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
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:

\[
\text{argmax } P(W, S \mid X) \\
W, S
\]
Mathematical Formulation
Large-Scale Identity Capture

- **Approximations**

\[
\text{argmax}_{W,S} P(W,S | X) \sim \text{argmax}_{W,S} \max_{Q \in Q_{W,S}} P(X|Q) P(X|S)^\alpha P(W)^\beta P(Q) P(S)
\]

- **Combined Speech and Speaker Recognition Score**

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

- ASV
- ASR
- AM
- Transition Probabilities
- LM
- Prior Knowledge of Speaker
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

Disambiguation

N-best list of names

Harry Heck

Larry Heck

Gary Heck

SV1
SV2
SV3
SVK

VoiceP 1
VoiceP 2
............
VoiceP K

Larry Heck

SR1
SR2
SR3

Combine & Resort

The given
Larry Heck

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
  \[ C_v = 2GK \]
  \[ C_v = 2 \times 2000 \times 6 = 24K \]
  \[ C_v^* = G + 2KM \]
  \[ C_v^* = 2000 + 2 \times 6 \times 5 = 2.06K \]

- ASV Computations for Multi-Pass Approach
  \[ C_v = 2GKN \]
  \[ C_v = 2 \times 2000 \times 6 \times 10 \times 2 = 480K \]
  \[ C_v^* = G + (1 + N \times A)MK \]
  \[ C_v^* = 2000 + (1 + 10 \times 2) \times 5 \times 6 = 2.63K \]
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

<table>
<thead>
<tr>
<th>N</th>
<th>Hypothesis</th>
<th>ASR Score</th>
<th>ASV Score</th>
<th>Combined Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chris Graft</td>
<td>0</td>
<td>1.32</td>
<td>428</td>
</tr>
<tr>
<td>2</td>
<td>Chris Craft</td>
<td>-60</td>
<td>4.72</td>
<td>1475</td>
</tr>
<tr>
<td>3</td>
<td>Chris Krauss</td>
<td>-209</td>
<td>-0.86</td>
<td>-490</td>
</tr>
<tr>
<td>4</td>
<td>Chris Kress</td>
<td>-359</td>
<td>1.96</td>
<td>278</td>
</tr>
<tr>
<td>5</td>
<td>Christi Crouse</td>
<td>-461</td>
<td>-1.04</td>
<td>-800</td>
</tr>
<tr>
<td>6</td>
<td>Bruce Graf</td>
<td>-529</td>
<td>-0.61</td>
<td>-727</td>
</tr>
<tr>
<td>7</td>
<td>Craig Kraft</td>
<td>-564</td>
<td>0.18</td>
<td>-507</td>
</tr>
<tr>
<td>8</td>
<td>Chris Groves</td>
<td>-613</td>
<td>-0.98</td>
<td>-930</td>
</tr>
<tr>
<td>9</td>
<td>Christine Craft</td>
<td>-640</td>
<td>-1.33</td>
<td>-1074</td>
</tr>
<tr>
<td>10</td>
<td>Curtis Craft</td>
<td>-651</td>
<td>0.71</td>
<td>-421</td>
</tr>
</tbody>
</table>
## Example
### Large-Scale Name-based ID Claim Capture

<table>
<thead>
<tr>
<th>N</th>
<th>Hypothesis</th>
<th>ASR Score</th>
<th>ASV Score</th>
<th>Combined Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Chris Graft</td>
<td>-60</td>
<td>4.72</td>
<td>1475</td>
</tr>
<tr>
<td>1</td>
<td>Chris Craft</td>
<td>0</td>
<td>1.32</td>
<td>428</td>
</tr>
<tr>
<td>4</td>
<td>Chris Krauss</td>
<td>-359</td>
<td>1.96</td>
<td>278</td>
</tr>
<tr>
<td>10</td>
<td>Chris Kress</td>
<td>-651</td>
<td>0.71</td>
<td>-421</td>
</tr>
<tr>
<td>3</td>
<td>Christi Crouse</td>
<td>-209</td>
<td>-0.86</td>
<td>-490</td>
</tr>
<tr>
<td>7</td>
<td>Bruce Graf</td>
<td>-564</td>
<td>0.18</td>
<td>-507</td>
</tr>
<tr>
<td>6</td>
<td>Craig Kraft</td>
<td>-529</td>
<td>-0.61</td>
<td>-727</td>
</tr>
<tr>
<td>5</td>
<td>Chris Groves</td>
<td>-461</td>
<td>-1.04</td>
<td>-800</td>
</tr>
<tr>
<td>8</td>
<td>Christine Craft</td>
<td>-613</td>
<td>-0.98</td>
<td>-930</td>
</tr>
<tr>
<td>9</td>
<td>Curtis Craft</td>
<td>-640</td>
<td>-1.33</td>
<td>-1074</td>
</tr>
</tbody>
</table>
Recognition Error Rate vs. # of Hyps
Name-based Identity Claim with 2-pass Approach

EER = 4.85%

ASR+ASV Improvement = 35.2% (rel.)
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

\[ST = SR + \alpha SV\]
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

EER = 5.7%

ASR+ASV Improvement = 22–28% (rel.)
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)