Integrating Speech & Speaker Recognizers:
Large Scale Identity Claim Capture for Speaker Verification

Larry Heck
Dominique Genoud

June 22, 2001

Outline
Large-Scale ID Claim Capture

- Introduction
  - Motivations
  - Problem Statement
  - Approach: Integrating Speaker and Speech Recognizers
- Improving Speech Recognition Search
- Computationally Efficient Implementation: 183x speedup
- Experiments
  - 1st+last name ID Claim (1M users): 35% ASR improvement

Outline
Large-Scale ID Claim Capture

- Introduction
  - Motivations
  - Problem Statement
  - Approach: Integrating Speaker and Speech Recognizers
- Improving Speech Recognition Search
- Computationally Efficient Implementation: 183x speedup
- Experiments
  - 1st+last name ID Claim (1M users): 35% ASR improvement

Introduction
Motivations: Large-Scale ID Capture

- Customers want natural, convenient interface for claiming their ID
  - "What is your name?" (More Preferred)
  - "What is your home telephone number?"
  - "What is your SSN?"
  - "What is your account number?"
  - "Please enter your account number" (DTMF) (Less Preferred)
- Recognition of Name-based ID Claims over large populations difficult
  - 60-70% ASR performance for 1 million names
- Need to minimize errors at the system (user) level:

  → ASR error often experienced as ASV false reject

Introduction
Problem Statement & Approach

- Problem Statement
  - How can we recognize ID claims over large populations (millions)
  - How can we disambiguate which user (e.g., “John Smith”)
- Approach:
  - Introduce speaker recognition score into ASR search
  - View as special case of integrating speech & speaker recognizers:

Jointly optimize word sequence (ID claim) and speaker:

$$\underset{W,S}{\operatorname{argmax}} \; P(W,S \mid X)$$

Mathematical Formulation
Large-Scale Identity Capture

- Approximations

$$\underset{W,S}{\text{argmax}}\ P(W,S \mid X) \sim \underset{W,S}{\text{argmax}}\ \underset{Q \in Q_{W,S}}{\text{MAX}}\ P(X|Q)\, P(X|S)^{\alpha}\, P(W)^{\beta}\, P(Q)\ P(S)$$

- Combined Speech and Speaker Recognition Score

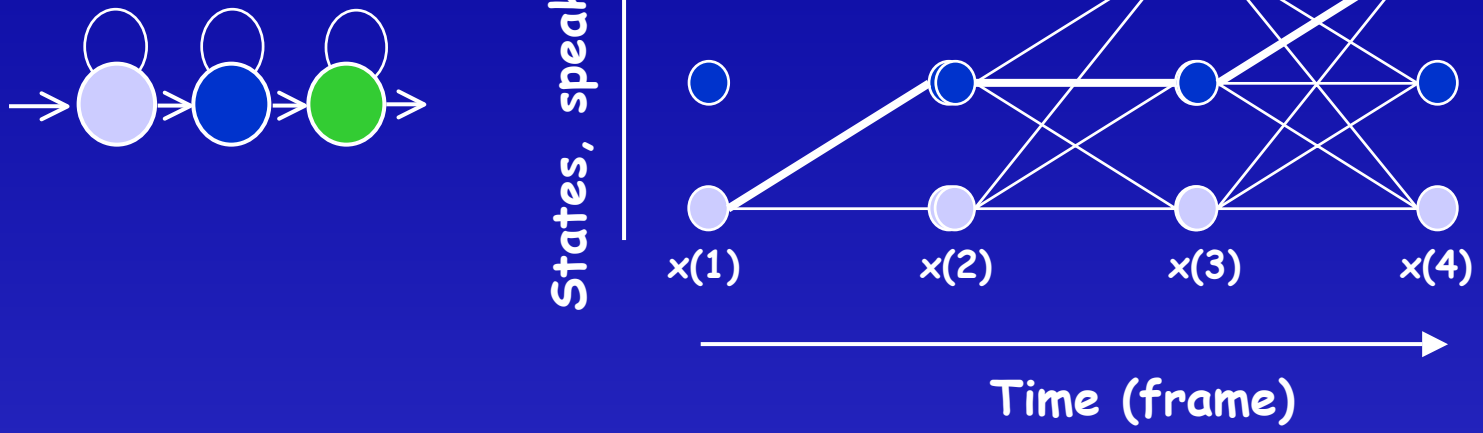
$$S_T = \alpha \log P(X|S) + \log P(X|Q) + \log P(Q) + \beta \log P(W) + \log P(S)$$

