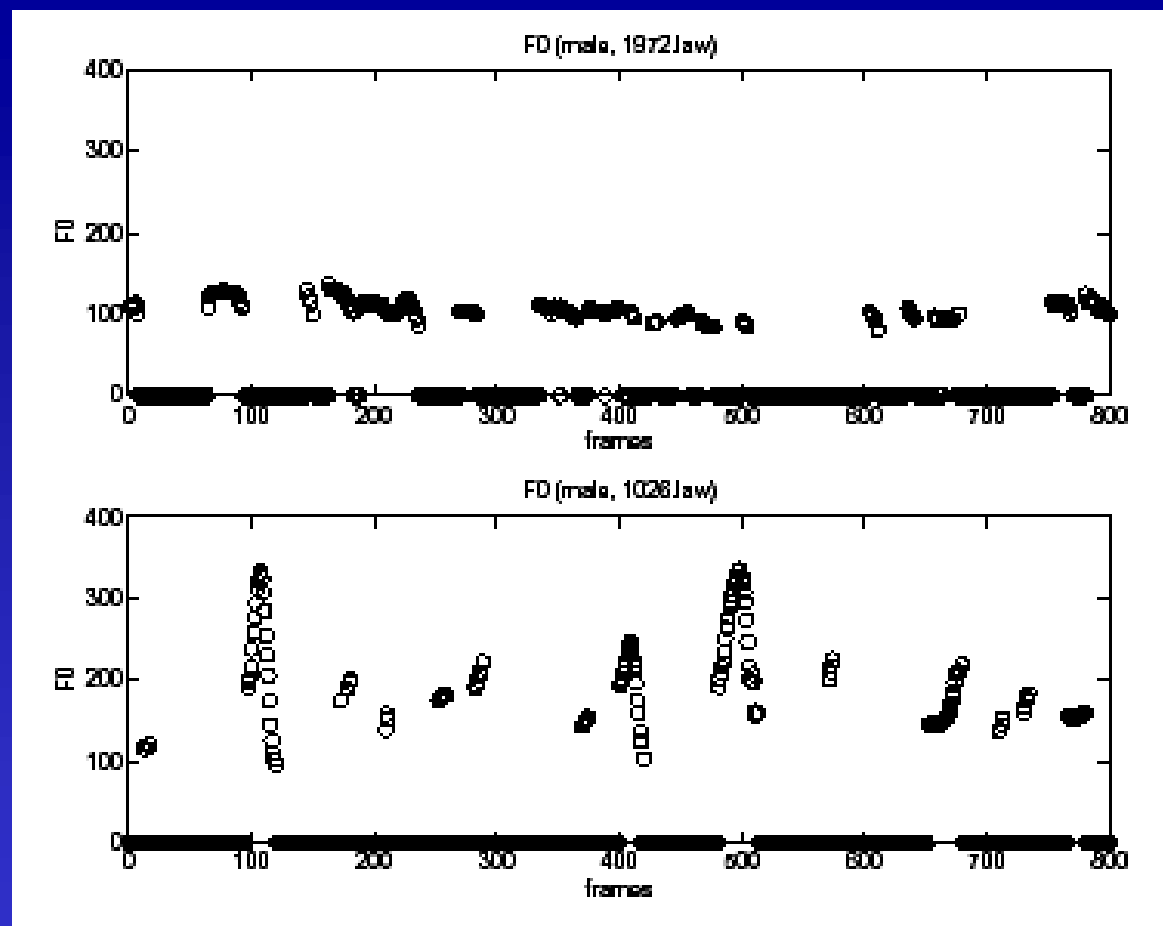


Speaker-Dependent Prosodic Models

Introduction

- Prosodics include:
 - Speaking rate
 - Pause rate
 - Timing
 - Pitch (f0)
 - Melody
 - Rate of change
 - Global declinations

- Prosodics useful for:
 - Conveying meaning
 - Detecting sentence & topic boundaries (Shriberg et al., 2000)
 - Speaker ID??



