Speaker Verification – The present and future of voiceprint based security

Prof. Eliathamby Ambikairajah
Head of School of Electrical Engineering & Telecommunications, University of New South Wales, Australia
Outline

• Introduction
• Speaker Verification Applications
• Speaker Verification System
• Performance measure
• NIST Speaker Recognition Evaluation (SRE)
• Discussion
Introduction

“How are you?”

Speech Recognition
“How are you?”

Language Recognition
English

Linguistic

Speaker Recognition
Hsing Ming

Emotion Recognition
Happy

Accent Recognition
Taiwanese

Paralinguistic

• Speech conveys several types of information
  – Linguistic: message and language information
  – Paralinguistic: emotional and physiological characteristics
Speech Recognition

Language Recognition

Speaker Recognition

Emotion Recognition

Accent Recognition

Speaker Identification
determines who is speaking given a set of enrolled speakers

Speaker Verification
determines if the unknown voice is from the claimed speaker

Speaker Diarization
partition an input audio stream into homogeneous segments according to the speaker identity
Speaker Identification
determines who is speaking given a set of enrolled speakers

Speaker Verification
determines if the unknown voice is from the claimed speaker

Speaker Diarization
partition an input audio stream into homogeneous segments according to the speaker identity
Speaker Verification Applications - Biometrics

Access control

Physical facilities

Transaction authentication

Telephone credit card purchases
In automatic speaker verification,
- The front-end converts speech signal into a more convenient representation (typically a set of feature vectors)
- The back-end compares this representation to a model of a speaker to determine how well they match
UBM: represent general, speaker independent model to be compared against a person-specific model when making an accept or reject decision.
Speaker Verification System – Speaker Enrolment

**Step 1**

- **Creating a male UBM**
  - Universal Background Models (UBM)
    - Generic Male
    - Generic Female

**Step 2**

- **Creating male speaker-specific models**
  - Target male speaker data
    - Speaker $x_1$
    - Speaker $x_2$
    - Speaker $x_M$
    - Feature Extraction
    - Model Adaptation
    - Speaker $x_1$ Model
    - Speaker $x_2$ Model
    - Speaker $x_M$ Model
Detailed Speaker Verification System

- Cepstral Mean Subtraction (CMS)
- RelAtive SpecTrAl (RASTA)
- Feature Warping
- Feature Mapping

- Nuisance Attribute Projection (NAP)
- Joint Factor Analysis (JFA)
- i-vectors
- Within Class Covariance Normalisation (WCCN)
- Linear Discriminant Analysis (LDA)
- Probabilistic Linear Discriminant Analysis (PLDA)

- Zero-normalisation (Z-norm)
- Test-normalisation (T-Norm)
Front-end: Feature Extraction

25ms Frame 1
25ms Frame 2
25ms Frame 3
25ms Frame N

Windowing → Feature Extraction → Feature Vector

Feature Vector: C₀, C₁, C₂, ..., Cₙ

Feature Normalisation

Normalised Feature Vector: C₀, C₁, C₂, ..., Cₙ

NORMALISED FEATURES

C₀ distribution

Normalised C₀ distribution
Normalised Feature vectors (Frame 1)
\[ C_0 \ C_1 \ C_2 \ \cdots \ C_n \]
Normalised Feature vectors (Frame 2)
\[ C_0 \ C_1 \ C_2 \ \cdots \ C_n \]
Normalised Feature vectors (Frame P)
\[ C_0 \ C_1 \ C_2 \ \cdots \ C_n \]

Temporal Derivative

Delta Feature vectors (Frame 1)
\[ d_0 \ d_1 \ d_2 \ \cdots \ d_n \]
Delta Feature vectors (Frame 2)
\[ d_0 \ d_1 \ d_2 \ \cdots \ d_n \]
Delta Feature vectors (Frame P)
\[ d_0 \ d_1 \ d_2 \ \cdots \ d_n \]

Temporal Derivative

Acceleration Feature vectors (Frame 1)
\[ a_0 \ a_1 \ a_2 \ \cdots \ a_n \]
Acceleration Feature vectors (Frame 2)
\[ a_0 \ a_1 \ a_2 \ \cdots \ a_n \]
Acceleration Feature vectors (Frame P)
\[ a_0 \ a_1 \ a_2 \ \cdots \ a_n \]

Frame 1 Features: (e.g: 39 dimensions)
\[ C_0 \ C_1 \ C_2 \ \cdots \ C_n \ d_0 \ d_1 \ d_2 \ \cdots \ d_n \ a_0 \ a_1 \ a_2 \ \cdots \ a_n \]
Detailed Speaker Verification System

- Feature Extraction
- Feature Normalisation
- Speaker Modelling
- Model Normalisation
- Classification (Scoring)
- Score Normalisation
- Decision Making

- Cepstral Mean Subtraction (CMS)
- RelAtive SpecTrAl (RASTA)
- Feature Warping
- Feature Mapping