Outline
Large-Scale ID Claim Capture

- Introduction
  - Motivations
  - Problem Statement
  - Approach: Integrating Speaker and Speech Recognizers
- Improving Speech Recognition Search
- Computationally Efficient Implementation: 183x speedup
- Experiments
  - 1st+last name ID Claim (1M users): 35% ASR improvement

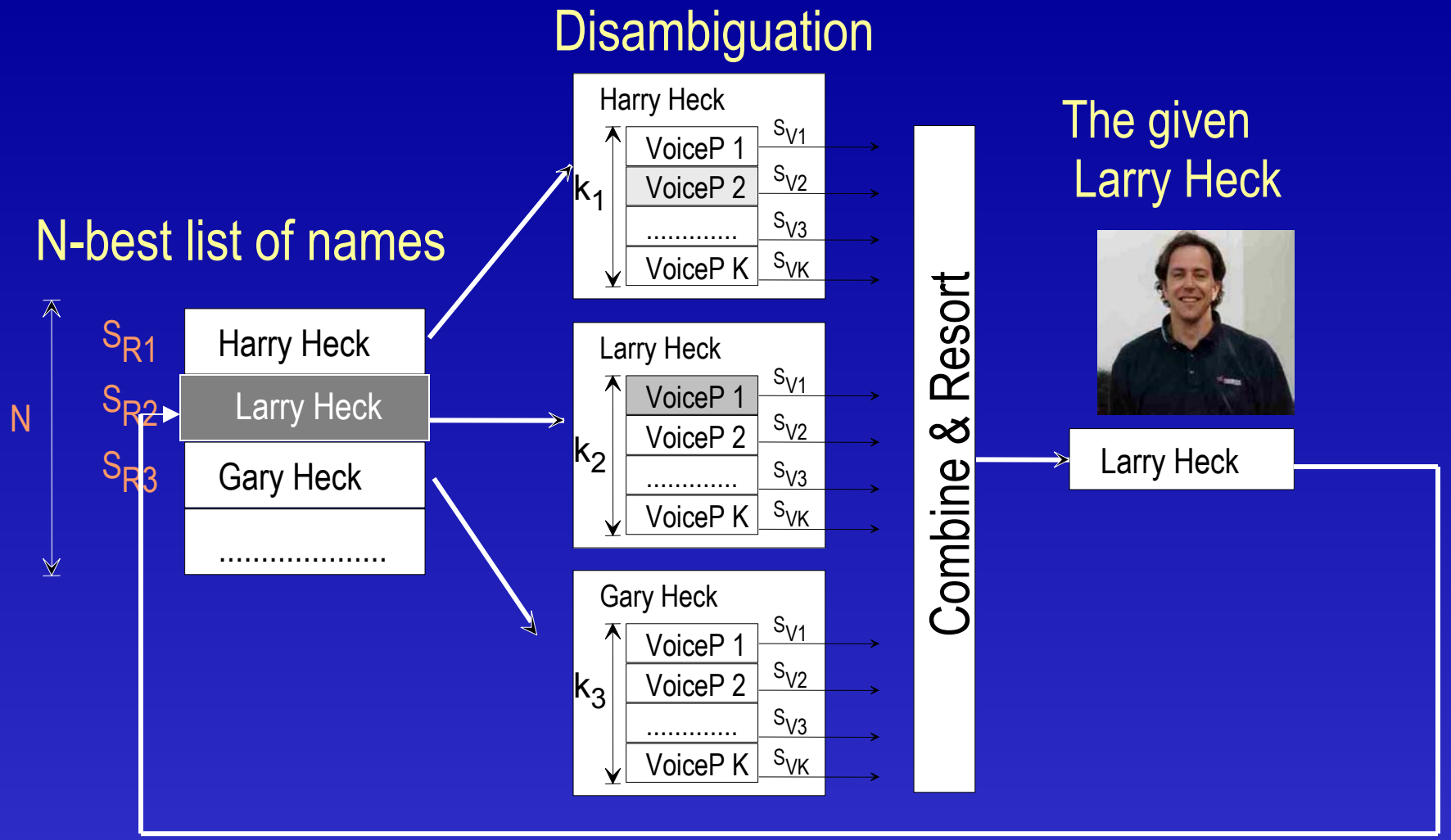
Integration of ASV into Search
1-Pass Approach: ASR + ASV in Viterbi

