The 2012 NIST Speaker Recognition Evaluation

Craig Greenberg, Vince Stanford,
Alvin Martin, Meghana Yadagiri
George Doddington

NIST Multimodal Information Group

Outline

- Introduction
- What's different in SRE12?
- Evaluation rules
- Evaluation tests
- Participants
- Results
  - Common Conditions
    - Comparison of Cnorm and Cllr results
  - Performance Factors
  - History
  - Correlation Among Systems
- Summary

Introduction

- SRE12 is latest in series of NIST evaluations of automatic speaker detection begun in 1996
  - Most recent NIST SRE occurred in 2010
- 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 contains a *target speaker id* and a *test segment*

Introduction

- Evaluation rules similar to those in past, with notable exceptions
- 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 2012

- Training
- Test
- System Output / Metric

What's Different in 2012

- Training
  - Most of the training data released in advance of the evaluation
  - More and more varied training data for each speaker
    - Majority of target speakers in all evaluation tests had more than one segment
    - In many cases target speakers had mic int, mic phn, and tel phn training data
  - Joint knowledge of the target speakers is allowed
    - In the past, trials processed without knowledge of other target speakers

What's Different in 2012

- Test
  - Known and unknown non-target speakers
  - Duration
    - 300, 100, 30 sec
  - Noise
    - Additive, Environmental
  - No ASR provided
  - No masking noise on interviewer channel

What's Different in 2012

- System Output
  - Log-likelihood ratio outputs required
  - No decision required
- Official Metric
  - Average cost using two different sets of parameters

What's Different - Decision Cost Function

$$C_{\text{Det}} = C_{\text{Miss}} \times P_{\text{Target}} \times P_{\text{Miss|Target}} + C_{\text{FalseAlarm}} \times (1 - P_{\text{Target}}) \times \left(P_{\text{FalseAlarm|KnownNonTarget}} \times P_{\text{Known}} + P_{\text{FalseAlarm|UnknownNonTarget}} \times (1 - P_{\text{Known}}) \right)$$

C_{Miss} := the cost of a miss
$C_{\text{FalseAlarm}}$:= the cost of a false alarm
P_{Target} := the *a priori* probability that the segment speaker is the target speaker
P_{Known} := the *a priori* probability that the non-target speaker is one of the evaluation target speakers

		C_{Miss}	C_{FA}	$P_{\text{Target-A1}}$	$P_{\text{Target-A2}}$	P_{Known}
Test Segment Condition	Core	1	1	0.01	0.001	0.5
	Extended					0.5
	Summed					0.5
	Known					1
	Unknown					0

What's Different - Decision Cost Function

Normalizing, we get

$$C_{\text{Norm}} = C_{\text{Det}} / C_{\text{Default}}$$

where

$$C_{\text{Default}} = min \begin{Bmatrix} C_{\text{Miss}} \times P_{\text{Target}} \\ C_{\text{Fa}} \times (1 - P_{\text{Target}}) \end{Bmatrix} = C_{\text{Miss}} \times P_{\text{Target}}$$

What's Different - Decision Cost Function

Equivalently,

$$C_{\text{Norm}}(\beta) = P_{\text{Miss}\,|\,\text{Target}} \times \beta \times \left\{ \begin{array}{r} P_{\text{Known}} \times P_{\text{FalseAlarm}\,|\,\text{KnownNontarget}} + \\ (1 - P_{\text{Known}}) \times P_{\text{FalseAlarm}\,|\,\text{UnknownNontarget}} \end{array} \right\}$$

where

$$\beta = \left(\frac{C_{\text{FalseAlarm}}}{C_{\text{Miss}}}\right)\left(\frac{1 - P_{\text{Target}}}{P_{\text{Target}}}\right)$$