- Nuisance Attribute Projection (NAP)
- Joint Factor Analysis (JFA)
- i-vectors
- Within Class Covariance Normalisation (WCCN)
- Linear Discriminant Analysis (LDA)
- Probabilistic Linear Discriminant Analysis (PLDA)

- Zero-normalisation (Z-norm)
- Test-normalisation (T-Norm)
Speaker Modelling

- Probability density function approximated by 3-component Gaussian mixture models
- Each Gaussian mixture consists of a mean ($\mu$), covariance ($\Sigma$) and weight ($w$)

FEATURE SPACE

MODELLING PROBABILITY DISTRIBUTION

All Weights must sum to 1
Database for creating UBM (example)

- **Training set**
  - 56 male speakers (each speaker consists of 2 minutes of active speech) for creating the UBM

- **Target set**
  - 20 male speakers (each speaker consists of 2 minutes of active speech) for speaker-specific model

- **Test set**
  - 250 male utterances (each speaker has many test utterances) with the known identity
Universal Background Model (UBM) consists of 1024 Gaussian mixtures.

Target speaker model consists of 1024 Gaussian mixtures.

✓ Gaussian mixture consists of a mean ($\mu$), covariance ($\Sigma$) and weight ($w$).
The UBM and each speaker model is a GMM.
Each of them will be represented by a vector of weights, a matrix of means and a matrix of covariances.
Decision Making

Feature Extraction

Universal Background Models
- Generic Male
- Generic Female

Determine level of Match
- Likelihood of Generic Male

Determine level of Match
- Likelihood of John

Score, \( L = \log \frac{\text{Likelihood S came from speaker model}}{\text{Likelihood S did not come from speaker model}} \)

\( L \leq \theta \)

Reject/Accept
Score Normalisation

Different Systems perform speaker verification in parallel

Speech

Speaker Verification System 1

Speaker Verification System 2

Speaker Verification System N
Score Normalisation

May not fall in the same range. i.e., NOT directly comparable

Speech

Speaker Verification System 1 → Score 1

Speaker Verification System 2 → Score 2

Speaker Verification System N → Score N
Score Normalisation

Normalised scores ($L$) are comparable
Fusion

Final score will be a weighted sum of score from each system

\[
\text{Final Score} = w_1 s_1 + w_2 s_2 + \ldots + w_N s_N
\]
Performance measure

• Types of error:
  – Misses: valid identity is rejected
    o Probability of miss: ratio of the number of falsely rejected speaker tests to the total number of correct speaker trials.
  – False alarms: invalid identity is accepted
    o Probability of false alarm: ratio of the number of falsely accepted speaker tests to the total number of impostor trials
Performance measure - Detection error trade-off (DET) curve

Each point on the curve corresponds to a different $\theta$
Performance measure - Detection error trade-off (DET) curve

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

Equal Error Rate (EER) = 1%

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

High Security
- False rejections alienate customers
- Any customization is beneficial

High Convenience
- False acceptance is costly
- Users may tolerate rejections for security

Balance

False Acceptance Rate (in %) vs. Miss Rate (in %)
NIST Speaker Recognition Evaluation (SRE)

• Ongoing text independent speaker recognition evaluations conducted by NIST
  (http://www.itl.nist.gov/iad/mig/tests/spk/)
  – driving force in advancing the state-of-the-art
  – Conditions for different amounts of data
    o 10 sec.
    o 3-5 minutes
    o 8 minutes
      o Separate channel and summed channel conditions
  – English-speakers, non-English speakers, multilingual speakers
NIST SRE Trends

• 1996 – First SRE in current series
• 2000 – AHUMADA Spanish data, first non-English speech
• 2001 – Cellular data, Automatic Speech Recognition (ASR) transcripts provided
• 2005 – Multiple languages with bilingual speakers, room mic recordings, cross-channel trials
• 2008 – Interview data
• 2010 – High and low vocal effort, aging, HASR (Human-Assisted Speaker Recognition) Evaluation
• 2012 – Broad range of test conditions, with added noise and reverberation, target speakers defined beforehand
Basic System

Speech → Feature Extraction → MAP adaptation → Log-likelihood → UBM
Trends

• In 2004’s: Classification

Speech → Feature Extraction → MAP adaptation → Extract Supervectors → SVM Scoring
Trends

• In 2005’s: Channel compensation - NAP

Diagram:

1. Speech
2. Feature Extraction
3. MAP adaptation
4. Extract Supervectors
5. NAP
6. SVM Scoring
Trends

• In 2007’s: Channel compensation - JFA

![Diagram showing the process of speech feature extraction and analysis]

- Feature Extraction
- Baum-Welch statistics estimation
- JFA Hyperparameters ($V$, $U$, $D$)
- Factor analysis
- Speaker Factor Extraction
- WCCN
- SVM
Trends

• In 2009’s: Channel compensation – i-vector

Speech → Feature Extraction → Baum-Welch statistics estimation → Factor analysis → i-vector extraction → WCCN → LDA → Cosine Distance Scoring
Trends

• In 2009’s: Channel compensation – PLDA