- In Viterbi search, combine ASR & ASV scores at frame-level



- Problem: large search space (states X speakers)

Integration of ASV into Search
Multi-Pass Approach



Outline
Large-Scale ID Claim Capture

- Introduction
  - Motivations
  - Problem Statement
  - Approach: Integrating Speaker and Speech Recognizers
- Improving Speech Recognition Search
- Computationally Efficient Implementation: 183x speedup
- Experiments
  - 1st+last name ID Claim (1M users): 35% ASR improvement

Computations
Multi-Pass Approach

- Total Computations

  $C_T = C_R + C_V$

- ASV computations for 1 model

  $Cv = 2\ G\ K$

  $= 2 * 2000 * 6 = 24K$

  $Cv^* = G + 2\ K\ M$

  $= 2000 + 2*6*5 = 2.06K$

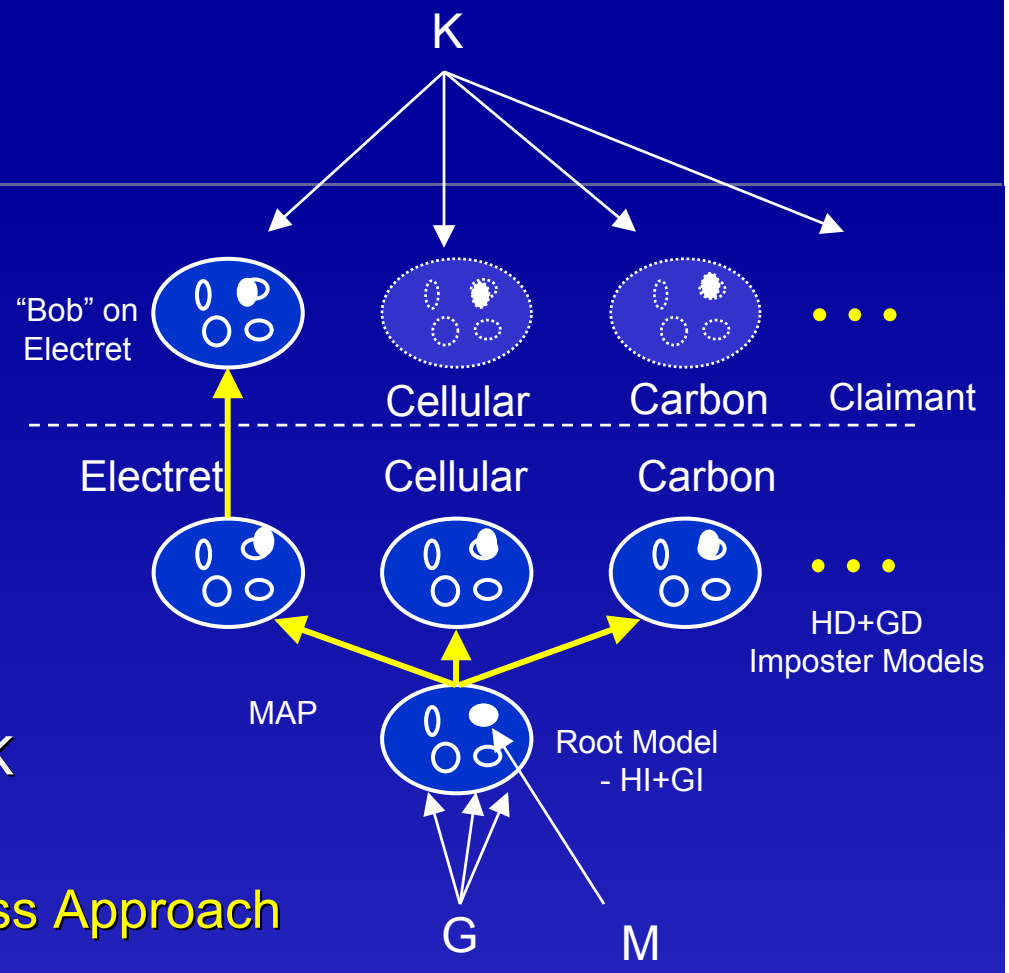
- ASV Computations for Multi-Pass Approach

