SRI's NIST 2006 Speaker Recognition Evaluation System

Sachin Kajarekar, Luciana Ferrer, Martin Graciarena, Elizabeth Shriberg, Kemal Sönmez, Andreas Stolcke, Gokhan Tur, Anand Venkataraman

SRI International, Menlo Park, CA, USA

Collaborators:

Yosef Solewicz (Bar-Ilan U.), Andy Hatch (ICSI)

And other ICSI team members

Outline

- System overview
  - Acoustic and stylistic systems
  - Improvements since SRE05
- Data issues and analyses
  - Language label confusion and mislabeling
  - Nonnative speakers and noise
- Post-evaluation updates
  - Combiner
  - ASR improvements
  - SRE04 data use
  - Covariance normalization
  - Contribution of stylistic systems
  - Updated system performance
- Summary and conclusions

System Overview

Overview of Submitted Systems

Individual Systems (several improved over last year)

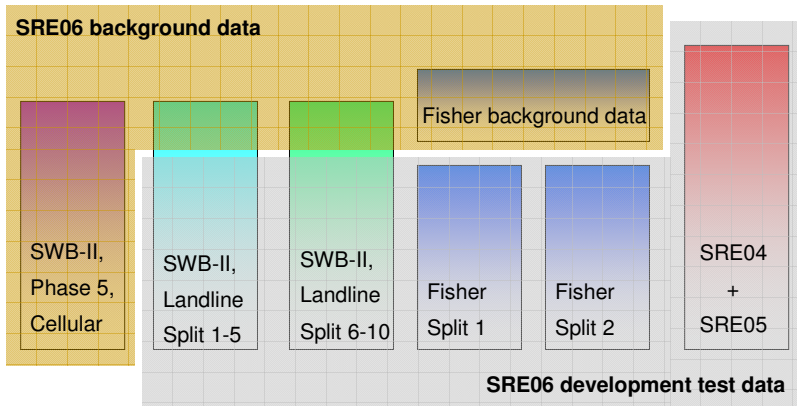
Type	Features	Model	Trials Scored
Acoustic	MFCC	GMM	ALL
	MFCC	SVM	ALL
	Phone-loop MLLR 4 transforms	SVM	Non-English
	Full MLLR 16 transforms	SVM	English-only
Stylistic	State Duration	GMM	English-only
	Word Duration	GMM	English-only
	Word+duration N-gram	SVM	English-only
	GNERFs + SNERFs	SVM	English-only

Submission used all systems with different combiners

Submission	Systems	Combiner
SRI_1 (primary)	SRI (7)	Regression SVM with ABIE
SRI_2	SRI (7)	Regression SVM w/o ABIE
SRI_3	SRI (7) + ICSI (5) + SRI/ICSI (1)	Regression SVM with ABIE

All submissions include results for 1conv4w-1conv4w and 8conv4w-1conv4w

Development Datasets



- Part of SWB-II, landline data was ignored because it had overlap with ASR training data
- TNORM for SRE06 was used from Fisher split 1
- Combiner for SRE06 was trained on SRE04, thresholds estimated on SRE05

Unchanged from SRE05

- **ASR system**
  - 2-pass decoding system (about 3xRT)
  - 2003 acoustic models, no Fisher data used in training
- 3 SID subsystems were used unchanged from last year
- Acoustic: **Cepstral bag-of-frames system**
  - 13 Mel frequency cepstral coefficients (C1-C13) after cepstral mean subtraction
  - Appended with delta, double-delta, and triple-delta coefficients
  - Feature normalization (Reynolds, 2003)
  - 2048-component gender and handset independent speaker independent (SI) model using gender and handset balanced data
  - GMM-UBM model
- Stylistic: **Word** and **state duration** models
  - Duration features extracted from ASR alignments
  - Word-level: vector of word-conditioned phone durations (variable length)
  - State-level: vector of phone-condition HMM state durations (3 per phone)
  - GMM-UBM model
- All system used TNORM for score normalization

Improved Cepstral SVM

- Feature extraction conditioned on 3 broad phonetic categories and 3 HMM states (combination of 8 systems)
  - Phone classes: vowels, glides+nasals, & obstruents
  - Based on ASR alignments
- PCA and PCA-complements features combined
  - Weights trained on Fisher data
- Eliminated mean-only SVM, kept mean-divided-by-stdev SV
- No ASR-conditioning for non-English data

System	SRE05 eng 1-side		SRE05 eng 8-side	
	DCF	EER	DCF	EER
Old cepstral SVM	0.2640	7.12	0.0979	2.91
Phone-conditioned	0.2026	5.33	0.0847	2.52

