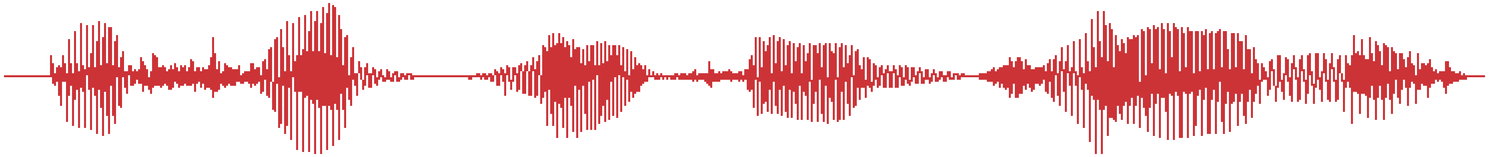
Challenges of Large-Scale Speaker Recognition



by

Homayoon Beigi

Beigi@RecognitionTechnologies.com

http://www.RecognitionTechnologies.com

(COST275 Keynote Speech)

Recognition Technologies, Inc.

300 Hamilton Avenue
White Plains, NY, U.S.A.

Speaker's Background

- **Recognition Technologies, Inc.** - *President* - 2002-present
- **Internet Server Connections, Inc.** - *Vice President* - 2000-present
- **Columbia University** – *Adjunct Professor*
  Courses: Signal Recognition, Speech Recognition, and Digital Control
- **IBM T.J. Watson Research Center** – Research Staff Member - 1991-2000
- **Columbia University** – BS, MS & PhD - 1990

Introduction

- What are the different manifestations and modalities of Speaker Recognition?
- Some circumstances under which Speaker Recognition make sense.
- Where do we need Large-Scale Speaker Recognition?
- Where do we stand with Large-Scale Recognition?
- What are the challenges of Large-Scale Speaker Recognition?
- What is Recognition Technologies doing to address this problem?

Manifestations of Speaker Recognition

- **Speaker Identification** – *suffers the most in large-scale scenarios*
- **Speaker Verification** – *cohort computations become problematic*
- **Speaker Classification** – *suffers the same way as Identification does*
- **Speaker Tracking** – *same treatment as Verification*
- **Speaker Detection** – *could be interpreted as ID or Verification – vague term*
- **Speaker Segmentation** – *not affected much as far as I can predict*

Modalities of Speaker Recognition

- **Text Dependent** – *Fixed text is spoken (not as attractive as other choices)*
- **Text Independent** – *The specific text is not used in the recognition (Language Independent ?)*
  - **Language Independence** – largely, *but may use different processing for different. languages*
- **Text Prompted** – *usually done randomly or based on some formula*
- **User Selected** – *may be treated like a password (user provides the question – not too practical)*
- **Speech Biometrics** – *may be used to come up with text prompting – most ideal*

Questions

- What are the different manifestations and modalities of Speaker Recognition?
- **Some circumstances under which Speaker Recognition make sense.**
- Where do we need Large-Scale Speaker Recognition?
- Where do we stand with Large-Scale Recognition?
- What are the challenges of Large-Scale Speaker Recognition?
- What is Recognition Technologies doing to address this problem?

When to Have Speaker Recognition

- Finger Print not available (damaged fingers) – *2% of the population (NIST)*
- Iris damage – Some of the blind
- Population Resistance – 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 mask 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
- Cellular Telephone and PDA-type device security

Questions

- What are the different manifestations and modalities of Speaker Recognition?
- Some circumstances under which Speaker Recognition make sense.
- **Where do we need Large-Scale Speaker Recognition?**
- Where do we stand with Large-Scale Recognition?
- What are the challenges of Large-Scale Speaker Recognition?
- What is Recognition Technologies doing to address this problem?

Why Large-Scale?

