
International Biometrics Consortium
Arlington, VA
23 September 2003

Larry P. Heck, PhD
Vice President, R&D
Nuance Communications
Outline

• Brief introduction to commercial speaker recognition

• Design requirements for commercial verifiers
  – General considerations
  – Voice User Interface ("VUI") design

• Summary
Outline

• Brief introduction to commercial speaker recognition

• Design requirements for commercial verifiers
  – General considerations
  – Voice User Interface (“VUI”) design

• Summary
The Commercial Potential of Voice Biometrics

• Speaker recognition is often referred to as a voice biometric

• Voice biometric is very attractive for companies because:
  – speech is a natural signal to produce
  – does not require a specialized input device
  – ubiquitous: telephones and microphone equipped PC

• Voice biometric can be combined with other forms of security
  – Something you have - e.g., badge
  – Something you know - e.g., password
  – Something you are - e.g., voice

Strongest security

Nuance Communications
The Voice Biometric: Gaining Momentum
Sampling of Key Deployments

1985
Access Control
TI Corporate Facility
(Texas Instruments)

1990
Law Enforcement
Home Incarceration
(ITT Industries)

1995
Telecom
Sprint’s Voice FONCAR
(Texas Instruments)

2000
Telecom
Swisscom
(ScanSoft)

Access
Mac OS9
(Apple)

Financial
Charles Schwab
(Nuance)

Government
UK Offenders
(Nuance)

Financial
Bradesco
(Nuance)

Government
Social Security
(Nuance)

Services
Union Pacific
(ScanSoft)

Insurance
The Hartford
(Nuance)

Law Enforcement
Prison Call Monitoring
(T-Netix)

Small Scale Deployments (100s)

Large-Scale Deployments (1M+)

Nuance Communications
Outline

- Brief introduction to commercial speaker recognition
- Design requirements for commercial verifiers
  - General considerations
  - Voice User Interface (“VUI”) design
- Summary
Design Requirements
General Considerations

- **Fast** (50 simultaneous verifications on a PIII 500MHz)
- **Accurate and Robust** (> 99% correct reject @ > 95% correct accept)
- **Compact Speaker Models** (< 20K/model)
- **Scalable and Fault Tolerant** (1M+ users with standard DBs)
- **Easy to deploy** (doc/tools for security/call-center mgrs & speech dev. , VoiceXML)
- **Internationalized** (support for many languages/regions)
- **Flexible** (various operating modes such as text-prompted/dependent/independent)
- **Fully integrated w/ state-of-the-art automatic speech recognizer (ASR)**
Accuracy Requirement

How Accurate is the Voice Biometric?

<table>
<thead>
<tr>
<th>Condition</th>
<th>Data Source</th>
<th>Microphones</th>
<th>Training Data</th>
<th>Probability of False Reject (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-dependent (Combinations) Clean Data</td>
<td>Telephone Data</td>
<td>Single</td>
<td>Large amount of</td>
<td>0.1%</td>
</tr>
<tr>
<td>Text-dependent (Combinations)</td>
<td></td>
<td>microphone</td>
<td>train/test speech</td>
<td></td>
</tr>
<tr>
<td>Text-dependent (Digit strings) Clean Data</td>
<td>Telephone Data</td>
<td>Multiple</td>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>Text-independent (Conversational) Telephone Data</td>
<td></td>
<td></td>
<td></td>
<td>10%</td>
</tr>
<tr>
<td>Text-independent (Read sentences) Military radio Data</td>
<td></td>
<td>Multiple radios &amp; microphones</td>
<td>Moderate amount of training data</td>
<td>25%</td>
</tr>
</tbody>
</table>
Accuracy Requirement

Effect of Microphone Mismatch

- In the NIST evaluation, performance was measured when speakers used the same and different telephone handset microphone types (carbon-button vs electret).

- With microphone mismatch, equal error rate increases by over a factor of 2.
Accuracy Requirement
Comparison to Other Biometrics

From CESG Biometric Test Programme Report (http://www.cesg.gov.uk/biometrics/)

- Face
- FP-chip
- FP-chip (2)
- FP-optical
- Hand
- Iris
- Vein
- Voice

**Figure 6. Detection error trade-off: Best of 3 attempts**
Design Requirements

General Considerations

- **Fast** (50 simultaneous verifications on a PIII 500MHz)
- **Accurate and Robust** (> 99% correct reject @ > 95% correct accept)
- **Compact Speaker Models** (< 20K/model)
- **Scalable and Fault Tolerant** (1M+ users with standard DBs)
- **Easy to deploy** (doc/tools for security/call-center mgrs & speech dev., VoiceXML)
- **Internationalized** (support for many languages/regions)
- **Flexible** (various operating modes such as text-prompted/dependent/independent)
- **Fully integrated w/ state-of-the-art automatic speech recognizer (ASR)**
Scalable & Fault Tolerant Requirement