Improved MLLR SVM

- Removed gender mismatch resulting from ASR gender-ID errors
- Always generate male and female transforms for all speakers, and combine feature vectors
- Non-English data uses MLLR based on phone-loop recognition
- No longer combine phone-loop and full MLLR for English speakers
- For details see Odyssey '06 talk (Friday morning)

System	SRE-05 eng 1-side		SRE-05 eng 8-side	
	DCF	EER	DCF	EER
Old MLLR SVM	0.2487	9.85	0.1181	5.53
New MLLR SVM	0.1770	5.25	0.0818	2.42

Improved Syllable-Based Prosody Model

- Replaced word-conditioned NERFs (WNERFs) with part-of-speech conditioned NERFs (GNERFs) for better generalization
- Switched SVM training criterion from classification to regression
- Reengineered prosodic feature engine for portability and speed (Algemy)
- Changed the binning method from discrete to continuous

System	SRE05 eng 1-side		SRE05 eng 8-side	
	DCF	EER	DCF	EER
Old: SNERF+WNERF	0.5307	14.00	0.2846	6.74
New: SNERF+GNERF	0.4523	11.92	0.1747	4.46

Improved Word N-gram SVM

- Classified instances of words according to pronunciation duration
  - 2 duration bins: "slow" and "fast"
  - Threshold is average word duration in background data
  - Applied to 5000 most frequent words only
- Modeled N-gram frequencies over duration-labeled word tokens
- Gains carried over to combination with GMM word-duration models

System	SRE05 eng 1-side		SRE05 eng 8-side	
	DCF	EER	DCF	EER
Old: word N-gram SVM	0.8537	24.58	0.4878	11.39
New: word+duration N-gram SVM	0.7841	21.12	0.3945	9.40

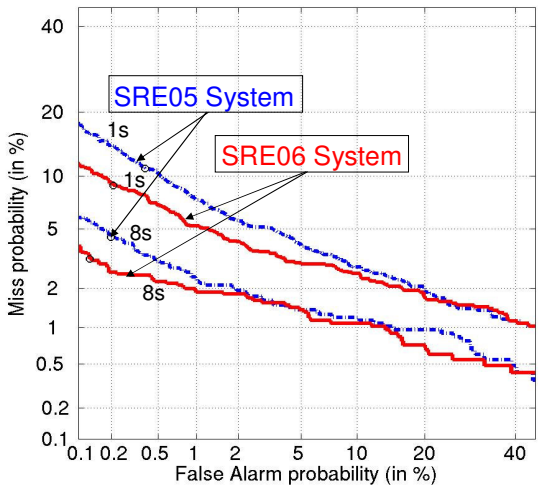
System Combination with Automatic Bias Identification and Elimination (ABIE)

- Based on work by Yosef Solewicz (Bar-Ilan University)
- SVM estimates a bias correction term based on auxiliary features
- Aux features designed to detect training/test mismatch
  - Mean & stdev of cepstrum and pitch
  - Difference of same between training and test
- Trained on samples near the decision boundary of baseline system
- Scaled output of correction SVM is added to baseline score
- Also: gains with regression versus classification SVM

System	SRE05 **CC** 1-side			SRE05 **eng** 1-side		
	Act DCF	Min DCF	EER	Act DCF	Min DCF	EER
SVM-classif combiner	0.1407	0.1062	3.42	0.1476	0.1135	3.62
SVM-regress combiner	0.1278	0.1097	3.47	0.1358	0.1169	3.66
ABIE SVM-regress combiner	0.1280	0.0986	3.18	0.1366	0.1077	3.46

Overall Pre-Eval Improvement on SRE05

(compared to last year's system)



System	SRE05 CC 1-side	
	DCF	EER
SRE05	0.2054	4.35
SRE06	0.1279	3.47
Rel. impr.	38%	20%

System	SRE05 CC 8-side	
	DCF	EER
SRE05	0.0937	1.93
SRE06	0.0598	1.80
Rel. impr.	36%	7%

Data Issues and Analysis

Data Issue: Language Label Confusion

- Initial submission had unexpectedly poor performance
- Major problem found: SRE06 data used language labels in waveform headers that were different from SRE05
  - Documented in email but not in eval plan or on web page
  - Even NIST was confused about the meaning of labels (e.g., "BEN")
- Problem for sites using different systems depending on language!
  - SRI and ICSI systems processed some English data as non-English
  - ASR-based models were not applied to a subset of the trials
  - Other sites not affected because processing was language-independent