$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Prosodic Models

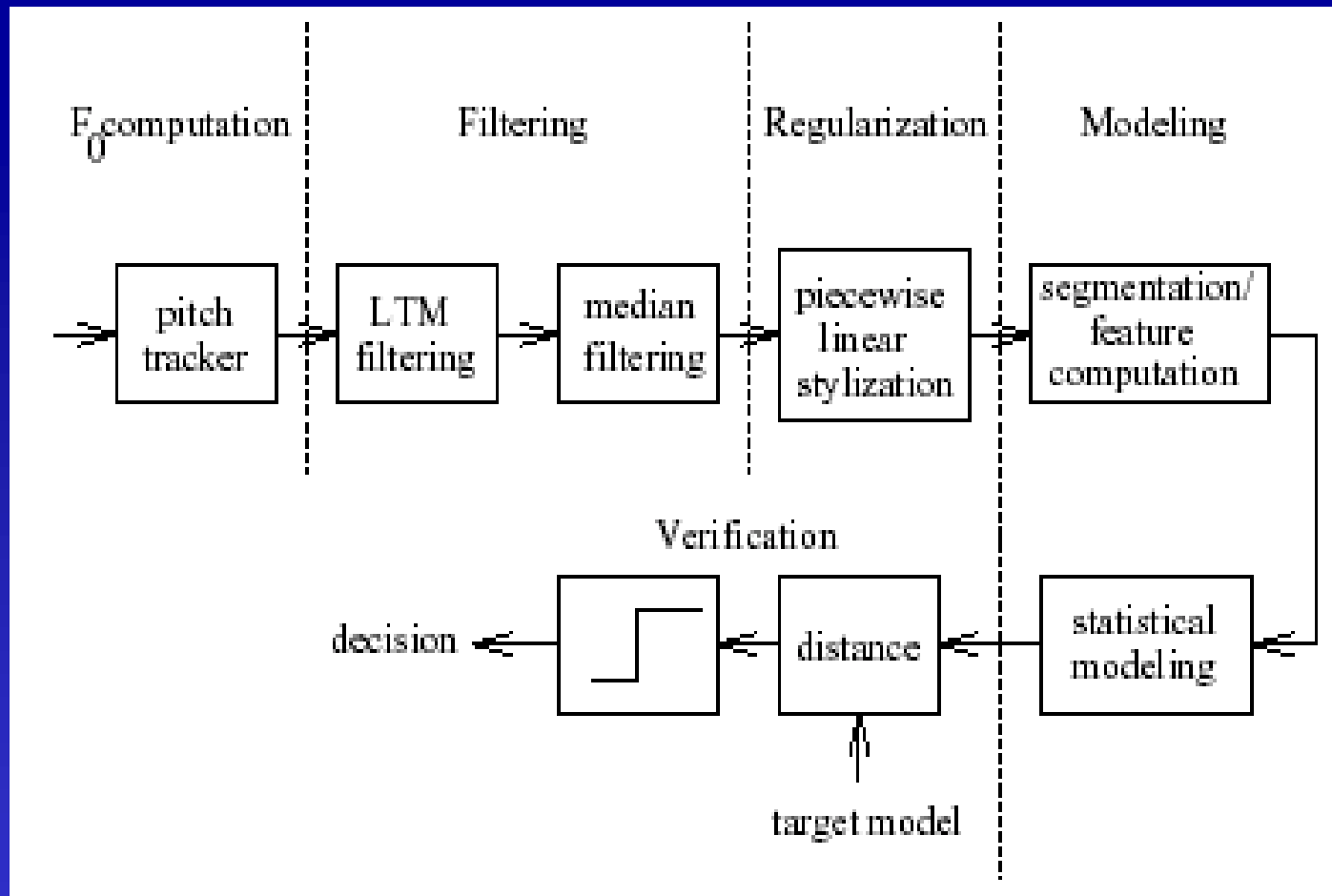
Prior Work

- Most work based on modeling of pitch (f0) statistics
 - B.S. Atal (1972) “Automatic speaker recognition based on pitch contours”
 - M.J. Carey et al. (1996) “Robust prosodic features for speaker identification”
 - K. Sonmez et al. (1997,1998) “Modeling dynamic prosodic variation for speaker verification”
- Early work augmented feature vector with raw f0
- More recent work (e.g., Sonmez) modeled f0 separately and included other factors (pause & voicing durations)