Nuance Client-Server Architecture

- Cost-effective operation
- High service availability
- Easy to maintain
- Load balanced per utterance
- Proven high volume deployments
Outline

• Brief introduction to commercial speaker recognition

• Design requirements for commercial verifiers
  – General considerations
  – Voice User Interface ("VUI") design

• Summary
Some consider an unconstrained (text-independent) VUI the “holy grail” for the voice biometric

• **Is this unconstrained interface the best for commercial apps?**
  – For most commercial applications (account access) ➔ No
  – Reason: end-users want to be guided in the interaction

• **What do customers want in their VUI?**
  – End-user makes up something to say each time? No
  – End-user is asked to say a specific phrase? Yes, better
  – End-user has dialog w/ system & some control? Yes, best

**End-users prefer dialog-based VUI**
Design Requirements
Dialog Design: General Principles

• **Dialog should be designed to be secure and convenient**
  – Security often compromised by users if dialog not convenient

    Example: 4-digit PIN
    Security = 1 out of 10,000 false accepts? NO!

    Users compromise security of PINs to make them easier to remember (writing down in wallet, on-line, etc.)

• **Dialog should be maximally constrained but flexible**
  – More constraints → better accuracy for fixed length training

  – Example: balance between constraints on acoustic space while maintaining flexibility → digit sequences

**Dialog Design Goal**  *Constrained but flexible dialog to maximize security while maintaining convenience*
Goal: Maximize Security while Maintaining Convenience

What components are needed to achieve above goal?

- Tightly integrated automatic speech recognizer (ASR)
  - Identity claim capture
  - Simultaneous ID claim and verification
  - Enhanced Security
    - Security against recordings
    - Simultaneous knowledge verification (two-factor security)
  - Advanced error recovery

- Variable Length Verification
Dialog Design
Integrated ASR: Capturing the Identity Claim

• Motivation
  – Many apps involve access to personal accounts and services
    ➔ brokerages & bank accounts, calling cards, insurance & medical records
  – Each app requires the users to ID themselves
  – Most natural VUI ➔ end-users simply speak their name

• Problem
  – ASR of names over large (1M+) populations is difficult
  – How to disambiguate users with common names (e.g., “John Smith”)?

• Approach
  – Utilize voice biometric to improve ASR of names
Dialog Design

Integrated ASR: Capturing the Identity Claim

Disambiguation

Combine & Resort

The correct Larry Heck

ASR output: N-best

Disambiguation

Harry Heck

Larry Heck

Gary Heck

Spkr 1

Spkr 2

............

Spkr K

k1

k2

k3

SV1

SV2

SV3

SVK

Larry Heck

Combine & Resort

Larry Heck

Nuance Communications
**Nuance’s It’s Me™**

- Leverages tight integration between Verifier and ASR engine:
  - Cuts ASR error by 62% on 1M names (7% reprompt-rate)
  - Automatically determines which “John Smith” is talking

**Integrated ASR: Capturing the Identity Claim**

![Bar graph showing 62% less errors with It’s Me™]

- **1M Name Identity Claim**
  - (1000 first X 1000 last)
  - **NL Error, %**
    - **Standard**
    - **It’s Me**

62% less errors
Can we simultaneously get ID claim & verify user?

- Buffer identity claim utterance
- Recognize identity claim and retrieve corresponding model
- “Re-process” data by verifying same utt. against model

**Dialog Design**

**Integrated ASR: ID Claim and Verification**

---

**Start Buffering Data**

My name is John Doe

**Start Verification (‘john_doe.model’)**

My name is John Doe

**Stop Verification & Stop Buffering Data**
Can we use ASR to protect against recordings?

- Prompt user to repeat random phrase
  - Example: “Please say 82-32-71, 82-32-71”
  - Serves as “liveness” test
- Requires modification of enrollment dialog
  - (typically) longer enrollment to adequately cover acoustics
What is “knowledge verification”?  
- Compare spoken response to personal question against known client information

Example ➔ “What is your date of birth?”

