Speaker Recognition – Practical Issues (A Tutorial)

by

Homayoon Beigi
Beigi@RecognitionTechnologies.com
http://www.RecognitionTechnologies.com

Recognition Technologies, Inc.
300 Hamilton Avenue
White Plains, NY, U.S.A.

Copyright: Recognition Technologies, Inc.
Instructor's Background

- **Recognition Technologies, Inc.** - *President* - 2003-present
- **Internet Server Connections, Inc.** - *Vice President* - 2001-present
- **Columbia University** – *Adjunct Professor* – 1995-present
  - *Courses*: Signal Recognition, Speech Recognition, and Digital Control
- **IBM T.J. Watson Research Center** – Research Staff Member - 1991-2001
- **Columbia University** – BS, MS & PhD, CTR - 1991

Copyright: Recognition Technologies, Inc.
Some Popular Biometrics

- Deoxyribonucleic Acid (DNA)
- Ear (Image-Based and Audio-Based)
- Face
- Fingerprint and Palm
- Hand and Finger Geometry
- Iris
- Retina
- Thermography
- Vein
- Gait
- Handwriting
- Keystroke
- Speaker
- Multimodal
Advantages of Speaker Recognition over other Biometrics (why should we use speaker recognition?)

- **Finger Print not available** (damaged fingers) – 2% of the population (NIST)
- **Iris damage** – Some of the blind (congenital or due to illness)
- **Public Acceptance** – Image and Finger-Print for Criminals only!
  - The U.S. requirement for taking the photo and finger-print of all tourists – Brazil's response :-)
  - Legacy suggests that criminals are photographed and fingerprinted
- **Hard to justify collection for Image, Finger-Print, Iris, Retinal Recognition** – SR Not as forward
  - Other techniques are used for Recognition only; Telephone speech is multi-purposed
- **Long Distance Identification and Verification** – Telephone, widely available interface
- **Media** – Speaker Tracking and Identification may be used for diarization of large media databases
- **Cellular Telephone and PDA-type device security** – No need for extra hardware
Disadvantages of Speaker Recognition

- **Channel Mismatch** – *Different microphone technologies and transmission quality*
- **Signal Variability**
  - More Data is needed to Cover All Possible Variations
  - **Intra-Class vs. Inter-Class Variations** – *Difficulties (common to most biometrics)*
- **Time Lapse Effects** – *Aging*
- **Background Conditions** – *Ambient Noise and Acoustics*
- **Illnesses** – *Affecting Vocal Tract*
Confusion in Terminology

- **Speaker Recognition** *(Verification, Identification, ...)* -- *Best Choice*
- **Speaker Biometrics**
- **Voice Recognition** *(Verification, Identification, ...)* -- *Clashes with Legacy*
- **Voiceprint Identification**
- **Biometric Speaker Identification**
- **Talker Recognition** *(Identification, Clustering, ...)* -- *Seldom Used*
- **Speech Biometrics** -- *Not Exactly the Same Thing – it includes knowledge of the content*
Manifestations of Speaker Recognition

Simple Branches

- **Speaker Verification** – (*Speaker Authentication*)
- **Speaker Identification** – (*Closed-Set and Open-Set*)
- **Speaker & Event Classification** – (*Gender, Age, Music, Noise, ...*)

Compound Branches

- **Speaker Segmentation** – An extension, including audio segmentation and limited *Speaker Identification* to tag the audio segments
- **Speaker Detection**
- **Speaker Tracking**
Modalities of Speaker Recognition
(Mostly Applicable to Verification)

**Text Dependent** – *Fixed text is spoken (not as attractive as other choices)*

**Text Prompted** – *Prompt is usually chosen randomly or based on some formula*

**User Selected** – *may be treated like a password (user picks the question – not too practical)*

**Speech Biometrics** – *may be used to come up with text prompting – most ideal*

**Text Independent** – *The specific text is not used in the recognition (Language Independent ?)*

**Language Independence** – largely, *but may use different processing for different languages*
Highly Multidisciplinary

- Vocal Characteristics, Phonetics, Linguistics, and Phonology
- Signal Processing – *Includes many subfields*
- Information Theory
- Optimization Theory – *Constrained Nonlinear Optimization*
- Bayesian Statistics and Learning – *MLE, MLLR, MAP, etc.*
- Parameter Estimation and Learning – *HMM, SVM, NN, etc.*
- Artificial Intelligence – *Suboptimal Searches, Decision Trees, etc.*
Speech Generation and Perception

Nervous System

Abstraction Encoding → Language Encoding → Neuro-Muscular Activity → Vocalization

50 bps 200 bps 2000 bps 30-50 kbps

Information Rate

Transmission Channel

Vocal Tract

Inner Ear Excitation → Neural Activity → Language Decoding → Abstraction Decoding

The Ear

Intended Message

From Beigi-2010 – adapted from Rabiner and Juang 1990

Copyright: Recognition Technologies, Inc.
Speech Signal

- Non-Stationary Signal
- Discrete Representation used for Automatic Speaker Recognition

**Sampling Theorem**

- Nyquist Critical Frequency: $f_c$
- Sampling Frequency (Sampling Rate): $f_s$
- Nyquist Rate: $2f_c$