  $Cv = 2\ G\ K\ N\ A$

  $= 2 * 2000 * 6 * 10 * 2 = 480K$

  $Cv^* = G + (1 + N\ A)\ M\ K$

  $= 2000 + (1 + 10*2) * 5 * 6 = 2.63K$



Outline
Large-Scale ID Claim Capture

- Introduction
  - Motivations
  - Problem Statement
  - Approach: Integrating Speaker and Speech Recognizers
- Improving Speech Recognition Search
- Computationally Efficient Implementation: 183x speedup
- Experiments
  - 1st+last name ID Claim (1M users): 35% ASR improvement

Experiments
Large-Scale Name-based ID Claim Capture

- Goal: simulate ID claim capture over large population with users speaking their names

- Testset:
  - 1000 utterances of personal names (first and last) spoken over long distance telephone lines
  - Callers on mixed handsets (landline, cellular)
  - 500 unique speakers

- Grammar:
  - ~1 Million first+last names from white pages of U.S. city telephone directory

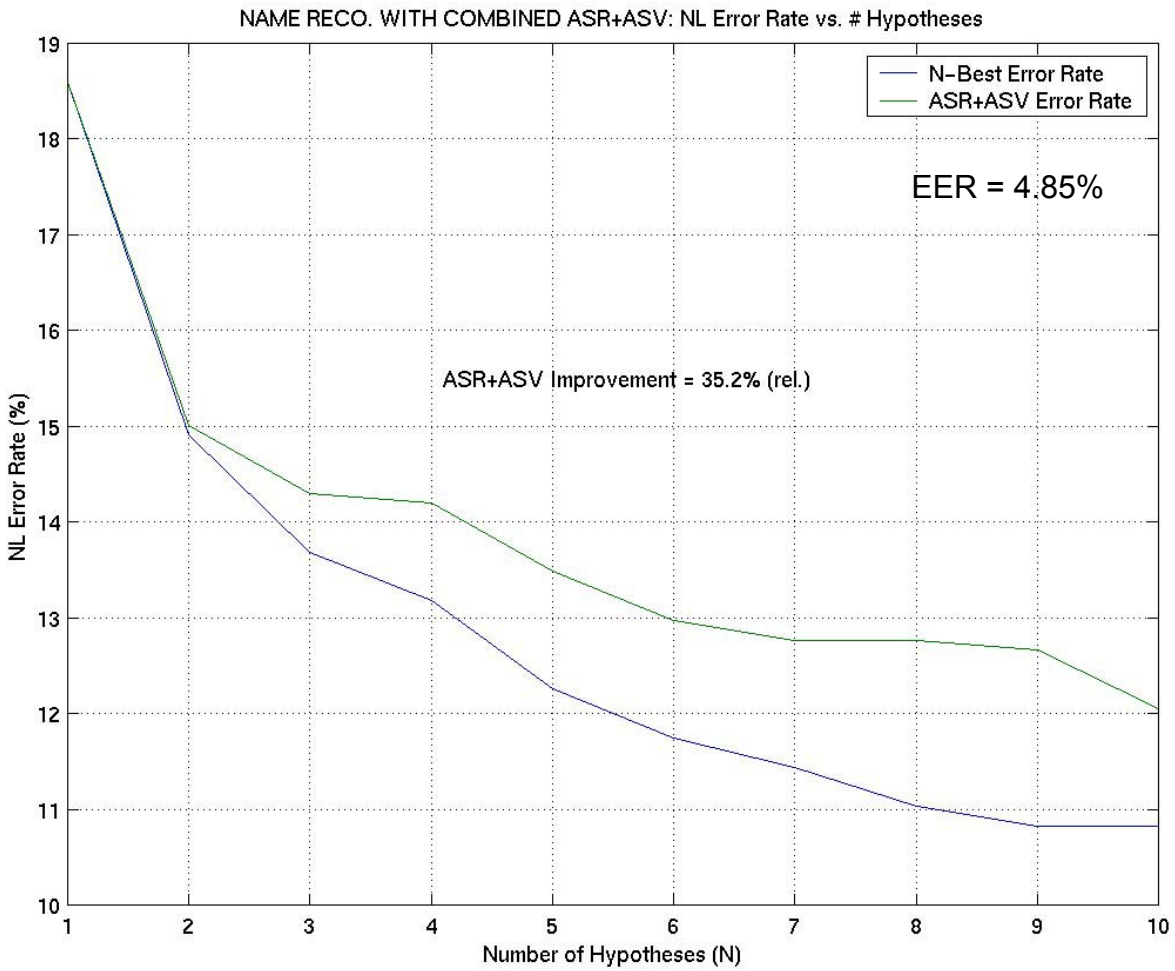
Example
Large-Scale Name-based ID Claim Capture

	N	Hypothesis	ASR Score	ASV Score	Combined Score
Incorrect	1	Chris Graft	0	1.32	428
Correct	2	Chris Craft	-60	4.72	1475
	3	Chris Krauss	-209	-0.86	-490
	4	Chris Kress	-359	1.96	278
	5	Christi Crouse	-461	-1.04	-800
	6	Bruce Graf	-529	-0.61	-727
	7	Craig Kraft	-564	0.18	-507
	8	Chris Groves	-613	-0.98	-930
	9	Christine Craft	-640	-1.33	-1074
	10	Curtis Craft	-651	0.71	-421

