Automatic Speaker Recognition
Recent Progress, Current Applications, and Future Trends

Douglas A. Reynolds, PhD
Senior Member of Technical Staff
M.I.T. Lincoln Laboratory

Larry P. Heck, PhD
Manager, Speaker Verification R&D
Nuance Communications

Presented at the AAAS 2000 Meeting
Humans, Computers and Speech Symposium
19 February 2000

This work was sponsored by the Department of Defense under Air Force contract F19628-95-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.
Outline

• Introduction (Reynolds)
• General theory (Reynolds)
• Performance (Heck)
• Applications (Heck)
• Conclusions and future directions (Heck)
Extracting Information from Speech

**Goal:** Automatically extract information transmitted in speech signal

- Speech Recognition
  - Words: “How are you?”
- Language Recognition
  - Language Name: English
- Speaker Recognition
  - Speaker Name: James Wilson
Introduction

Identification

• Determines who is talking from set of known voices

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

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

• Determine whether person is who he/she claims 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
Speech Modalities

Application dictates different speech modalities:

- **Text-dependent recognition**
  - Recognition system knows text spoken by person
  - Examples: fixed phrase, prompted phrase
  - Used for applications with strong control over user input
  - Knowledge of spoken text can improve system performance

- **Text-independent recognition**
  - Recognition system does not know text spoken by person
  - Examples: User selected phrase, conversational speech
  - Used for applications with less control over user input
  - More flexible system but also more difficult problem
  - Speech recognition can provide knowledge of spoken text
Introduction
Voice as a Biometric

- **Biometric**: a human generated signal or attribute for authenticating a person’s identity

- **Voice is a popular biometric**:
  - natural signal to produce
  - does not require a specialized input device
  - ubiquitous: telephones and microphone equipped PC

- **Voice biometric with other forms of security**
  - Something you have - e.g., badge
  - Something you know - e.g., password
  - Something you are - e.g., voice

Diagram:
- Are
- Have
- Know

Strongest security
Outline

- Introduction
- General theory
- Performance
- Applications
- Conclusions and future directions
General Theory
Components of Speaker Verification System

1. **Input Speech**
   - "My Name is Bob"

2. **Feature Extraction**
   - Bob’s "Voiceprint"

3. **Speaker Model**
   - Bob

4. **Decision**
   - Accept
   - Reject

5. **Impostor Model**
   - Impostor "Voiceprints"

6. **Identity Claim**
Two distinct phases to any speaker verification system

**Enrollment Phase**
- Enrollment speech for each speaker
  - Bob
  - Sally
- Voiceprints (models) for each speaker

**Verification Phase**
- Feature extraction
- Model training
- Verification decision
- Accepted!
- Claimed identity: Sally
Humans use several levels of perceptual cues for speaker recognition.

**Hierarchy of Perceptual Cues**

<table>
<thead>
<tr>
<th>High-level cues (learned traits)</th>
<th>Low-level cues (physical traits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantics, diction, pronunciations, idiosyncrasies</td>
<td>Acoustic aspect of speech, nasal, deep, breathy, rough</td>
</tr>
<tr>
<td>Socio-economic status, education, place of birth</td>
<td>Anatomical structure of vocal apparatus</td>
</tr>
<tr>
<td>Prosodics, rhythm, speed intonation, volume modulation</td>
<td>Personality type, parental influence</td>
</tr>
</tbody>
</table>

- There are no exclusive speaker identity cues
- Low-level acoustic cues most applicable for automatic systems

Easy to automatically extract

Difficult to automatically extract
General Theory
Features for Speaker Recognition

- Desirable attributes of features for an automatic system (Wolf ‘72)
  - Occur naturally and frequently in speech
  - Easily measurable
  - Not change over time or be affected by speaker’s health
  - Not be affected by reasonable background noise nor depend on specific transmission characteristics
  - Not be subject to mimicry

- Practical
- Robust
- Secure

- No feature has all these attributes

- Features derived from spectrum of speech have proven to be the most effective in automatic systems
• Speech production model: source-filter interaction
  – Anatomical structure (vocal tract/glottis) conveyed in speech spectrum

Glottal pulses  Vocal tract  Speech signal

SOURCE SPECTRUM  FILTER FUNCTION  OUTPUT ENERGY SPECTRUM

F\theta = 200 Hz
General Theory

Features for Speaker Recognition

- **Speech is a continuous evolution of the vocal tract**
  - Need to extract time series of spectra
  - Use a sliding window - 20 ms window, 10 ms shift

- **Produces time-frequency evolution of the spectrum**
General Theory
Speaker Models

Bob’s “Voiceprint”

“My Name is Bob”

Identity Claim

Feature extraction

Speaker Model

Impostor Model

Decision

Impostor “Voiceprints”

ACCEPT

REJECT
• Speaker models (voiceprints) represent voice biometric in compact and generalizable form

• Modern speaker verification systems use **Hidden Markov Models (HMMs)**
  
  – HMMs are statistical models of how a speaker produces sounds
  
  – HMMs represent underlying statistical variations in the speech state (e.g., phoneme) and temporal changes of speech between the states.
  