**Whittaker-Kotelnikov-Shannon (WKS) Sampling Theorem**

Requires that $f_s \geq 2f_c$

**Use Ideal Sampler**
Signal Approximations and Error Sources

- Periodic Sampling
  - Aliasing
  - Folding
- Amplitude Quantization
  - Linear Pulse Code Modulation (LPCM) – e.g. 16-bit samples
  - ITU-T G.711
    - UPCM (μ-law PCM)
    - APCM (a-law PCM)
- Truncation Error
- Jitter
Signal Representation (Waveform)
Window ~ 23 ms
(43 Hz Band)

Signal Representation
(Narrowband Spectrogram)
Window ~ 6 ms
(172 Hz Band)

Signal Representation
(Wideband Spectrogram)
Control System View of Speech Production

(from Beigi-2010)
Major Lobes of the Cerebral Cortex

(from Beigi-2010)
Brodman Areas of the Cerebral Cortex

(from Beigi-2010)
Speech Production in the Cerebral Cortex

(from Beigi-2010)
Production

**Broca's area, in the brain, is responsible for speech production**

- It is situated near the motor control areas for the vocal tract and the lips
- It also has communication with the language perception sections

**The Vocal Tract produces a carrier signal plus the information intended by the Broca's area**

- Speaker recognition, in general, is only concerned with the carrier signal
- Speech Recognition cares about the intended message produced by Broca's area
Speech Production Anatomy

(From Gray's Anatomy 1918)
Speech Production Anatomy

(From Gray's Anatomy 1918)
Speech Production Model

(from Flanagan 1972)
Perception

- **Language production and Perception are done mostly in the left hemisphere**
  - True for over 95% of the population and most languages

- **Pitch-based Languages are also processed in the right hemisphere**

- **Speaker Recognition is done mostly in the right (less dominant) hemisphere**
  - Humans use tempo, pitch and other musical discourse to conduct speaker recognition
  - Automated Speaker Recognition algorithms generally do not consider musical discourse

- **Is using prosody a good idea for speaker recognition?**
The Human Ear

(from Gray's Anatomy 1918)
Speech Transmission Between the Ear and the Cerebral Cortex

(from Beigi-2010)
Pitch Mapping between the Auditory Cortex

(from Beigi-2010)
Auditory Cortex

**Primary Auditory Cortex**
- Neurons are Tonotopically arranged
- Similar functionality in both hemispheres

**Secondary Auditory Cortex**
- Left Hemisphere: maps auditory signals to phonetic patterns
- Right Hemisphere: maps auditory signal to harmonic, melodic and rhythmic patterns

**Tertiary Auditory Cortex**
- Left Hemisphere: extraction of lexical semantics from phonetic patterns
- Right Hemisphere: extraction of musical discourse
Speech Centers in the Left Hemisphere

(from Beigi-2010)
Phonetics & Phonology
(Elements)

**Phone** – *Elementary sounds in a language*

**Phoneme** – *Semantically significant sounds of a language*

**Allophone** – *Different phones conveying the same phonemic information*
Phonetics & Phonology
(Speech Production Elements)

- **Initiation** – *Function of the air stream and the direction of air flow*
- **Phonation** – *Function of the acoustic energy generated by vocal folds & Larynx*
- **Articulation** – *Place and degree of stricture and aspect of articulation*
- **Coordination** – *Collaboration of articulatory organs for advanced sounds*
Phonetics & Phonology
(Types of Phonation)

- **Pulmonic Consonants**
  - Glottal
  - Epiglottal
  - Pharyngeal
  - Uvular
  - Velar
  - Palatal
  - Retroflex
  - Alveolar
  - Palato-Alveolar
  - Post Alveolar
  - Dental
  - Labiodental
  - Bilabial

- **Non-Pulmonic Consonants**
  - Clicks
  - Voiced Implosives
  - Ejectives
  - Vowels – periodic
  - Whisper
  - Whistle
Speaker Recognition – practical issues

Phonetics & Phonology (English)

**Phonology** – *The study of phonetics in the framework of a specific language*

**49 Phones used by English speakers**

- 19 Vocoids: 12 Vowels, 6 Diphthongs and one Triphthong
- 4 Approximants: 2 Liquids and 2 Glides
- 21 Standard Consonants
- 4 Syllabic Consonants
- 1 Glottal Stop

But, phonologically, there are only 39 phonemes in English,

- 11 Vowels
- 4 Diphthongs
- 4 Semi-Vowels – 2 Liquids (w, l) and 2 Glides (r, y)
- 3 Nasal Consonants (m, n, ng)
- 6 Voiced (b, d, g) and Unvoiced (p, t, k) Stop Consonants
- 8 Fricatives: Voiced (v, th, z, zh) and Unvoiced (f, θ, s, sh)
- 2 Affricates ʃ and ʒ
- 1 Whisper – h
Prosodic Information in Languages

- **Intonation (Pitch)** – e.g., Mandarin and Cantonese
- **Metric Properties of Speech** – Tempo and Continuity
- **Stress (Relative Loudness)** – Insult or Insult
- **Rhythm** – Mora-based rhythm (Turkish)
- **Whisper** – Amerindian languages such as Comanche
Pitch in Mandarin (Ma)