$$\operatorname{argmax}_S P(S|\mathbf{X}, \mathbf{W}, \mathbf{F}, \mathbf{C}) = \operatorname{argmax}_S \underbrace{\frac{P(\mathbf{X}|S, \mathbf{W}, \mathbf{F}, \mathbf{C})}{P(\mathbf{X}|\mathbf{W}, \mathbf{F}, \mathbf{C})}}_{TD-Speaker} \cdot \underbrace{\frac{P(\mathbf{W}|S)}{P(\mathbf{W})}}_{SD-LM} \cdot \underbrace{\frac{P(\mathbf{F}|S, \mathbf{W})}{P(\mathbf{F}|\mathbf{W})}}_{SD-Prosody} \cdot \underbrace{\frac{P(\mathbf{C}|S)}{P(\mathbf{C})}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Prosodic Models

Statistics of Local F0 Dynamics*



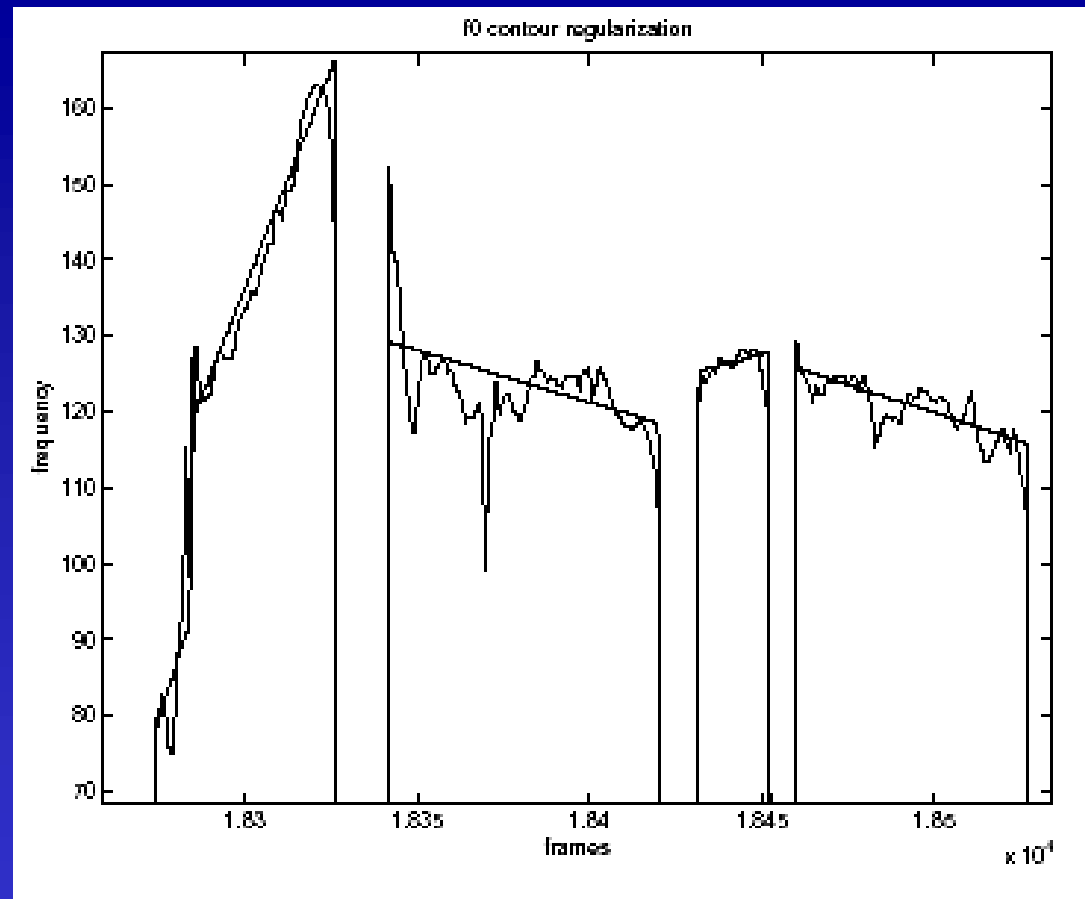
* Sonmez, Shriberg, Heck, & Weintraub (1998)