Where is knowledge verification (KV) useful?  
- Securing the initial enrollment of speaker model
  - Implicit enrollment (KV for 1st few accesses to collect enough speech)
  - Fall-back in case speaker verifier is unsure or fails
  - Enhanced security (combine KV & voice biometric for 2-factor verification)

First introduced in April ’98 (Heck & Teunen, 1st Nuance User’s Conf.)
Dialog Design

Integrated ASR: Knowledge Verification

You’re accepted by the system

Accept

Reject

Recognize “Who you are”

Speaker Models

Speaker Verification

Combine

Recognize “What you know”

Knowledge Verification

Knowledge Base

Nuance Communications
Methods to Combine Knowledge:

- Sequential ("and" of decisions)
  \[ \text{FAR} = \text{FAR}(sv) \times \text{FAR}(kv) \]
  \[ \text{FRR} = \text{FRR}(kv) + (1-\text{FRR}(kv)) \times \text{FRR}(sv) \]

- Parallel ("or" of decisions)
  \[ \text{FAR} = \text{FAR}(sv) + \text{FAR}(kv) \]
  \[ \text{FRR} = \text{FRR}(sv) \times \text{FRR}(kv) \]

- Weighted Scores
Example: Sequential Combination of Decisions
- Easy to implement, focuses on improving overall security

$$\text{FAR} = \text{FAR}(sv) \times \text{FAR}(kv)$$
$$0.01\% = 0.1\% \times 10\%$$

$$\text{FRR} = \text{FRR}(kv) + (1 - \text{FRR}(kv)) \times \text{FRR}(sv)$$
$$1.1\% = 0.1\% + (1 - 0.1\%) \times 1\%$$
Goal: Maximize Security while Maintaining Convenience

What components are needed to achieve above goal?

• Tightly integrated automatic speech recognizer (ASR)
  ➔ Identity claim capture
  ➔ Simultaneous ID claim and verification
  ➔ Enhanced Security
    - Security against recordings
    - Simultaneous knowledge verification (two-factor security)
  ➔ Advanced error recovery

• Variable Length Verification
**Problem:** “90% of errors come from 10% of end-users!”

Solution: **Use Variable Length Verification (VLV)**

- automatically re-prompt when unsure
- Maintain accuracy while cutting avg. verification time by 1/3
- Reduce overall error rates substantially

*Example:* European high-stakes gambling deployment

- 90% call automation
- no observed errors w/ ~2000 trials
Variable Length Verification:

- “Re-prompt Rate (RPR)” controlled by two thresholds

RPR = Pr(spkr) \((FRR_1 - FRR_2)\) + Pr(imp) \((FAR_2 - FAR_1)\)

Original:
- FAR = 0.4%
- FRR = 1.5%

Nuance Communications
• Convenience is *primary* in design
  – User will only tolerate max of 3 turns
  – Average number of turns needs to be close to 1 turn
  – Users are more tolerant of longer utts per turn

• How do we put all the previous features together?
  1. Try to capture as many of these as possible in 1st turn:
     a) Identity claim
     b) Voice biometric
     c) Knowledge
     d) “Liveness”
  2. Try to capture (c-d) above on each subsequent turn
  3. Use variable length verification to keep avg # turns low
• Presented design requirements for commercial verifiers
  – General considerations
  – Focused on Voice User Interface (“VUI”) design

• Stated that best VUI is *dialog-based*

• Presented advanced features to support VUI design
  – Tightly integrated *automatic speech recognizer (ASR)*
    ➔ Identity claim capture
    ➔ Simultaneous ID claim and verification
    ➔ Enhanced security
      - Security against recordings
      - Simultaneous *knowledge verification* (two-factor security)
    ➔ Advanced error recovery
  – Variable Length Verification

• Provided strategy for combining above features in VUI

Questions?

Larry P. Heck, PhD

www.nuance.com
Applications and Deployments

Deployment Steps

1. Initial Data Collection
2. Tune
3. Limited Deployment
4. Tune
5. Rollout
How do you collect the data?

• “Probing/sampling” approach?
  – Employ persons to call app. under supervision
  – Not widely used (too difficult to collect enough data)

• Assessment from actual in-field data?
  – Much easier to get volumes of data and more realistic
  – **Impostor trials**: common enrollment utterance for impostor trials
  – **True Speaker Trials**: Sort scores. Manually transcribe poor-scoring utts.

• Need ~50 callers/gender
• Need to observe 30 errors of each type/condition (“rule of 30”)
• Each speaker enroll/verifies several times (across multiple channels)
Evaluation metrics

DET Curve

Application operating point depends on relative costs of the two errors

Equal Error Rate (EER) is often quoted as a summary performance measure

Wire Transfer:
False acceptance is very costly
Users may tolerate rejections for security

Toll Fraud:
False rejections alienate customers
Any fraud rejection is beneficial

Equal Error Rate (EER) = 1 %
In addition to EER, a decision cost function (DCF) is also used to measure performance.

$$DCF(\theta) = C(\text{miss})Pr(\text{spkr})Pr(\text{miss} | \theta) + C(\text{fa})Pr(\text{imp})Pr(\text{fa} | \theta)$$