Example
Large-Scale Name-based ID Claim Capture

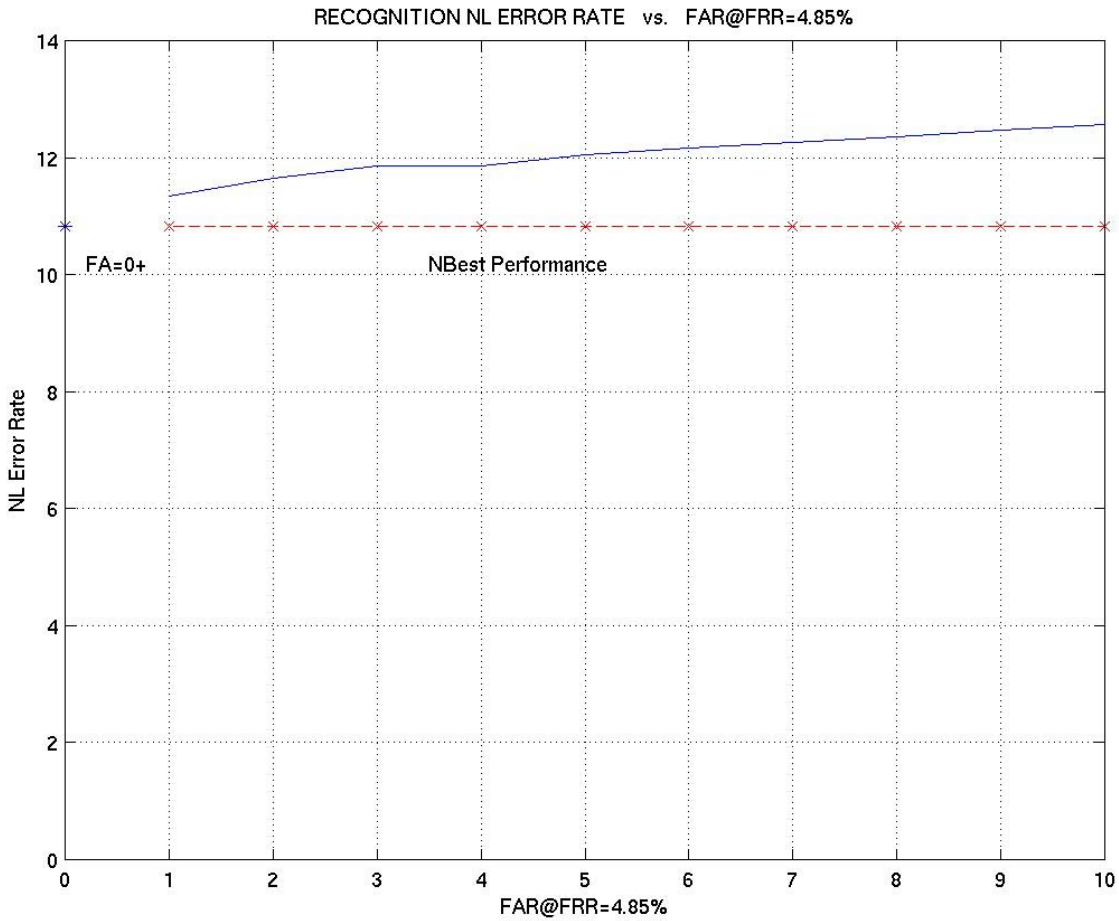
	N	Hypothesis	ASR Score	ASV Score	Combined Score
Correct	2	Chris Graft	-60	4.72	1475
	1	Chris Craft	0	1.32	428
	4	Chris Krauss	-359	1.96	278
	10	Chris Kress	-651	0.71	-421
	3	Christi Crouse	-209	-0.86	-490
	7	Bruce Graf	-564	0.18	-507
	6	Craig Kraft	-529	-0.61	-727
	5	Chris Groves	-461	-1.04	-800
	8	Christine Craft	-613	-0.98	-930
	9	Curtis Craft	-640	-1.33	-1074

Recognition Error Rate vs. # of Hyps
Name-based Identity Claim with 2-pass Approach