- Large Government Applications
  - Social Security Eligibility Verification – *millions of participants*
  - Verification of Life Status for remote citizens – *e.g. Pension plans*
- Financial Applications
- Large Health Insurance Memberships
- Large Corporation VoiceMail Applications
- Telephone Order Credit Card Charges – Verify buyers in place of signature
- Remote Test Proctoring – *requires continuous verification*
- Any other system-wide applications requiring remote authentication and customization

Questions

- What are the different manifestations and modalities of Speaker Recognition?
- Some circumstances under which Speaker Recognition make sense.
- Where do we need Large-Scale Speaker Recognition?
- **Where do we stand with Large-Scale Recognition?**
- What are the challenges of Large-Scale Speaker Recognition?
- What is Recognition Technologies doing to address this problem?

Large-Scale
Research Status

- Data Collection Efforts
- Curse of Large Databases – Speakers continually closer to each-other and unmanageable
- Need more challenges :-) ?!!

Legacy Data

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

TIMIT/NTIMIT
Amerian English (LDC)

- 630 (438 M + 192 F)
- Clean Wideband Handset / Telephone Handset PSTN (Half Long Distance)
- Read out Sentences
- Controlled Clean Environment
- Only one session per speaker

SIVA
Italian (ELRA)

- 40 and 800 (50% M + 50% F)
- Telephone Handset PSTN
- Short Sentences (Prompted Words & Digits)
- Home/Office Environment
- 18 sessions – over 3 days for the 40 and single session for the 800

PolyVar
European French (ELRA)

- 143 (85 M + 58 F)
- Telephone Handset PSTN and ISDN
- Read and Prompted words, digits, sentences, question & spontaneous speech
- Home/Office Environment
- 1-229 sessions – 160 hours overall

POLYCOST
English & European (ELRA)

- 133 (75 M + 59 F)
- Telephone Handset ISDN
- Read out and prompted words, digit strings, read out sentences, free-style monologues
- Home/Office Environment
- More than 5 sessions per speaker over many days or weeks – non-native speakers

KING
American English (LDC)

- 51 Male speakers
- Wideband microphone as well as electret handsets through PSTN
- Read out and Prompted words, digit strings, read sentences, free-style descriptions of photos
- Clean speech and clean environment
- 10 sessions per speaker over weeks

YOHO
American English (LDC)

- 138 (106 M + 32 F)
- 3.8 kHz clean handset
- Prompted digit strings
- Clean speech in an office environment
- 4 enrollment and 10 verification sessions per speaker

Switchboard I & II
American English (LDC)

- 543 & 657 (~50% M + ~50% F)
- Various Telephone handsets through PSTN
- Conversational
- Home and Office environment
- 1-25 sessions per speaker – 5 minutes per each session
- SPIDRE is a subset of switchboard I selected for speaker ID

Cellular Switchboard
American English (LDC)

- 190 (~50% M + ~50% F)
- Various cellular handsets through GSM 1900
- Conversational speech
- Various natural environments
- 10 or more sessions per speaker over dats – 5 nminutes per session

Tactical Speaker ID (TSID)
American English (LDC)

- 40 (39 M + 1 F)
- 4 Military radio handsets and one electret microphone – HF, UHF, VHF and Wideband
- Read out sentences, digits and free-style speech
- Outdoors Environment
- 1 session per speaker