- Hemp
- Horse
- Mother
- Scold

Copyright: Recognition Technologies, Inc.
Gender and Age

- **Different Fundamental Frequency** – *Male, Female, Child*
- **Formant Locations change**
- **Aging also plays a role in changing the formant location of the same individual**
Gender & Age

Frequency (Hz)

Male
Female
Child
Formants
Gender & Age

Formant 1 (Hz) vs. Formant 2 (Hz) plot showing different vowels and their corresponding formant values.
Age

The same utterance spoken by,

44-Year Old Male

2.5-Year Old Male

Copyright: Recognition Technologies, Inc.
Features

**Linear Predictive Coding Features** – *LPC, Reflection, Log Area Ratio, ...*

**Filter Banks** – *EIH Model, Wavelet Filter Banks, Other Nonuniform Filter Banks, ...*

**Modulation Features** – *AM/FM, Empirical Mode Decomposition (EMD)*

**Mel Frequency Cepstral Coefficients (MFCC)**
- **1963 – Cepstra**: *Bogert, Healy and Tukey*
- **1937 – Melody**: *Stevens, Volkman and Newman*
- **Zero Mean** – *All Coefficients but c0*

**Suprasegmental Features**
Suprasegmental Features

- **Prosodic Features**
  - Pitch
  - Loudness (Sonority)

- **Metrical Features**
  - Stress
  - Rhythm

- **Temporal Features**
  - Co-Articulation
Signal Processing

- **Sampling**

- **Pre-Emphasis** – *Best if included in the sampling step*

- **Bandpass Filtering**

- **Filter Bank, Autocorrelation, FFT, Wavelet Transforms, etc.**

- **Local (Static) Feature Extraction** (*e.g., MFCC, LPCC*)

- **Dynamic Features** (*Delta-Cepstra and Delta-Delta-Cepstra*)
Sampling Process

\[ h(t) \rightarrow \text{Anti-Aliasing} \rightarrow \text{Hi-Pass Filter} \rightarrow \text{Pre-Emphasis} \rightarrow \text{Sampler} \rightarrow h(nT) \]
Signal Processing
(Essence of Audio)

- Pitch
- Mel Scale
- Bark Scale
- Loudness
- Timbre
Pitch (Mel Scale - Full)

(from Beigi-2010)
Pitch (Mel Scale -Partial – 1000Hz 40dB Threshold)

(from Beigi-2010)
Quantization

(from Beigi-2010)
Feature Extraction
(Spectral Analysis – e.g. MFCC)

**Direct Method** – *Moving Average (MA)*

- **Framing**

- **Windowing** – *Hamming, Hann(ing), Welch, Triangular, ...*

- **DFT** – *Spectral Estimation*

- **Frequency Warping** – e.g., *Mel or Bark*

- **Magnitude Warping**

- **Mel Frequency Cepstral Coefficient Computation (MFCC)**

- **Mel Cepstral Dynamics – Delta and Delta-Delta Cepstra**
Feature Extraction
(MFCC)
Feature Extraction
(Linear Predictive Methods – e.g. LPCC)

- **Linear Predictive Method** – *AutoRegressive (AR)*
  - Framing
  - **Windowing** – *Hamming, Hann(ing), Welch, Triangular, ...*
  - **AutoRegressive Estimation of the PSD**
  - **LPC Features** – *e.g., LPC Coefficients, PARCOR, Log Area Ration (LAR), ...*
  - **Frequency Warping** – *e.g., Mel or Bark*
  - **Magnitude Warping**
  - **Linear Predictive Cepstral Coefficient Computation (LPCC)**
  - **Feature Cepstral Dynamics** – *Delta and Delta-Delta Cepstra*

*(from Beigi-2010)*
Feature Extraction
(Perceptual Linear Predictive Method – PLP)

(from Beigi-2010)
Feature Extraction
(Wavelet Filterbanks)

- Mel-Frequency Discrete Wavelet Coefficients (MFDWC)
  - Use DWT instead of DCT to reduce the effect of frequency band spill-over
  - Process Sub-bands Separately
    - Relax the assumption that each frame contains only one phoneme

- Wavelet Octave Coefficients of Residues (WOCOR)
  - Hi-Pass Filter
  - Pre-Emphasis
  - Pitch Extraction
  - AutoRegressive Residue Computation
  - Compute Pitch-Sync Wavelet Coeffs
  - Expand Residual Signal
  - Subdivide Wavelet Coeffs
  - Compute WOCOR
Feature Extraction
(Other Features)

- Modulation Features
  - Amplitude Modulation
  - Frequency Modulation
  - Amplitude-Frequency Modulation
  - FEPSTRUM
  - Mel-Cepstrum Modulation Spectrum (MCMS)
- Empirical Mode Decomposition (EMD)
Training and Enrollment

- **Clustering** – Kmeans, LBG, KH, etc.