What's Different - Decision Cost Function

Finally, thanks to Niko Brummer, David van Leeuwen, Daniel Ramos, Joaquin Gonzalez-Rodriguez

$$C_{\text{Primary}} = \frac{\{C_{\text{Norm}(\beta_{A1})} + C_{\text{Norm}(\beta_{A2})}\}}{2}$$

$$\beta_{A_k} = \left(\frac{C_{\text{FalseAlarm}}}{C_{\text{Miss}}}\right)\left(\frac{1 - P_{\text{Target}-A_k}}{P_{\text{Target}-A_k}}\right)$$

		C_{Miss}	C_{FA}	$P_{\text{Target-A1}}$	$P_{\text{Target-A2}}$	P_{Known}
Test Segment Condition	Core	1	1	0.01	0.001	0.5
	Extended					
	Summed					
	Known					1
	Unknown					0

Evaluation Rules

- Each trial decision to be made based on:
  - The specified segment and the set of target speakers
  - Use of information about other test segments is NOT allowed
- Normalization over multiple ***target speakers*** IS allowed
- Normalization over multiple ***test segments*** 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 ALLOWED
  - No cross sex trials

Evaluation Tests (outline)

- Training Conditions
- Test Segment Conditions
- Evaluation Test Matrix
- Core Test – Common Conditions

Training Conditions

Identifier	*Description*
Core	All speech data, including microphone and telephone channel recordings, available for each target speaker.
Telephone	All telephone channel speech data available for each target speaker. This condition prohibits the use in any way of the microphone data from any of the designated target speakers. Microphone data from speakers other than those specified as target speakers may be used, for example, for background models, speech activity detection models, etc.
Microphone	All microphone channel speech data available for each target speaker. This condition prohibits the use in any way of the telephone data from any of the designated target speakers. Telephone data from speakers other than those specified as target speakers may be used, for example, for background models, speech activity detection models, etc.

Test Segment Conditions

Identifier	*Description*
Core	One two-channel excerpt from a telephone conversation or interview, containing nominally between 20 and 160 seconds of target speaker speech. Some of these test segments will have additive noise imposed.
Extended	The test segments will be the same as those used in Core. The number of trials in Extended tests will exceed the number of trials in Core tests.
Summed	A summed-channel excerpt from a telephone conversation or interview, containing nominally between 20 and 160 seconds of target speaker speech formed by sample-by-sample summing of its two sides.
Known	The trial list for the known test segment condition will be the same as in Extended. The system should presume that all of the non-target trials are by known speakers.
Unknown	The trial list for the unknown test segment condition will be the same as in Extended. The system should presume that all of the non-target trials are by unknown speakers.

Evaluation Test Matrix

		Training Condition		
		Core	**Microphone**	**Telephone**
Test Segment Condition	**Core**	**required**	optional	optional
	Extended	optional	optional	optional
	Summed	optional	-	-
	Known	optional	-	-
	Unknown	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

Common Conditions

A “common condition” is a subset of the evaluation trials that have some given property

The *purpose* of the common conditions is to encourage participants to focus research efforts

Core Test – Common Conditions

Five common conditions were specified for SRE12

They were all trials involving multiple segment training and

1) interview speech in test without added noise in test
2) telephone channel speech in test without added noise in test
3) interview speech in test with added noise in test
4) telephone channel speech in test with added noise in test
5) telephone channel speech intentionally collected in a noisy environment* in test and without added noise

*this was self-reported by the speaker

Numbers of Trials

Common Condition	Core (target / known non-target / unknown non-target)	Extended Trials (target / known non-target / unknown non-target)
1	2,897 / 46,601 / 61,871	3,860 / 10,985,377 / 11,349,426
2	7,354 / 445,041 / 105,196	7,354 / 10,312,118 / 2,088,834
3	3,851 / 49,032 / 20,048	5,127 / 12,444,672 / 4,804,500
4	7,176 / 411,843 / 4,872	7,176 / 9,471,219 / 124,830
5	3,883 / 209,532 / 2,406	3,883 / 5,119,130 / 77,745