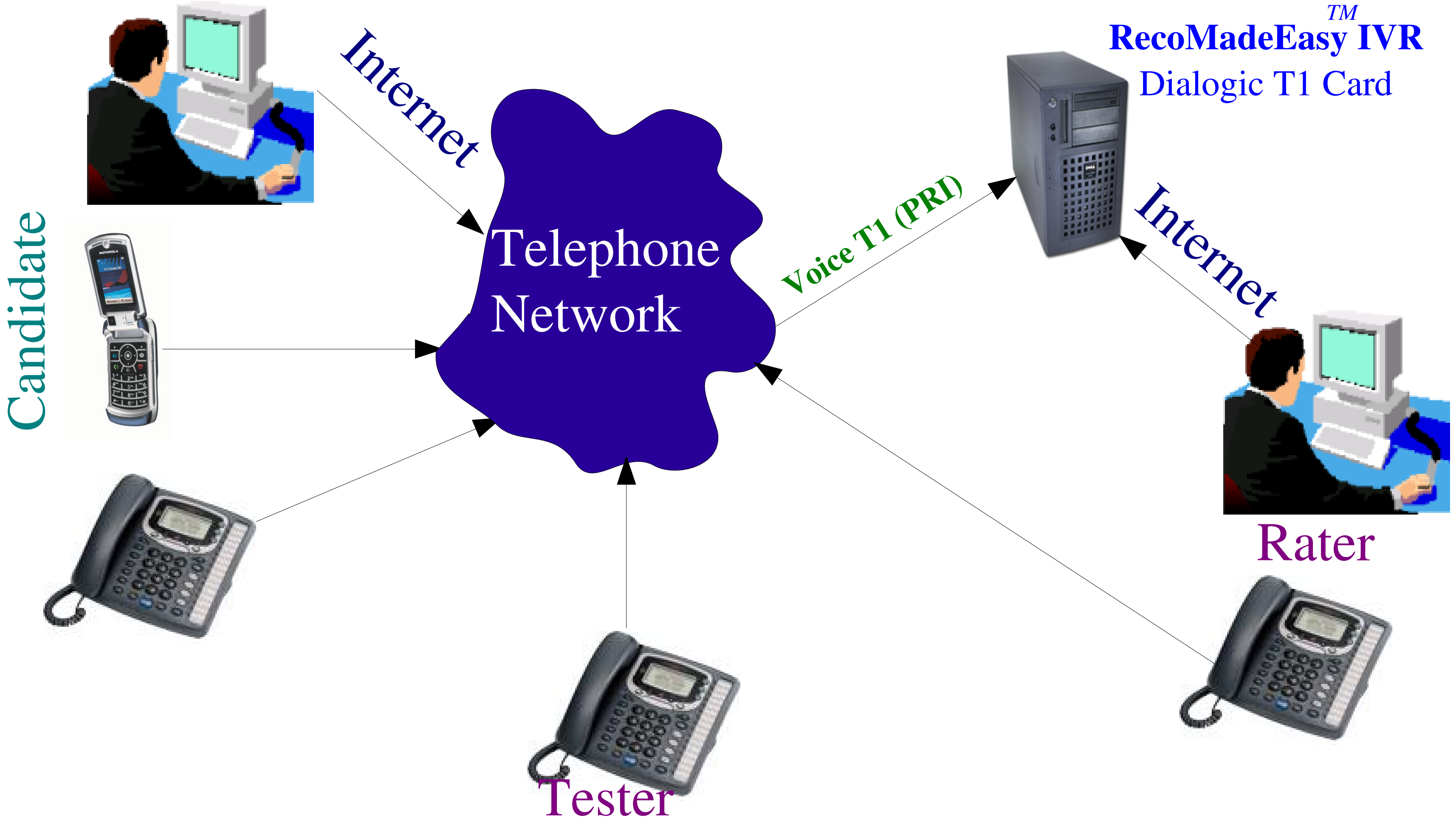
Speaker Recognition Corpus
American English (OGI)

- 100 (47 M + 53 F)
- Various telephone handsets through PSTN
- Prompted digits, phrases and momologues
- Home and Office Environments
- About 12 sessions per speaker over months

Advances toward Semi-Large-Scale Data Collection (ELRA)

- A total of 143 corpora
- Maximum of 4000 speakers
- A lot of attention to Telephone and cellular telephone handsets
- Some recordings done through sessions over many months
- Large Corpora in British English, German, Spanish, Italian, French, Danish and Finish

Language Proficiency Testing
Recognition Technologies, Inc.



Language Proficiency Testing
Recognition Technologies, Inc.

- 12,080 Total Sessions collected over 18 months
- Each session contains one unique speaker (candidates)
- About 100 speakers (testers) are repeatedly heard in all sessions
- Mixed telephone handsets over PSTN, ISDN, Internet and a few cellular
- In 48 different languages, although at least about 1 minute in English per session
- Data is useful for three practical applications
  - Continuous Speaker Verification
  - Speaker Segmentation
  - Language Detection

Curse of Large Data-Sets

- Target number of speakers are in the order of hundreds of thousands and millions
- The discrete nature of classes gives way to a more continuous nature – hurting results
- Error rate increases with the number of classes
- Identification by matching against all possible models is not practical
- Computing cohorts becomes harder as the number of speakers increases
- Real-Time computation becomes a big issue
- Optimal searching becomes an important issue