- **Training** – Large Database

- **Expectation Maximization**

- **Enrollment** – MAP, MLLR, ...

\[
p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}
\]

where \[
\begin{align*}
x, \mu & \in \mathbb{R}^d \\
\Sigma & : \mathbb{R}^d \rightarrow \mathbb{R}^d
\end{align*}
\]
Parameter Estimation and Learning

**Unsupervised Clustering** – *K*-means, VQ, Modified *K*-means, LBG, KH, etc.

**Mean and Variance Estimation** – Unbiased Estimation, MLE, MLLR, ME, MAP.

**Data Mapping and Projection** – PCA, LDA, NLCA, IMELDA, Factor Analysis

**Hierarchical Clustering** – Agglomerative, Divisive
Modeling

Hidden Markov Models
- Single-State Degenerate Model – GMM
- Multi-State Ergodic Continuous Mixture Model

Support Vector Machines – Determinations of a Kernel Function, Dimensionality

Neural Networks
- Feedforward – Auto Associative Neural Nets (AANN)
- Time Delay Neural Nets – TDNN
- Probabilistic Random Access Memory (pRAM)
- Hierarchical Mixtures of Experts (HME) – Uses EM Algorithm
Hidden Markov Modeling

- Memoryless (Zero-Memory) Models – Single-State Degenerate Model (GMM)
- Sources: Ergodicity, Unifilar and Non-Unifilar
- Model Design and States
- Trellis Diagram Representation
- Baum-Welch and Forward-Backward Algorithms
- Viterbi (Forward) Algorithm
Support Vector Machine Modeling

- Decision Functions
  - Kernels
  - Direct Decision Functions
  - Indirect Decision Functions

- 2-Class Problem
  - Linearly Separable
  - Linearly Inseparable

- N-Class Problem
  - Unclassifiable Regions
  - Batch Systems
  - Decision Trees
  - Fuzzification

- Variations
  - Least Squares
  - Linear Programming
  - Robust

- Training and Optimization
  - Expedited Training
    - Problem Decomposition
    - Data Preselection
Neural Network Modeling

- The Perceptron
- Feedforward Networks
  - Auto Associative Neural Networks (AANN)
  - Training (Learning)
  - Global Solution – Simulated Annealing
- Recurrent Neural Networks (RNN)
  - Long Short-Term Memory (LSTM)
- Time-Delay Neural Networks (TDNN)
- Hierarchical Mixtures of Experts (HME)

Parameter Estimation
- Practical Issues
  - Over-Training
  - Local Traps

Copyright: Recognition Technologies, Inc.
beigi@RecoTechnologies.com
Speaker Modeling

- Individual Speaker Models
- Competing Models
  - Background Models
  - Cohorts
- Data Pooling – Speaker-Independent Models
- Speaker Adaptation
Speaker Segmentation

- **Bayesian Information Criterion (BIC)**
- **Model-Based Techniques**
- **General Likelihood Ratio (GLR)**
Speaker Verification

- Text-Dependent
- Text-Prompted
- Text-Independent
- Knowledge-Based
Speaker Identification

- Closed-Set Identification
- Open-Set Identification
- Large-Scale Identification
Other Speaker Recognition

- Speaker Classification
  - Gender Classification (Identification)
  - Age Classification (Identification)
  - Event Classification (Identification)
- Speaker Detection
- Speaker Tracking
- Farfield Speaker Recognition
- Whispering Speaker Recognition
Specialized Traits and Methods

- **Distance Computation** – Log Likelihood Ratio, ...
- **Identification vs. Verification** (1 to N and 1 to 1 Comparison)
- **Classification** May Use Different Types of Features
- **Segmentation** – BIC, GLR, ...
- **Detection** – Segmentation, Identification/Verification
- **Tracking** – Segmentation, Verification
Verification

- **Test Sequence**: $X$
- **Competing Model(s)**: $S_1$
- **Target Model**: $S_D$
- **Likelihood Ratio (LR)**: $p(X \mid S_D) / p(X \mid S_I)$
- **Log Likelihood Ratio (LLR)**: \[ \lambda = \ln(p(X \mid S_D)) - \ln(p(X \mid S_I)) \]
- **Log Likelihood Ratio (LLR) Threshold**: $\lambda_o$ would be the operating threshold
- **Equal Error Rate (EER)**
Identification

- **Test Sequence**  \( X \)
- **Competing Model(s)**  \( S_i \)
- **Target Models**  \( S_t \)
- **Likelihoods**  \( p(X \mid S_t) \) and \( p(X \mid S_i) \)

*Sorted list of Identities based on the likelihood values*
Receiver Operating Characteristic (ROC)
Curve

Equal Error Rate (EER)
Detection Error Trade-off (DET) Curve

Equal Error Rate (EER)
Sample Identification Results

Total Number of Speakers: 75
Rank 0 is associated with Unknown
Rank 1 is a correct Identification
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption
- Spoofing
- Data Quality
- Time-Lapse Effects and Aging
- Large-Scale Recognition
- Standards
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption
- Spoofing
- Data Quality
- Time-Lapse Effects and Aging
- Large-Scale Recognition
- Standards
Channel Mismatch
Sample Channel Types

- Land-Line (PSTN) Carbon Button Microphone
- Land-Line (PSTN) Electret Microphone
- Cellular CDMA
- Cellular GSM
- VoIP 64kbps
- VoIP 128kbps
Channel Mismatch -- Handling
Spectral Filtering and Cepstral Liftering

- Cepstral Mean Subtraction (CMS) – aka, Cepstral Mean Normalization (CMN)
- Cepstral Mean and Variance Normalization (CMVN)
- Histogram Equalization (HEQ)
- Cepstral Histogram Normalization (CHN)
- Auto Regressive Moving Average (ARMA)
- RelAtive SpecTrAl (RASTA) Filtering
- J-RASTA
- Kalman Filtering
Speaker Recognition – practical issues

Channel Mismatch
CMS (CMN)

Copyright: Recognition Technologies, Inc.