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Prosodic Models

Statistics of Local F0 Dynamics*

F0 stylization*



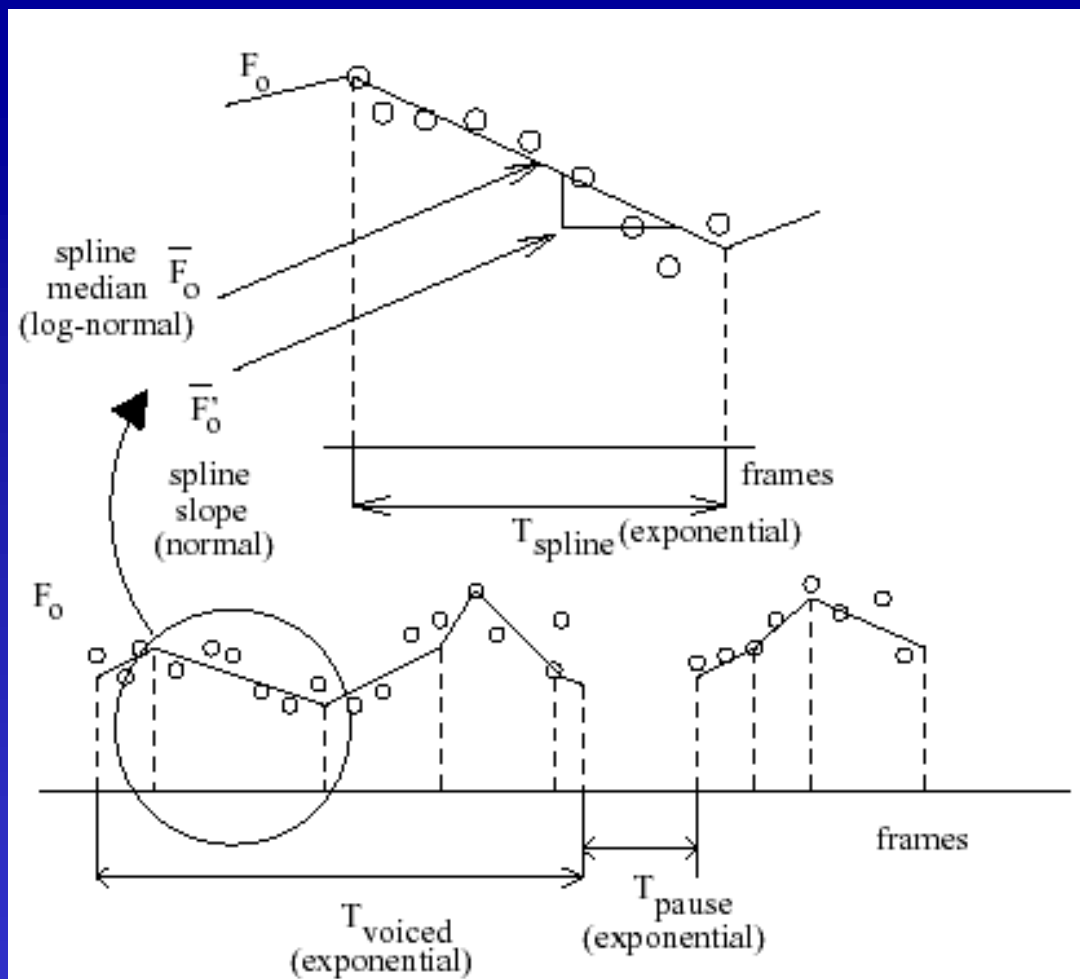
* Sonmez, Shriberg, Heck, & Weintraub (1998)