What if the impostor has a recording (Speech Biometrics)

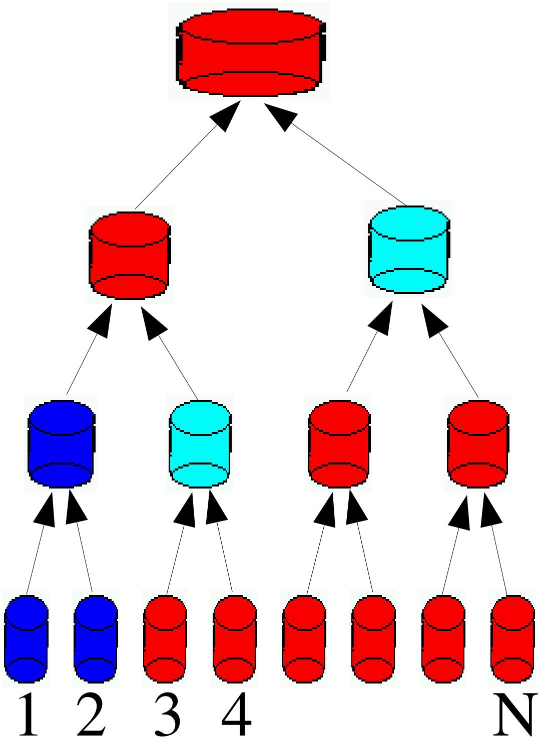
- Text-Independent Speaker Recognition
- Speech Recognition (ASR)
- Natural Language Understanding (NLU)
- Knowledge-Based Systems
- Interactive Voice Response (IVR) system

How Is It Done Then?

- More efficient Identification – *using hierarchical techniques*
- A Voice model as well as a distance measure for comparing speakers
- A centralized database is needed – *client-server models*
- Use in conjunction with Speech Biometrics for accuracy
- Speech Biometric system used prompt the speakers – *avoid spoofing using recoded voice*
- Use standard Interactive Voice Response (IVR) systems for the automation

Hierarchical Model

See EuroSpeech 1999 paper by Homayoon Beigi, et. al.