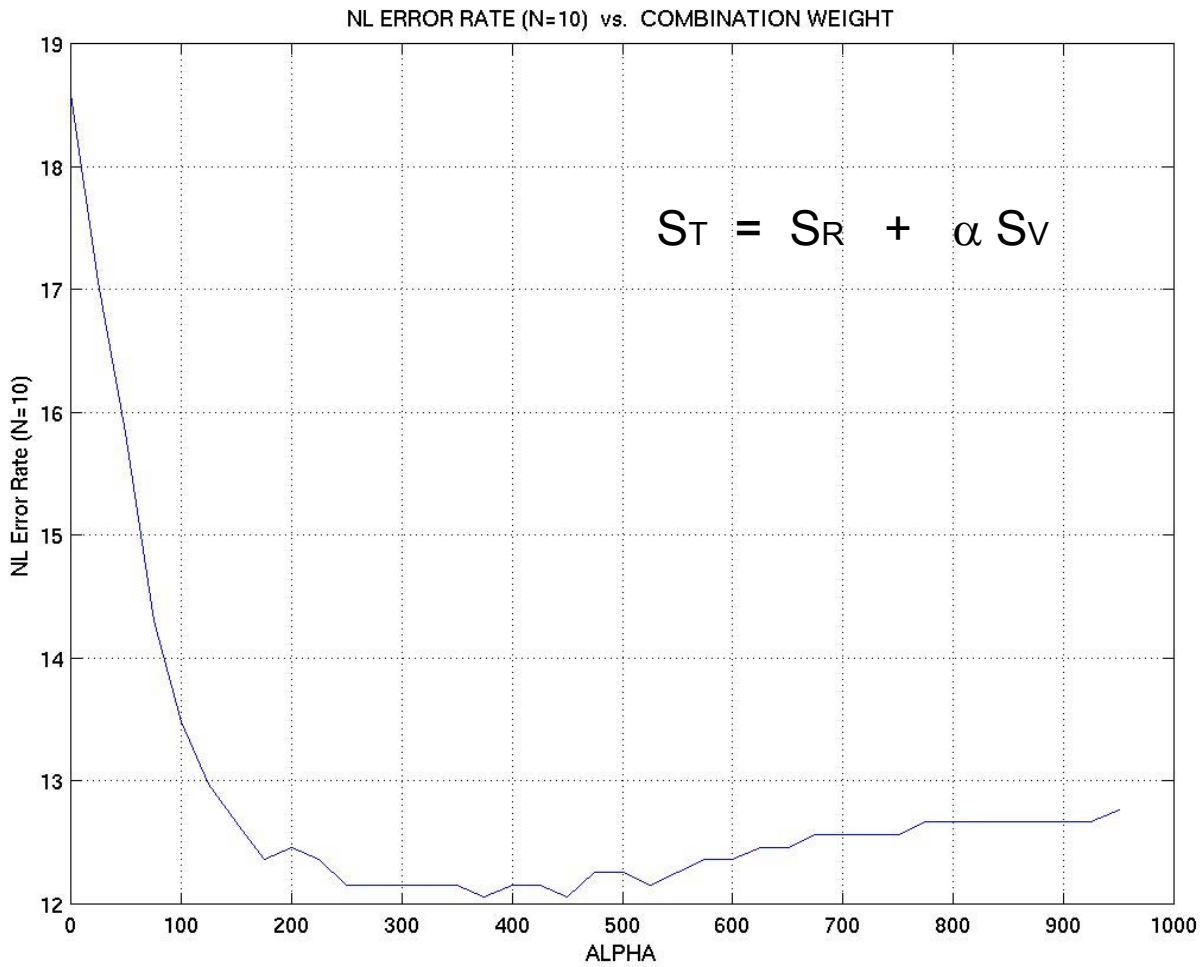


Sensitivity to Verifier Performance
Name-based Identity Claim with 2-pass Approach