Aug 6, 2010
Channel Mismatch Handling
Other Techniques

- Vocal Tract Length Normalization (VTLN)
- Feature Warping
- Speaker Model Synthesis (SMS)
- Feature Mapping
- Score Normalization
  - Z-Norm
  - T-Norm
  - H-Norm
  - HT-Norm
  - AT-Norm
  - C-Norm
  - D-Norm
Channel Mismatch -- Capacity
Band Limitation – Telephony (Landline)

22kHz Sampling Rate

8kHz Sampling Rate

(from Beigi-2010)
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption

- Spoofing
- Data Quality
- Time-Lapse Effects and Aging
- Large-Scale Recognition

- Standards
Numerical Stability

- Use of logarithms
- Logadd tables for speed
- Integer Arithmetic is not prevalent any longer
- Mel and Bark Scale
- Matrix Inversion avoidance
- Scaling
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption
- Spoofing
- Data Quality
- Time-Lapse Effects and Aging
- Large-Scale Recognition
- Standards
Privacy

- Data Encryption – Storage
- Data Encryption – in Transmission
- Audio should not be recoverable
- Encrypted ID
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption
- Spoofing
- Data Quality
- Time-Lapse Effects and Aging
- Large-Scale Recognition
- Standards
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption
- Spoofing
- Data Quality
- Time-Lapse Effects and Aging
- Large-Scale Recognition
- Standards
Spoofing
What if the impostor has a recording

Text-Independent Speaker Recognition
Speech Recognition (ASR)
Natural Language Understanding (NLU)
Knowledge-Based Systems
Interactive Voice Response (IVR) system
Spoofing Handling -- Speech Biometrics Verification Process

Specific Questions

IVR System

Random Questions to Set off Impostors

ASR

NLU

Text-Independent Speaker Verification

Authentication Result

Centralized Database

Speaker Recognition – practical issues

Copyright: Recognition Technologies, Inc.
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption
- Spoofing
- Data Quality
- Time-Lapse Effects and Aging
- Large-Scale Recognition
- Standards
Legacy Data

- TMIT/NTIMIT (LDC)
- SIVA (ELRA)
- POLYVAR (ELRA)
- PPOLYCOOST (ELRA)
- KING (LDC)
- YOHO (LDC)
- Switchboard I & II (LDC)
- Cellular Switchboard (LDC)
- Tactical Speaker ID (LDC)
- Speaker Recognition (OGI)
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption
- Spoofing
- Data Quality
- **Time-Lapse Effects and Aging**
- Large-Scale Recognition
- Standards
Time-Lapse – First Re-Test

Number of Days lapsed between First Test and First Re-Test

- days lapsed
- mean

Copyright: Recognition Technologies, Inc.
Time-Lapse – Second Re-Test

Number of Days lapsed between First and Second Re-Tests

Copyright: Recognition Technologies, Inc.
Time-Lapse – Test Timeline

~50 days       ~50 days

First Test     First Re-Test     Second Re-Test

First Test

<table>
<thead>
<tr>
<th>Q</th>
<th>Q</th>
<th>Q</th>
<th>...</th>
<th>Q</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>...</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

First Re-Test

<table>
<thead>
<tr>
<th>Q</th>
<th>Q</th>
<th>Q</th>
<th>...</th>
<th>Q</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>...</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Second Re-Test

<table>
<thead>
<tr>
<th>Q</th>
<th>Q</th>
<th>Q</th>
<th>...</th>
<th>Q</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>...</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Time-Lapse – Verification Scenario

- For verification tests, for each setting, a total of \(22^2 = 484\) tests were run.
- Each speaker was treated as an impostor for all other speakers.
- Each session contains one unique speaker (candidate).
- Audio was collected randomly from over 100 stations with different sound systems.
- Microphone types are a combination of close-talk, farfield, analog input with imperfect grounding, USB, etc.
- Severe Mis-match conditions exist.
- All proficiency tests in this study were done in English.
- A part of the first test was used for enrollment.
- A different part of the audio was used for verification in the first seating.
Time-Lapse – Identification

Error Rate (%) vs Trial

Copyright: Recognition Technologies, Inc.
Time-Lapse – Verification

EER
First Trial 3%
Second Trial 17%
Third Trial 17%
Time-Lapse – Identification Augmented Data
Speaker Recognition – practical issues

Time-Lapse – Verification – Augmented Data

EER
First Trial 3%
Second Trial 10%
Third Trial 10%

Copyright: Recognition Technologies, Inc.
Time-Lapse – Identification

Error Rate (%) vs Trial

- Usual Enrollment
- Augmented-Data Enrollment
- Adapted Enrollment 1
- Adapted Enrollment 5

Copyright: Recognition Technologies, Inc.