- $C(\text{miss}) = \text{cost of a miss}$
- $Pr(\text{spkr}) = \text{prior probability of true speaker attempt}$
- $C(\text{fa}) = \text{cost of a false alarm}$
- $Pr(\text{imp}) = 1 - Pr(\text{spkr}) = \text{prior probability of impostor attempt}$

For application specific costs and priors, compare systems based on minimum value of DCF.
Evaluation Design
Data Selection Factors

- Performance numbers are only meaningful when evaluation conditions are known.

| Speech quality                  | - Channel and microphone characteristics
|                                | - Ambient noise level and type
|                                | - Variability between enrollment and verification speech
| Speech modality                | - Fixed/prompted/user-selected phrases
|                                | - Free text
| Speech duration                | - Duration and number of sessions of enrollment and verification speech
| Speaker population             | - Size and composition
|                                | - Experience

The evaluation data and design should match the target application domain of interest.
What components can be tuned?

• Operating point (threshold)
  – Setting operating point *a Priori* is very difficult!
  – Speaker-independent and/or speaker-dependent thresholds?
  – Picking correct operating point is key to a successful deployment!

• Dialog Design
  – Customer feedback and/or usage patterns can be used to simplify dialog design (e.g., removing confirmation steps, reducing reprompt rate)

• Impostor Models (Acoustic)
  – Training with real application data results in more competitive impostor models with better representation of linguistics & noise & channels.
Applications and Deployments

Deployment Steps: Limited Deployment/Rollout

What steps are there to deployment?

• Begin with limited set of actual users
  – Representative of entire caller population
  – Representative sampling of (telephone) network
  – Representative of noise and channel mismatch conditions

• After rollout, track the following statistics:
  – Successful enrollment sessions (# of speaker models)
  – Successful verification sessions
  – In-grammar/Out-of-grammar analysis (recognition)
  – Verification rejects (correct & false) for each speaker
  – Duration of sessions
The general area of speaker recognition can be divided into two fundamental tasks:

- **Identification**
- **Verification**

Any work on speaker recognition should identify which task is being addressed.
Terminology
Identification

• Determines whom is talking from set of known voices

• No identity claim from user (one to many mapping)

• Often assumed that unknown voice must come from set of known speakers - referred to as **closed-set** identification
Terminology
Verification/Authentication/Detection

- Determine whether person is who they claim to be
- User makes identity claim (one to one mapping)
- Unknown voice could come from large set of unknown speakers - referred to as **open-set** verification
- Adding “none-of-the-above” option to closed-set identification gives open-set identification

Is this Bob’s voice?

?
Introduction
Examples: Joint Speaker and Speech Recognition

**ID Claim Capture**
Users speak their ID
ID Claim & SV in 1 step

**ASR of Co-channel Speech**
Single Microphone (e.g., handsfree in-car)
ASR of Target Talker Speech Only

**Surveillance & Speaker Tracking**
Background Monitoring of Customer-Agent Interaction
Idiosyncratic Word-Usage/Pronunciations Useful

**SV “Liveness” Test**
Prompted Text Verification
Protects Against Recordings

**Speaker+Knowledge Verification**
“Who you are” + “What you know”
⇒ stronger security
Use ASR for knowledge verification
Problem Formulation
Combining Speaker & Speech Recognizers

- **Objective:** simultaneously minimize the word and speaker error rates

- **Approach:** maximize \( P(W,S|O) \)

- **Bayesian formulation with** \( O = \{X,F,C,...\} \)

\[
\text{argmax}_{\{W,S\}} P(W,S|O) = \frac{P(X|W,C)}{P(X|W,C)} \frac{P(X|F,C,W,S)}{P(X|W,C)} \frac{P(F|C,W,S)}{P(X|W,C)} \frac{P(C|W,S)}{P(X|W,C)} \frac{P(W|S)}{P(X|W,C)} \frac{P(S)}{P(X|W,C)}
\]

- Low-level features (MFCC): \( X = \{x1, x2, \ldots xT\} \)
- Words/phrases/phonemes: \( W = \{W1, W2, \ldots, Wn\} \)
- Prosodic Events: \( F = \{F1, F2, \ldots, Fp\} \)
- Channel information: \( C = \{\text{handheld/handsfree landline & wireless, PC}\} \)
## Example Application: Name-Based ID Claim

### N-Best List Rescoring/Resorting

<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 Application: Name-Based ID Claim

#### N-Best List Rescoring/Resorting

<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 Craft</td>
<td>-60</td>
<td>4.72</td>
<td>1475</td>
</tr>
<tr>
<td>1</td>
<td>Chris Graft</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>
Voice Authentication Application
Examples

Secure access to Speak@Ease Auto Attendant

Secure access to product information

UK Government tracking of youth offenders

Caller authentication for banking

Secure access to online employee data

Secure employee PIN reset

Secure field service automation

Secure access to personalized information