Combining ASV with ASR still helps with FAR = 10%



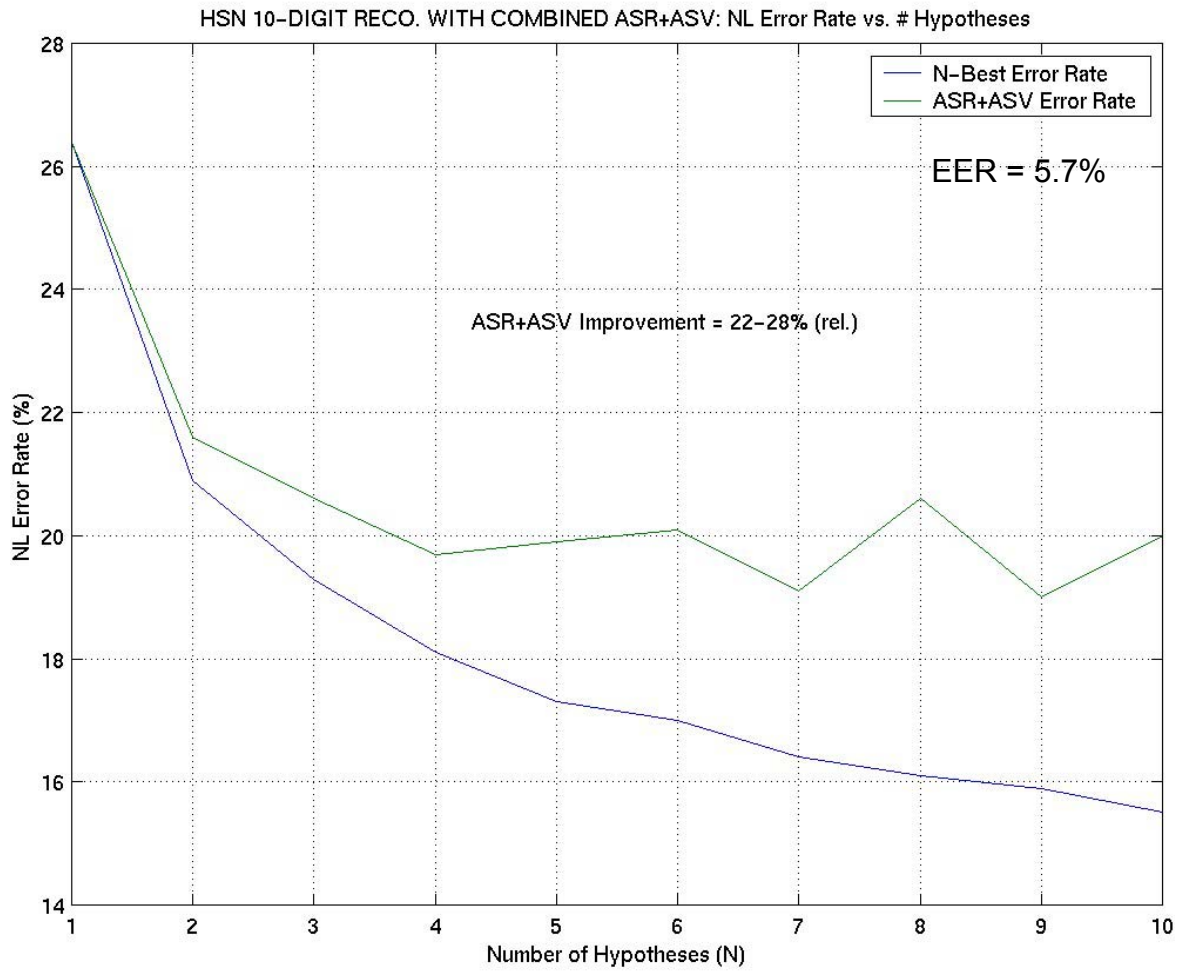
Sensitivity to Combination "Weight"
Name-based Identity Claim with 2-pass Approach



Experiments
Digit-based Identity Claim

- Goal: simulate ID claim capture over large population with users speaking their numbers
- Testset:
  - 1000 utterances of telephone numbers spoken over long distance telephone lines
  - Callers on mixed handsets, but 75% electret/women
  - Very noisy (television, household sounds)
  - 500 unique speakers
- Grammar:
  - Simple (single) digit loop with fixed length of 10 digits

Recognition Error Rate vs. # of Hyps
Digit-based Identity Claim with 2-pass Approach



Summary
Large-Scale Identity Capture Using ASR+ASV

- Introduced integrated speech/speaker recognition approach for ID claim capture
  - Formulated problem as joint optimization P(W,S | X)
  - Derived expressions for efficient search
  - 1-pass versus 2-pass
  - Computationally efficient implementation: 183x speedup
- ID Claim Capture Results
  - Names with 1 Million+ entries: 35% reduction in NL Err.
  - 10-digit ID Claim (telephone number): 22-28% reduction in NL Err
  - Performance is insensitive to ASR+ASV combination "weight"
  - Relative inaccurate speaker recognizer still provides significant ASR improvement (30% reduction in NL Err. @ 10% FAR)