Aug 6, 2010
Speaker Recognition – practical issues

Time-Lapse – Verification – 1 Iteration MAP

EER
First Trial 14%
Second Trial 7%
Third Trial 10%

Copyright: Recognition Technologies, Inc.
Practical Issues

Channel Mismatch
Numerical Stability
Privacy
Biometric Encryption

Spoofing
Data Quality
Time-Lapse Effects and Aging
Large-Scale Recognition

Standards
Large-Scale Speaker Recognition

Large Government Applications

- Social Security Eligibility Verification, Border Crossing, etc. – millions of participants
- Forensic Applications
  - Verification of Life Status for remote citizens – e.g. Pension plans

Financial Applications – Fraud Protection, Account Access, etc.

Large Health Insurance Memberships – Access to Medical Records, etc.

Large Corporation VoiceMail Applications

Telephone Order Credit Card Charges – Verify buyers in place of signature

Remote Test Proctoring – Requires continuous verification

Other System-Wide Applications – Requiring Remote Authentication or Customization
Some Applications

- Financial
- Forensic and Legal
- Surveillance
- Access Control
- Audio/Video Indexing (Diarization)
- Teleconferencing
- Proctorless Oral Testing
Practical Issues

- Channel Mismatch
- Numerical Stability
- Privacy
- Biometric Encryption

- Spoofing
- Data Quality
- Time-Lapse Effects and Aging
- Large-Scale Recognition

Standards
Standards

No real attention to speaker recognition standards

SVAPI – 1.0 and 2.0 – dead

BIOAPI – Concentrates on fingerprint and other non-sequential biometrics

Media Resource Control Protocol – MRCP 2.0

Session Initiation Protocol (SIP)

IETF signaling protocol for multimedia

UniMRCP – open source implementation

No audio data support – e.g., Audio sent through Real-time Transport Protocol (RTP)

TTS (SSML) and ASR (GRXML)

VoiceXML 3.0 – public draft

Common Biometric Exchange Formats Framework (CBEFF)

INCITS M1, ISO/IEC, and NIST
Standard Audio Format Encapsulation (SAFE)
Starting Question to Ask

What Should be Standardized at This Stage of Development in Speaker Recognition?
SAFE
Starting Question to Ask

What Should be Standardized at This Stage of Development in Speaker Recognition?

Audio Format?

Speaker Models?

Results of Recognition?

Interaction with Engines?
SAFE
Starting Question to Ask

What Should be Standardized at This Stage of Development in Speaker Recognition?

- Audio Format? **Definitely**
- Speaker Models?
- Results of Recognition?
- Interaction with Engines?
SAFE
Starting Question to Ask

What Should be Standardized at This Stage of Development in Speaker Recognition?

Audio Format?  
- Definitely

Speaker Models?  
- Not Yet

Results of Recognition?

Interaction with Engines?
SAFE
Starting Question to Ask

What Should be Standardized at This Stage of Development in Speaker Recognition?

- Audio Format?  
  - Definitely

- Speaker Models?  
  - Not Yet

- Results of Recognition?  
  - Yes

- Interaction with Engines?
SAFE
Starting Question to Ask

What Should be Standardized at This Stage of Development in Speaker Recognition?

Audio Format?  Definitely
Speaker Models?  Not Yet
Results of Recognition?  Yes
Interaction with Engines?  Yes
SAFE
Proposal and Status

- An Audio Encapsulation Standard
- Meeting Specific Requirements – Discussed Later
- Currently Considered by ANSI/INCITS for standardization
  - Public Review Period has been Completed
- Submitted to ISO/JTC1 SC37 (Project 19794-13)
- Recommended by the Chair of VBWG of W3C to be considered either as a Complete Standard or a Minimum Requirement for VoiceXML 3.0

Copyright: Recognition Technologies, Inc.
SAFE
Goals (Audio Format Only)

A Basic List of Audio Formats Meeting All Interchange Requirements
SAFE
Goals (Audio Format Only)

A Basic List of Audio Formats Meeting All Interchange Requirements

With Minimal Redundancy for the Sake of Clarity, Simplicity, and Compactness
SAFE
Goals (Audio Format Only)

A Basic List of Audio Formats Meeting All Interchange Requirements

With Minimal Redundancy for the Sake of Clarity, Simplicity, and Compactness

Preference Given to Open-Source and Royalty-Free Formats
SAFE
Goals (Audio Format Only)

- A Basic List of Audio Formats Meeting All Interchange Requirements
- With Minimal Redundancy for the Sake of Clarity, Simplicity, and Compactness
- Preference Given to Open-Source and Royalty-Free Formats
- Ease of Adoption
SAFE
Goals (Audio Format Only)

A Basic List of Audio Formats Meeting All Interchange Requirements

With Minimal Redundancy for the Sake of Clarity, Simplicity, and Compactness

Preference Given to Open-Source and Royalty-Free Formats

Ease of Adoption

Stability of Implementation
SAFE
Goals (Audio Format Only)

A Basic List of Audio Formats Meeting All Interchange Requirements

With Minimal Redundancy for the Sake of Clarity, Simplicity, and Compactness

Preference Given to Open-Source and Royalty-Free Formats – as Suggested by Kazuyuki

Ease of Adoption

Stability of Implementation

Relative Quality – Compared to Contenders

Data Security (Encryption/Decryption) Outside the Scope
SAFE
Sampling Process

- Analog Signal
- Sampling
- Storage

**Type**
- Periodic
- Variable-Rate
- Delayed
- Multirate
- Cyclic-Rate
- Random
- Pulse-Width Modulated

**Bit Rate (bps)**

*Periodic*: Bit Rate (bps) is Prop. to Sampling Freq. (Hz)

