Information Technology Laboratory

Information Access Division (IAD)

NIST

National Institute of Standards and Technology

The 2010 NIST Speaker Recognition Evaluation (SRE10)

Alvin Martin, Craig Greenberg

NIST Multimodal Information Group

SRE10 Workshop

24-25 June 2010, Brno, Czech Republic

Outline

- Introduction
- What's different in SRE10?
- Evaluation rules
- Test conditions
- Participants
- Results
  - Core Test – Common Conditions
  - History
  - Performance factors
  - Non-Core tests
- Summary

Introduction

- SRE10 is latest in series of NIST evaluations of automatic speaker detection begun in 1996
  - Most recent NIST SRE occurred in 2008
- Basic task is speaker detection

- Given a target speaker and a test speech segment, determine if the target is speaking in the test segment
  - A trial consists of a *model* (target training data) and a *test segment*
  - Outputs required for each trial are:

    a *decision* (*'T'* or *'F'*) and

    a *score* (preferably a *log-likelihood ratio*)

Introduction (cont’d)

- Evaluation rules similar to those in past
- A **core test** was required of all participants
- Other tests included variations of the train and test segment conditions, and were optional
- Evaluation open to all interested participants willing to follow evaluation rules

What's Different in 2010

- Data
  - Vocal effort for most Mixer 6 speakers
    - Phone calls made with high, low, and normal vocal effort
  - Greybeard Corpus for limited testing of aging effect
    - Phone calls from speakers who also participated in earlier test sets
  - Greatly increased number of trials
    - core trials: 570176
    - core-extended trials: 6,451,524
- Official metric
  - Cost function emphasizing low false alarm rates (new *parameter settings*)
    - New cost function used in core and 8conv-core tests
    - Old cost function retained in other tests
- Extended trials
  - To enhance statistical significance at very low false alarm rates
  - Included all possible speaker pairs in non-target trials
- HASR (Human Assisted Speaker Recognition) offered
  - Trials included in the core test of main evaluation
  - To be discussed tomorrow

What’s Different – Data

- Mixer 6 – 430 speakers; for each there were generally
  - 3 LDC interview sessions
  - 3 vocal effort phone calls (high, low, normal vocal effort)
  - External phone calls, preferably from multiple phone lines

- Greybeard – 166 speakers
  - Current calls (2008)
  - Calls from earlier corpora (some as far back as 1990, most from early 2000’s)
  - *Scored separately from Mixer 6*
  - *Key is not being released to allow possible reuse*

What’s Different - Decision Cost Function(s)

$$C_{Norm} = \frac{((C_{Miss} * P_{Miss|Target} * P_{target}) + (C_{FA} * P_{FA|NonTarget} * P_{NonTarget}))}{C_{Default}}$$

Cost of a miss	C_{Miss} = 1 (core, 8conv-core), = 10 (other conditions)
Cost of a false alarm	C_{FA} = 1
Probability of a target	P_{Target} = 0.001 (core, 8conv-core), = 0.01 (other conditions)
Probability of a non-target	$P_{Nontarget} = 1 – P_{Target}$ = (0.999 or 0.99)
A normalization factor ($C_{Default}$) is defined to make 1.0 the score of a knowledge-free system that always decides “False”. $C_{Default} = \min(C_{Miss} * P_{Target} , C_{FA} * P_{Nontarget}))$ (= 0.001 or 0.1)	

What’s Different – Extended trials

- New cost function typically results in a minimum cost operating point with FA rate in the 0.01% - 0.1% range
  - “Rule of 30” implies ~30,000 non-target trials per condition needed at 0.1% FA rate
- Sites requested additional trials to enhance statistical significance
  - Extended trials provided ~300,000 non-target trials per condition where possible
  - Non-target trials included all possible speaker pairs, and in some cases all possible segment pairs

Evaluation Rules

- Each trial decision to be made independently
  - Based only on the specified segment and the speaker model
  - Use of information about other test segments and/or other target speakers is NOT allowed

- Normalization over multiple ***test segments*** NOT allowed
- Normalization over multiple ***target speakers*** NOT allowed

- Use of evaluation data for impostor modeling NOT allowed
- Use of manually produced transcripts or any other human interaction with the data NOT allowed

- Knowledge of the model speaker gender is ALLOWED
  - No cross sex trials

Test Conditions (outline)

- Training conditions
- Test segment conditions
- Evaluation Test matrix
- Core Test – Common Conditions

Training Conditions

Identifier	*Description*
10 sec	Telephone conversational excerpt containing about 10 seconds of the target speaker’s speech in the channel of interest
Core	Telephone conversational excerpt of about 5 minutes total duration, recorded over a telephone or room microphone channel, and involving the target speaker on its designated side; or Interview segment of about 3 or 8 minutes in total duration, recorded over a room microphone channel, and involving the target speaker and an interviewer
8conv	Eight telephone conversational excerpts of about 5 minutes in total duration, recorded over a telephone channel and involving the target speaker on their designated sides
8summed	Eight excerpts similar to 8conv but with each excerpt consisting of a single, summed channel, formed by sample-by-sample summing of its two sides. The eight interlocutors are all distinct.