Participating Sites and Systems

System Identifier	Site	Location
ABC	Agnitio	Spain/ S. Africa
	Brno University of Technology	Czech Republic
	CRIM	Canada
ANHYT	ANHYT	China
ATIP	Advanced Technologies for Information Processing	Germany
ATVS-QUT	Universidad Autonoma de Madrid	Spain
	Queensland University of Technology	Australia
CCNT	Zhejiang Unviersity	China
CGT	3M Cogent	USA
CLARKSON	Clarkson University	USA
CPQD	CPqD	Brazil
CRSS	University of Texas at Dallas	USA
DELTANCU	Delta Electronics, Inc.	Taiwan
	National Central University	Taiwan

Participating Sites and Systems

System Identifier	Site	Location
FIUPM	Polytechnic University of Madrid	Spain
GTTSE	University of the Basque Country	Spain
HKPU	Hong Kong Polytechnic Unviversity	Hong Kong
I3A	University of Zaragoza	Spain
I4U	Institute for Infocomm Research	Singapore
	University of Eastern Finland	Finland
	Radboud University Nijmegen	Netherlands
	CRSS, University of Texas at Dallas	USA
	ValidSoft Ltd	United Kingdom
	LIA, University of Avignon	France
	Idiap Research Institute	Switzerland
	Swansea University	United Kingdom
	University of New South Wales	Austrailia

Participating Sites and Systems

System Identifier	Site	Location
IBM-HAIFA	IBM Haifa	Israel
ICSI	International Computer Science Institute	USA
IDIAP	IDIAP Research Institute	Switzerland
IFLYUSTC	University of Science and Technology	China
IIR	Institute for Infocomm Research	Singapore
IISAS	Institute of Information Science, Academia Sinica	Taiwan
IITG	Indian Institute of Technology, Guwahati	India
IITH	Indian Institute of Technology, Hyderabad	India
IITMADRAS	Indian Institute of Technology, Madras	India
IOACAS	Institute of Acoustics, Chinese Academy of Science	China
LIA	University of Avignon	France
LIMSIVR	LIMSI	France
	Vocapia Research	France

Participating Sites and Systems

System Identifier	Site	Location
MANU	University of Manchester	United Kingdom
MITLL	MIT Lincoln Laboratory	USA
	MIT Computer Science and Artificial Intelligence Laboratory	USA
	Johns Hopkins University	USA
MJRC	Media Joint Research Center, Tsinghua University	China
NPT	Nuance	USA
NPT	Politecnico di Torino	Italy
NTUT	National Taipei University of Technology	Taiwan
OZU	Ozyegin University	Turkey
RUN	Radboud University, Nijmegen	Netherlands
SHDRAGON	Shanghai Dragon Voice	China
SIAT	Shenzhen Institute of Advanced Technology	China
SPMI	Speech Processing and Machine Intelligence Lab , Tsinghua University	China
SRI	SRI International	USA

Participating Sites and Systems

System Identifier	Site	Location
STMSGP	STMicroelectronics Asia Pacific	Singapore
SVID	Speech Technology Center	Russia
TALENTED	Xiamen Talented Software	China
TEC-CMU	Tecnologico de Monterrey	Mexico
	Carnegie Mellon University	USA
TENCENT	Tencent	China
THUEE	Dept. of Electrical Engineering, Tsinghua University	China
TIT	Tokyo Institute of Technology	Japan
TUBSU	Tubitak	Turkey
	Sabanci University	Turkey
UEF	University of Eastern Finland	Finland
UWS	Swansea University	United Kingdom
VLD	ValidSoft	United Kingdom

Participating Sites and Systems

Sites:	58
Systems:	
Identifiers	49
Total	212
core_core	120
core_extended	30 (9 sites)
core_{un}known	2 x 20 (7 sites)
{micro|tele}phone_*	2 x 9 (2 sites)
core_supplemental	3 (2 sites)
core_summed	1