*Variable-Rate*: Bit Rate (bps) has Indirect Rel. to Freq. (Hz)
SAFE
Audio Coding Scenarios

Lossless Representation – Amplitude and Frequency are Unchanged

Lossless Compression May be Applied – gzip, bzip2

beigi@RecoTechnologies.com
SAFE
Audio Coding Scenarios

- **Lossless Representation** – Amplitude and Frequency are Unchanged

- **Lossless Compression May be Applied** – gzip, bzip2

- **Amplitude Compression** – Freq. Stays the Same, Amplitude is Represented Nonlinearly

- **Lossless Compression May be Applied** – gzip, bzip2
SAFE
Audio Coding Scenarios

- **Lossless Representation** – *Amplitude and Frequency are Unchanged*
  - **Lossless Compression May be Applied** – gzip, bzip2

- **Amplitude Compression** – *Freq. Stays the Same, Amplitude is Represented Nonlinearly*
  - **Lossless Compression May be Applied** – gzip, bzip2

- **Variable-Rate Sampling** – *Aggressive Variable Bitrate Compression*
SAFE

Audio Coding Scenarios

- **Lossless Representation** – *Amplitude and Frequency are Unchanged*
  - Lossless Compression May be Applied – gzip, bzip2

- **Amplitude Compression** – *Freq. Stays the Same, Amplitude is Represented Nonlinearly*
  - Lossless Compression May be Applied – gzip, bzip2

- **Variable-Rate Sampling** – *Aggressive Variable Bitrate Compression*

- **Streaming** – *Usually includes multirate sampling and streaming*
SAFE
Audio Interchange Scenarios

Lossless Representation

Microsoft WAV Comes to Mind — A Wrapper which includes over 104 codecs

LPCM offers all that is needed — Just need to code the header information
SAFE
Audio Interchange Scenarios

**Lossless Representation**

- LPCM offers all that is needed — *Just need to code the header information*

**Amplitude Compression**

- G.711 and G.711.1 ITU-T define PCMU and PCMA for 64, 80, and 96kbps

- ADPCM was considered, but it has many flavors and is not open source
SAFE
Audio Interchange Scenarios

- **Lossless Representation**
  - LPCM offers all that is needed – *Just need to code the header information*

- **Amplitude Compression**
  - G.711 and G.711.1 ITU-T define PCMU and PCMA for 64, 80, and 96kbps

- **Variable-Rate Sampling**
  - **MP3 comes to mind** – *Patent driven and certainly not an open standard*
  - **OGG Vorbis** – *Open Source and better quality as MP3 for the same bit rate*
SAFE
Audio Interchange Scenarios

Lossless Representation

LPCM offers all that is needed – *Just need to code the header information*

Amplitude Compression

G.711 and G.711.1 ITU-T define PCMU and PCMA for 64, 80, and 96kbps

Variable-Rate Sampling

OGG Vorbis – *Open Source and better quality as MP3 for the same bit rate*

Streaming – *Usually includes variable-rate sampling and streaming*

OGG Media Stream – *Open Source with capability of streaming different audio types*
SAFE
Audio Interchange Scenarios

<table>
<thead>
<tr>
<th>Quality</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lossless</td>
<td>Linear PCM (LPCM)</td>
</tr>
<tr>
<td>Amplitude Compression</td>
<td>$\mu$-law (PCMU) and A-law (PCMA)</td>
</tr>
<tr>
<td>Aggressive variable bit-rate compression</td>
<td>OGG Vorbis</td>
</tr>
<tr>
<td>Streaming</td>
<td>OGG Media Stream</td>
</tr>
</tbody>
</table>
# SAFE

## Audio Format Header

<table>
<thead>
<tr>
<th>Type</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>U16</td>
<td>ByteOrder</td>
<td>The value is 0xFF00 and it is set by the audio file producer</td>
</tr>
<tr>
<td>U16</td>
<td>HeaderSize</td>
<td>Size of this header in bytes</td>
</tr>
<tr>
<td>Boolean</td>
<td>Streaming</td>
<td>This will 0 for non-streaming and 1 for streaming. This boolean variable is redundant since the AF_FORMAT for streaming audio is greater than 0xFFFF. However, it is used for convenience.</td>
</tr>
<tr>
<td>U16</td>
<td>Compression</td>
<td>Standard Data Compression Scheme</td>
</tr>
<tr>
<td>U64</td>
<td>FileLengthInBytes</td>
<td>In Bytes not including the header</td>
</tr>
<tr>
<td>U64</td>
<td>FileLengthInSamples</td>
<td>In Number of samples</td>
</tr>
<tr>
<td>U16</td>
<td>AudioFormat</td>
<td>See AF_FORMAT macros</td>
</tr>
<tr>
<td>U16</td>
<td>NumberOfChannels</td>
<td>Number of channels, <em>N.B.,</em> Channel data alternates</td>
</tr>
<tr>
<td>U32</td>
<td>SamplingRate</td>
<td>Sampling rate in samples per second – This is the audio sampling rate and not necessarily the sampling rate of the carrier which may be variable.</td>
</tr>
<tr>
<td>U64</td>
<td>AudioFullSecondsOf</td>
<td>It is the truncated number of seconds of audio</td>
</tr>
<tr>
<td>U32</td>
<td>AudioRemainderSamples</td>
<td>This is the number of samples of audio in the remainder which was truncated by the above variable</td>
</tr>
<tr>
<td>U16</td>
<td>BitsPerSample</td>
<td>Number of bits per sample, may be 0 for formats which use variable bits</td>
</tr>
</tbody>
</table>
# SAFE