  – Fast training algorithms (EM) exist for HMMs with guaranteed convergence properties.
Form of HMM depends on the application

- **Fixed Phrase** → **Word/phrase models**
  - "Open sesame"

- **Prompted phrases/passwords** → **Phoneme models**
  - /s/ /i/ /x/

- **Text-independent** → **single state HMM**
  - General speech
General Theory
Verification Decision

Bob’s “Voiceprint”

Feature extraction

Speaker Model

Impostor Model

Identity Claim

Impostor “Voiceprints”

"My Name is Bob"

ACCEPT

REJECT
Verification decision approaches have roots in signal detection theory

- **2-class Hypothesis test:**
  - **H0:** the speaker is an impostor
  - **H1:** the speaker is indeed the claimed speaker.

- **Statistic computed on test utterance $S$ as likelihood ratio:**
  \[ \Lambda = \log \frac{\text{Likelihood } S \text{ came from speaker HMM}}{\text{Likelihood } S \text{ did not come from speaker HMM}} \]

\[ \Lambda > \theta \quad \text{accept} \]
\[ \Lambda < \theta \quad \text{reject} \]
Outline

- Introduction
- General theory
- Performance
- Applications
- Conclusions and future directions
Verification Performance
Evaluating Speaker Verification Systems

- There are many factors to consider in evaluating speaker verification systems

| Speech quality                      | - Channel and microphone characteristics  
|                                   | - 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  

The evaluation data and design should match the target application domain of interest
Verification Performance
Evaluating Speaker Verification Systems

Example Performance Curve: Detection Error Tradeoff (DET) Curve

Application operating point depends on relative costs of the two error types:

- **Wire Transfer**: False acceptance is very costly. Users may tolerate rejections for security.
- **Toll Fraud**: False rejections alienate customers. Any fraud rejection is beneficial.

![Example DET Curve Diagram](image-url)

- **High Security**: Equal Error Rate (EER) = 1%
- **High Convenience**: Balance
- **Low Convenience**:
- **NIST** (National Institute of Standards & Technology) conducts annual evaluation of speaker verification technology (since ‘95)
- **Aim:** Provide a common paradigm for comparing technologies
- **Focus:** Conversational telephone speech (text-independent)

**Comparison of technologies on common task**

**Technology Developers**

**Data Provider**

**Evaluation Coordinator**

---

**Nuance Communications**

**MIT Lincoln Laboratory**
Verification Performance
Range of Performance

- **Text-dependent (Combinations)**
  - Clean Data
  - Single microphone
  - Large amount of train/test speech

- **Text-dependent (Digit strings)**
  - Telephone Data
  - Multiple microphones
  - Small amount of training data

- **Text-independent (Conversational)**
  - Telephone Data
  - Multiple microphones
  - Moderate amount of training data

- **Text-independent (Read sentences)**
  - Military radio Data
  - Multiple radios & microphones
  - Moderate amount of training data

Increasing constraints
Verification Performance
Human vs. Machine

• Motivation for comparing human to machine
  – Evaluating speech coders and potential forensic applications

• Schmidt-Nielsen and Crystal used NIST evaluation (DSP Journal, January 2000)
  – Same amount of training data
  – Matched Handset-type tests
  – Mismatched Handset-type tests
  – Used 3-sec conversational utterances from telephone speech
Outline

- Introduction
- General theory
- Performance
- Applications
- Conclusions and future directions
Applications

- Transaction authentication
  - Toll fraud prevention
  - Telephone credit card purchases
  - Telephone brokerage (e.g., stock trading)
Applications

- Access control
  - Physical facilities
  - Computers and data networks
Applications

• Monitoring
  – Remote time and attendance logging
  – Home parole verification
  – Prison telephone usage
Applications

- **Information retrieval**
  - Customer information for call centers
  - Audio indexing (speech skimming device)
Applications

- Forensics
  - Voice sample matching

![Recorded threat and Suspect waveforms](image)
Applications
Speaker + Speech Recognition

- Voice
- Data
- Knowledge

Authenticate Voice
Authenticate Knowledge

Accept
Reject

You’re accepted by the system

Speaker Verification + Speech Recognition + Knowledge Verification

Please enter your account number "5551234"
Say your date of birth "October 13, 1964"
You’re accepted by the system
Applications
First High-Volume Deployment

Benefits
• Security
• Personalization

Application
• Speaker verification and identification based on home phone number
• Provides secure access to customer record & credit card information

Implementation
• Nuance Verifier™
• Edify telephony platform
• Deployed July 1999

Size & Volume
• 250k customers enrolled currently @20K calls/day
• 5 million customers will enroll by Q2 ‘00 @170K calls/day

Nuance Communications
MIT Lincoln Laboratory
• Introduction

• General theory

• Performance

• Applications

• Conclusions and future directions
Conclusions

Speaker recognition is one of the few recognition areas where machines can outperform humans.

Speaker recognition technology is a viable technique currently available for applications.

Speaker recognition can be augmented with other authentication techniques to increase security.
Future Directions

Research will focus on using speaker recognition for more unconstrained, uncontrolled situations

- Audio search and retrieval
- Increasing robustness to channel variability
- Incorporating higher-levels of knowledge into decisions

Speaker recognition technology will become an integral part of speech interfaces

- Personalization of services and devices
- Unobtrusive protection of transactions and information