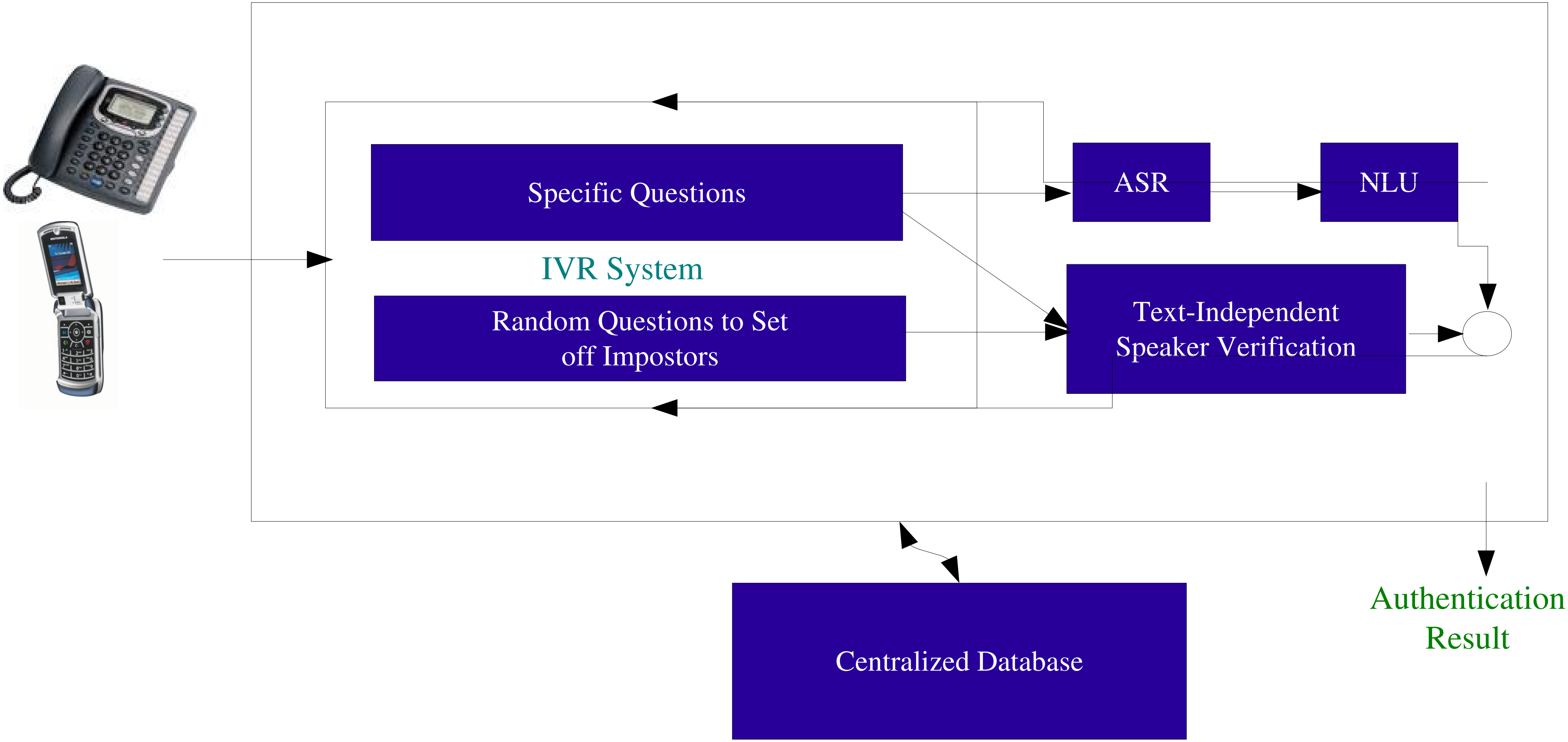
Is used as a part of the set of Complementary Models for

- N-ary (Binary) speaker tree
- Some nodes on the tree may be used for rejection models
- Aggressive Complementary Models possible for very large-scale systems
- Background models may be used like other models

Speech Biometrics
Enrollment

- Use an enrollment form to obtain name, address and other vital information
- The system records all the utterances by user in the process of enrollment
- The data is used by the Knowledge-Base and Speaker Recognition systems
- The enrollee may present extra questions to be asked

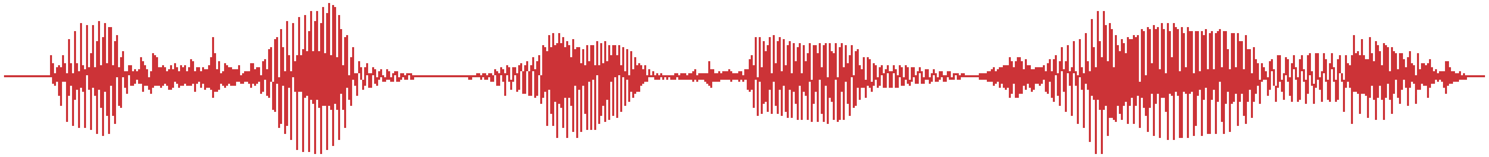
Speech Biometrics
Verification Process



Conclusion

- **Need a self-contained model for each speaker**
- **A distance measure to allow comparison between speaker models**
- **A good method for creating a hierarchical representation of the speaker database**
- **A background model resembling the speaker models**
- **Complementary models to help determine cohorts for open-set recognition**
- **Centralized database with a client-server recognition scheme**
- **Rich data for generating base models created from as many speakers as possible**
- **Rich channel model and possibly channel detection and separation**
- **Should we still re-think the front-end processing for Speaker Recognition?**

Challenges of Large-Scale Speaker Recognition



by

Homayoon Beigi

Beigi@RecognitionTechnologies.com

http://www.RecognitionTechnologies.com

(COST275 Keynote Speech)

Recognition Technologies, Inc.

300 Hamilton Avenue
White Plains, NY, U.S.A.