Audio Interchange Scenarios

<table>
<thead>
<tr>
<th>Macro</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF_FORMAT_UNKNOWN</td>
<td>0x0000</td>
</tr>
<tr>
<td>AF_FORMAT_LINEAR_PCM</td>
<td>0x0001</td>
</tr>
<tr>
<td>AF_FORMAT_MULAW</td>
<td>0x0002</td>
</tr>
<tr>
<td>AF_FORMAT_ALAW</td>
<td>0x0003</td>
</tr>
<tr>
<td>AF_FORMAT_OGG_VORBIS</td>
<td>0x0004</td>
</tr>
<tr>
<td>AF_FORMAT_OGG_STREAM</td>
<td>0x1000</td>
</tr>
</tbody>
</table>
SAFE
Lossless Compression

Uncompressed

Lempel-Ziv (LZ77) – gzip

Burrows-Wheeler Block Sort + Huffman Coding – bzip2

<table>
<thead>
<tr>
<th>Macro</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF_COMP_NONE</td>
<td>0x0000</td>
</tr>
<tr>
<td>AF_COMP_GZIP</td>
<td>0x0001</td>
</tr>
<tr>
<td>AF_COMP_BZIP2</td>
<td>0x0002</td>
</tr>
</tbody>
</table>
## SAFE -- Lossless Compression

### LPCM – Example 44kHz

<table>
<thead>
<tr>
<th>Compression</th>
<th>Factor</th>
<th>Size</th>
<th>Coding Time</th>
<th>Decoding Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAC</td>
<td>2.2</td>
<td>0.46</td>
<td>0.7</td>
<td>1.7</td>
</tr>
<tr>
<td>gzip (level 6)</td>
<td>1.4</td>
<td>0.7</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>bzip2</td>
<td>2.0</td>
<td>0.5</td>
<td>1.4</td>
<td>3.2</td>
</tr>
</tbody>
</table>

## Mu-Law -- Example

<table>
<thead>
<tr>
<th>Compression</th>
<th>Factor</th>
<th>Size</th>
<th>Coding Time</th>
<th>Decoding Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAC</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gzip (level 6)</td>
<td>1.5</td>
<td>0.67</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>bzip2</td>
<td>2.2</td>
<td>0.5</td>
<td>2.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>
How About Fusion with Other Biometrics?

- **Fusion with Face Recognition is Quite Natural**
- **Event Detection Using Image Cues as well as Audio Cues**
- **Video Indexing Using Speaker Identity** – *Broadcast News, etc.*
Audio & Video Fusion

Audio & Video Fusion

FE

FEATURES

SEG

FEATURES

RECO

TIME STAMPED WORDS

SPKR-ID

SPKR-ID

DISPLAY

FUSED ID

VIDEO

FRAMES

FACE DETECTION & LANDMARK ID

FRAME SCORES

RANKED SEGMENT VIDEO IDS

DISPLAY

FUSED ID

AUDIO

Copyright: Recognition Technologies, Inc.
Audio & Video Fusion

**Example of when fusion helps** – *Correct Speaker is OH*

<table>
<thead>
<tr>
<th>RANK</th>
<th>AUDIO</th>
<th>VIDEO</th>
<th>FUSED</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UM</td>
<td>JW</td>
<td>OH</td>
</tr>
<tr>
<td>2</td>
<td>OH</td>
<td>OH</td>
<td>AK</td>
</tr>
<tr>
<td>3</td>
<td>UF</td>
<td>AK</td>
<td>GB</td>
</tr>
<tr>
<td>4</td>
<td>AK</td>
<td>SF</td>
<td>JM</td>
</tr>
<tr>
<td>5</td>
<td>JM</td>
<td>GB</td>
<td>JW</td>
</tr>
<tr>
<td>6</td>
<td>GB</td>
<td>JM</td>
<td>SF</td>
</tr>
</tbody>
</table>
Concurrent Transcription, Speaker Segmentation and Identification

PCM INPUT

FEATURES

EXTRACT FEATURE VECTORS

FEATURES

TRANSCRIPTION ENGINE

TIME-STAMPED WORDS

TURNS

SEGMENTATION

FEATURES

FORM UTTERANCE FROM SEGMENT

DATA STORE

IDENTIFICATION ENGINE

SPEAKER NAMES

SPEAKER TRAINING FILES

ENROLLMENT ENGINE

OFFLINE PROCESS

SPEAKER DATABASE

VERIFICATION ENGINE

SPEAKER ID

VERIFIED SPEAKER ID

DISPLAY

INTERLEAVE TURNS AND WORDS

ISOLATED SPEECH SEGMENTS
For more information refer to:

Fundamentals of Speaker Recognition
Out in September 2010
Springer-Verlag
References