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Prosodic Models

Statistics of Local F0 Dynamics*

- f0 stylization
- Model (segments):
 - median log(f0) $\sim N(\mu_0, \sigma_0^2)$
 - slope f0' $\sim N(\mu_1, \sigma_1^2)$
 - duration $T_{\text{spline}} \sim \varepsilon(\mu_2)$
 - Voiced duration $T_{\text{voiced}} \sim \varepsilon(\mu_3)$
 - Pause duration $T_{\text{pause}} \sim \varepsilon(\mu_4)$
- Results (NIST 1998)
 - Baseline: MAP-adapted GMM w/ UBM+Hnorm (MFCC-based)
 - Linearly combined scores with baseline system
 - ~10% improvement



* Sonmez, Heck, Shriberg, & Weintraub (1997-1998)

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Prosodic Models

Idea: Dynamic Models of Prosody for “Involuntary” Speech

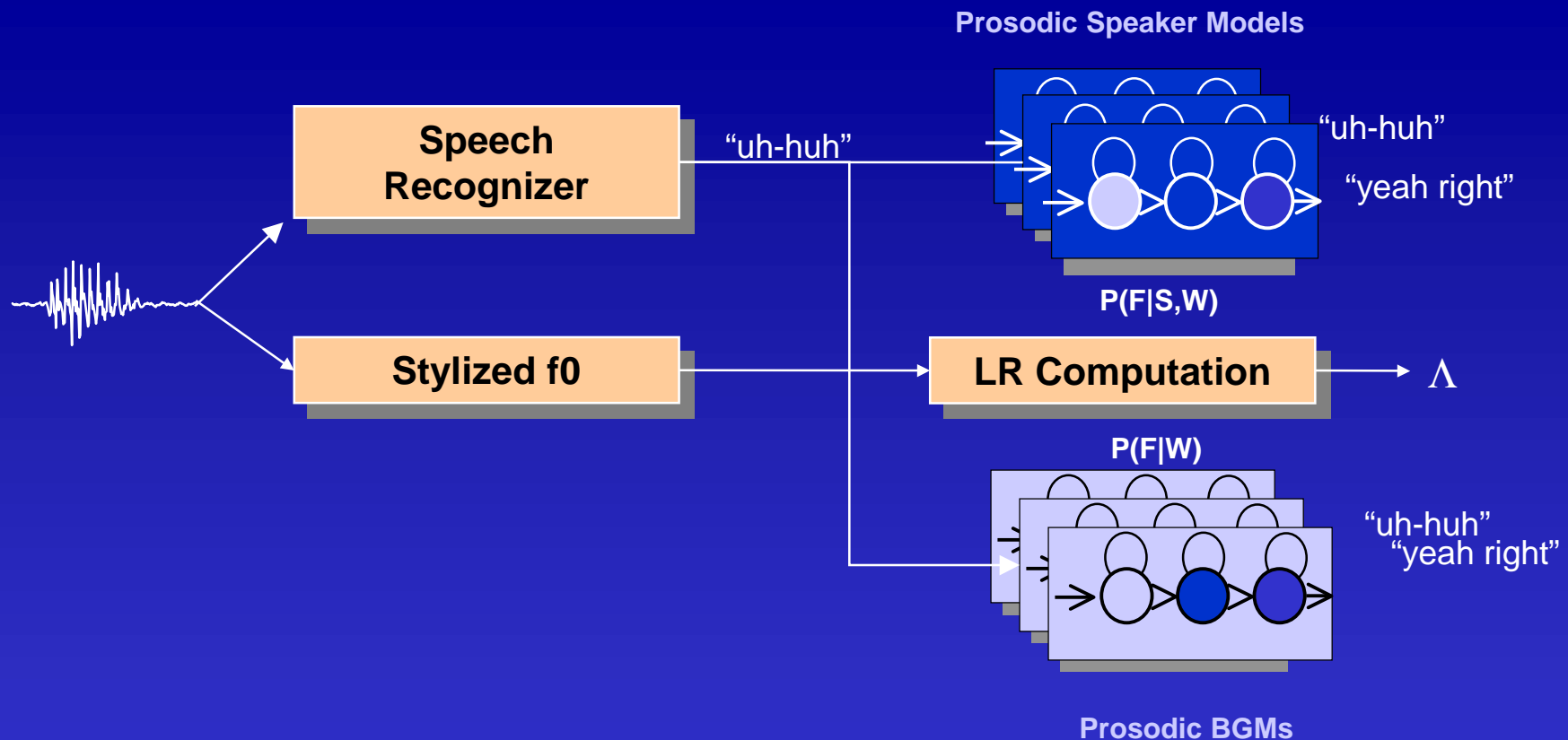
- (With ASR) Compute stylized f0 representations of frequently occurring “involuntary” words/phrases
 - Fillers/disfluencies (“um”, “uh”, “so”)
 - Backchannel words/phases (“yep”, “yeah”, “right”, “uh-huh”)
 - Identity Claims (“my name is Larry Heck”, phone #s, birth date)
 - Frequently visited dialog states (“Voyager”)
- (No ASR) Compute stylized f0 representations for each sentence
 - Find “classes” of canonical f0 profiles (compact w/ broad coverage)
 - Decision trees, various clustering algorithms
- Train/adapt dynamic models (e.g., HMMs) of stylized f0 segments for each word/phrase/sentence
 - Speaker-Independent: large population of speakers: $P(F|W)$
 - Speaker-Dependent: adapt $P(F|W)$ w/ training data for speaker: $P(F|S,W)$
- Implement as likelihood ratio detector for prosodic knowledge source:

$$\frac{P(F|S, W)}{P(F|W)}$$

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

Speaker-Dependent Prosodic Models

Idea: Dynamic Models of Prosody for “Involuntary” Speech



$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \underbrace{\frac{P(X|S, W, F, C)}{P(X|W, F, C)}}_{TD-Speaker} \cdot \underbrace{\frac{P(W|S)}{P(W)}}_{SD-LM} \cdot \underbrace{\frac{P(F|S, W)}{P(F|W)}}_{SD-Prosody} \cdot \underbrace{\frac{P(C|S)}{P(C)}}_{SD-Channel} \cdot P(S)$$

High-Level Information for Speaker Recognition Summary

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \frac{P(X|S, W, F, C)}{P(X|W, F, C)} \cdot \frac{P(W|S)}{P(W)} \cdot \frac{P(F|S, W)}{P(F|W)} \cdot \frac{P(C|S)}{P(C)} \cdot P(S)$$

- Introduced component factorization with explicit dependencies on high-level information
 - TD Speaker Recognizer: current state-of-the-art (MFCC-based)
 - SD Language Models: idiosyncratic word/phrase usage & pronunciations
 - SD Prosodic Models
 - SD “Channel Profile”: Types of phones speaker uses
 - Prior for Speaker: Caller ID, calling patterns
- Presented overview of prior work on language & prosodic modeling
 - SD Language Models:
 - Word/phrase-based LMs improve low-level acoustic system
~50% ERR! IF > 30 mins training data
 - Phonetic-based LMs show significant promise (3.58% EER when used alone!)
 - Prosodic Modeling:
 - Augmenting feature vectors with raw f0 yielded no significant gains
 - Separate system using static modeling methods gave modest gains (~10% rel.).



High-Level Information for Speaker Recognition Summary

$$\operatorname{argmax}_S P(S|X, W, F, C) = \operatorname{argmax}_S \frac{P(X|S, W, F, C)}{P(X|W, F, C)} \cdot \frac{P(W|S)}{P(W)} \cdot \frac{P(F|S, W)}{P(F|W)} \cdot \frac{P(C|S)}{P(C)} \cdot P(S)$$

- Suggested connections between determination of (1) authorship of written text and (2) identity of speaker from spoken utterances:
 - Select features (e.g., words/phrases) that represent “*involuntary speaking style*” for both SD-LM and SD-prosodics
 - Fillers/disfluencies, backchannel, function words
 - Potentially make language models effective with less training data through LDA-based dimensionality reduction
- Suggested dynamic modeling approach for f0 prosodic:
 - HMMs of stylized f0 for fillers/disfluencies, backchannel words
 - LR Detector
- Significant potential in the integration of high-level information!