NIST
National Institute of
Standards and Technology

Test Segment Conditions

Identifier	***Description***
10 sec	Telephone conversational excerpt containing about 10 seconds of speech in the channel of interest
Core	Telephone conversational excerpt of about 5 minutes total duration, recorded over a telephone or room microphone channel; or Interview segment of about 3 or 8 minutes in total duration, recorded over a room microphone channel, and involving the subject and an interviewer
Summed	Summed channel telephone conversational excerpt of about 5 minutes in total duration, recorded over a telephone channel

Evaluation Test Matrix

		Test Segment Conditions		
		10sec	**core**	**summed**
Training Conditions	**10sec**	optional	-	-
	core	optional	**required**	optional
	8conv	optional	optional	optional
	8summed	-	optional	optional

- The ***core test*** is the single required condition
- Non-summed phone conversation segments were two-channel, with side of interest designated
- Interview segments each included interviewer’s close-talking mic channel, to support speaker separation
- ASR output of all speech segments was made available using phone or highest quality microphone channel available – Thanks to BBN!

Core Test – Common Conditions

Within the core test there are (9) “Common Conditions”

1) Interview speech trials with **matched mics** for train and test
2) Interview speech trials with **unmatched mics** for train and test
3) Trials involving interview training speech and normal vocal effort conversational telephone test speech
4) Trials involving interview training speech and normal vocal effort conversational telephone test speech recorded over a room microphone channel
5) Different number trials involving normal vocal effort conversational telephone speech in training and test

Core Test – Common Conditions (cont'd)

Within the core test there are (9) "Common Conditions"

6) Telephone channel trials involving normal vocal effort conversational telephone speech in training and **high vocal effort** conversational telephone speech in test
7) Room microphone channel trials involving normal vocal effort conversational telephone speech in training and **high vocal effort** conversational telephone speech in test
8) Telephone channel trials involving normal vocal effort conversational telephone speech in training and **low vocal effort** conversational telephone speech in test
9) Room microphone channel trials involving normal vocal effort conversational telephone speech in training and **low vocal effort** conversational telephone speech in test

Numbers of Trials

Common Condition	Trials Target (Non-target)	Extended Trials Target (Non-target)
1	2152 (60712)	4304 (795995)
2	7535 (212307)	15084 (2789534)
3	1633 (56410)	3989 (637850)
4	2366 (83536)	3637 (756775)
5	708 (29665)	7169 (408950)
6	361 (28311)	4137 (461438)
7	359 (27997)	359 (82551)
8	298 (28306)	3821 (404848)
9	290 (27230)	290 (70500)

Participating Sites and Systems

Participating Sites and Systems

System Identifier	Site	Location
ABC	Agnitio	South Africa
ABC	Brno University of Technology	Czech Republic
ABC	CRIM	Canada
ALP	Alpineon	Slovenia
ATVS	Universidad Autónoma de Madrid	Spain
BOUN	Bogazici University	Turkey
CCNT	Zhejiang University	China
CLIK	LIMSI-CNRS	France
CLIK	Carnegie Mellon University	USA
COGENT	Cogent Systems	USA
CRIM	CRIM	Canada

Participating Sites and Systems

System Identifier	Site	Location
CRSS	University of Texas at Dallas	USA
EHU	University of the Basque Country	Spain
HKCUPU	Chinese University of Hong Kong	China
HKCUPU	Hong Kong Polytechnic University	China
I3A	University of Zaragoza	Spain
I4U	Institute for Infocomm Research	Singapore
I4U	University of Science and Technology of China	China
I4U	Universtiy of Joensuu (Eastern Finland)	Finland
I4U	The University of New South Wales	Australia
I4U	Nanyang Techological University	Singapore
IBM	IBM	USA

Participating Sites and Systems

System Identifier	Site	Location
ICSI	International Computer Science Institute	USA
IFLY	University of Science and Technology of China	China
IIR	Institute for Infocomm Research	Singapore
IITKGP	India Institute of Technology Kharagpur	India
ILPGIP	PerSay, GM, IBM Haifa, Israeli Police	Israel
IOASLR	Institute of Acoustics, Chinese Academy of Sciences	China
IRITIN	Institut de Recherche en Informatique de Toulouse	France
IRITIN	Institut National de Recherche en Informatique et Automatique	France
L2FUPC	Laboratório de sistemas de Língua Falada	Portugal
L2FUPC	Universitat Politècnica De Catalunya	Spain
LIA	Université d'Avignon	France

Participating Sites and Systems

System Identifier	Site	Location
LPT	Loquendo	Italy
LPT	Politecnico di Torino	Italy
LRDE	LRDE-EPITA	France
MITLL	MIT Lincoln Laboratory	USA
NCMF	National Center for Media Forensics	USA
NTUT	National Taipei University of Technology	Taiwan
OZU	Ozyegin University	Turkey
PORT	Porticus Technolgoy	Lithuania
QUT	Quuensland University of Technology	Australia
RUN	Radboud University	Netherlands
SCL	Speech Communication Lab, University of Maryland	USA

Participating Sites and Systems

System Identifier	Site	Location
SLS	MIT Computer Science and Artificial Intelligence Laboratory	USA
SRI	SRI International	USA
STMSGP	STMicroelectronics Asia Pacific	Singapore
SVID	Speech Technology Center	Russia
SVIST	Shanghai Voice Info Science and Technology	China
TEC	Technológico de Monterrey	Mexico
THU	Tsinghua University	China
TITECH	Tokyo Insitute of Technology	Japan
TUL	Technical University of Liberec	Czech Republic
UAS	Tubitak-Uekae	Turkey
UAS	Sabanci University	Turkey

Participating Sites and Systems

System Identifier	Site	Location
UOB	University of Balamand	Lebanon
UPMFIM	Universidad Politéchnica de Madrid	Spain
UWB	University of West Bohemia	Czech Republic
UWS	Swansea University	United Kingdom
VLD	ValidSoft	United Kingdom
XMU	Xiamen University	China

Record number of participants. 5 continents. First-time reps: Japan, Turkey, & Russia	
Sites	58
System Identifiers	49
Total Core Systems	113