System	SRE06 CC 1-side		SRE06 CC 8-side	
	DCF	EER	DCF	EER
Original submission	0.2591	5.15	0.0790	1.78
Corrected submission	0.2220	4.21	0.0634	1.73

- Note: Results scored according to NIST's v2 answer key

Data Issue: Language Mislabeling

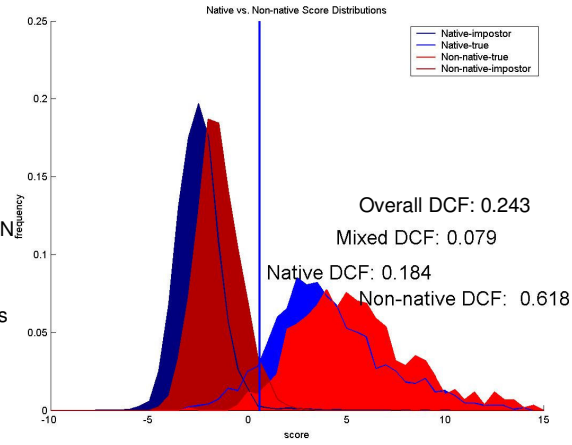
- Corrected submission still had much higher error than on SRE05
- We checked random segments in all conversations. Found errors in language labels: conversations labeled as English were not
- Found 267 conversations NOT in English, 3507 out of 22433 trials affected
- ALL sites could be affected by this
- SRI systems severely affected due to dependence on English-only ASR

Trials	SRE06 1-side		SRE06 8-side	
	DCF	EER	DCF	EER
V2 CC trials as labeled by NIST	0.2220	4.21	0.0634	1.73
V2 CC trials after removing nonEnglish ("sri-eng" from now on)	0.1682	3.54	0.0591	1.73

- Results in next few slides are on this "sri-eng" data set

Data Issues: Nonnative Speakers, Noise

- Listening revealed that majority (53%) of speakers in 1s condition nonnative (NonN)
- ASR looks poor for these talkers
- Trial breakdown in 1s condition (sri-eng) 20% NonN-NonN, 37% mixed, 43% Native-Native
- Score distributions show NonN-NonN trials have systematic positive bias: this destroys the actual DCF
- All systems are affected, but effect is stronger for stylistic systems
- Also noticed in listening: channel distortion and noise



Post-Evaluation Updates

Effect of SVM Combiner

- We originally chose a regression over classification SVM combiner due to marginal improvements on actual DCF on SRE05 (shown earlier)
- Unfortunately, classification was better than regression for SRE06
- Also unfortunately, ABIE combiner did not generalize well to SRE06

System	SRE06 sri-eng 1-side			SRE06 sri-eng 8-side		
	Act DCF	Min DCF	EER	Act DCF	Min DCF	EER
SVM-classif combiner	0.2432	0.1619	3.54	0.0568	0.0561	1.737
SVM-regress combiner	0.2686	0.1698	3.70	0.0652	0.0606	1.737
ABIE SVM-regress combiner	0.3111	0.1697	3.59	0.0728	0.0611	1.737

Effect of ASR System

- As an expedient we originally left ASR system unchanged since SRE04
  - Avoids need to reprocess all of background training data
  - Updated ASR showed only little benefit on development data
- But:
  - State-of-the-art only as of 2003
  - Only ~ 300 hours of Switchboard training data
  - Native English speakers only, poor performance on nonnative speakers
- We compared old ASR to ASR from BBN and from current SRI system
  - Trained on ~ 2000 hours of data, including Fisher
  - Only word hypotheses changed; same MLLR models used in all cases
  - To do: reprocess background data to retrain stylistic systems

	SRE06 sri-eng 1-side		SRE06 sri-eng 8-side	
MLLR SVM system	DCF	EER	DCF	EER
Old SRI ASR	0.2076	4.51	0.0872	2.28
BBN ASR (provided by NIST)	0.1939	4.56	0.0854	2.28
New SRI ASR	0.1887	4.40	0.0837	2.18

Use of SRE04 for Model Training and TNORM

- We were trying to avoid tunning on SRE05 (until it was “officially” allowed); used only SRE04 subset to test effect of SRE04 for background and TNORM
- Found little gain in that work from using SRE04 for background/TNORM
- We should have checked results on SRE05 when NIST allowed its use
- Using SRE04 background and/or TNORM does improve our systems, e.g.:

System	SRE05 sri-eng 1-side		SRE06 sri-eng 1-side	
	DCF	EER	DCF	EER
SNERF+GNERF system				
w/o SRE04 background & TNORM	0.4546	12.1	0.5144	12.34
with SRE04 background & TNORM	0.4373	11.1	0.4529	11.00
MLLR SVM system (new ASR)				
w/o SRE04 background	0.1887	4.40	0.0837	2.18
with SRE04 background	0.1851	4.18	0.0777	2.23

Effect of Within-Class Covariance Normalization

- All leading systems this year applied some form of session variability modeling
  - NAP (Solomonoff et al., Odyssey '04; Campbell et al. ICASSP '06)
  - Factor analysis likelihood ratios (Kenny et al., Odyssey '04)
  - "Modelling Session Variability …" (Vogt et al., Eurospeech '05)
- Similar issue is addressed by WCCN for SVMs
  - Hatch & Stolcke, ICASSP '06; Hatch et al., ICSLP '06
- Official submission only applied WCCN to MLLR system in combined SRI/ICSI (non-primary) submissions
- Plan to apply WCCN (or NAP) to several SVM subsystems

MLLR SVM system	SRE06 sri-eng 1-side	
	DCF	EER
w/o WCCN	0.2076	4.51
with WCCN	0.1845	4.24

Summary of Post-Eval Results

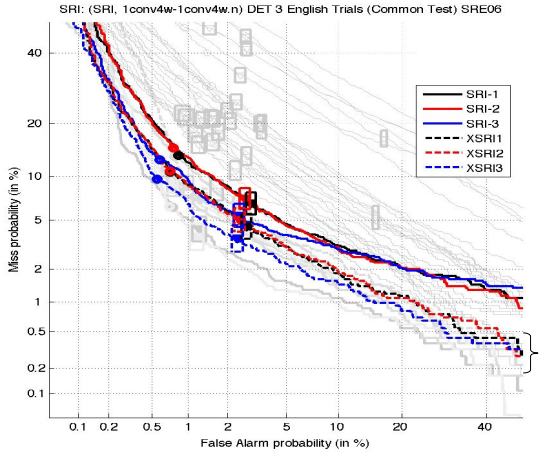
- Step 1: Fixed 'bugs' (processed all English data as English) → 'XSRI'
- Step 2: Improved systems (fixed suboptimal decisions)
  - Use SRE04 data for TNORM and SVM training in prosodic system
  - Use new ASR for MLLR system
  - Use SVM-classification combiner
- Step 3: Start applying WCCN (to MLLR system so far)

Using V4 Answer Key

System	SRE06 CC 1-side			SRE06 CC 8-side		
	Act DCF	DCF	EER	Act DCF	DCF	EER
Original submission (SRI)	0.3571	0.2169	4.75	0.0898	0.0817	1.84
1. Corrected (XSRI)	**0.3164**	**0.1764**	**3.67**	**0.0739**	**0.0670**	**1.74**
2. Improved systems	**0.2557**	**0.1652**	**3.34**	**-**	**-**	**-**
3. Improved + MLLR-WCCN	**0.2562**	**0.1537**	**3.29**	**-**	**-**	**-**

Eval and Post-Eval Results

- Original (SRI) and "corrected" (XSRI)
  (results for SRI1 and XSRI1 = first two rows of previous table)



Corrected (XSRI)

Contribution of Stylistic Systems

(using "improved systems", no WCCN, no new ASR for stylistic)

Systems included in combination	SRE05 eng 1-side		SRE05 eng 8-side	
	DCF	EER	DCF	EER
3 Cepstral	0.1377	4.07	0.05664	2.33
3 Cepstral + 4 Stylistic	0.1139	3.62	0.04774	1.99
Relative Improvement	17%	11%	16%	15%

Systems included in combination	SRE06 sri-eng 1-side		SRE06 sri-eng 8-side	
	DCF	EER	DCF	EER
3 Cepstral	0.1705	3.33	0.06512	1.93
3 Cepstral + 4 Stylistic	0.1597	3.22	0.05544	1.54
Relative Improvement	6%	3%	15%	20%

- Significant improvements from stylistic systems, but less for SRE06 1s
- Why? SRE06 new data:
  - harder for ASR
  - more nonnative speech → greater score shift for stylistic systems. Stylistic have good true/imposter separation, but threshold was off.

Summary and Conclusions

- Substantial improvements since SRE05
  - Cepstral SVM, MLLR SVM, NERFs, word N-grams, combiner
  - Overall ~ 36% lower DCF on SRE05 data
- But various problems this year, from bugs to suboptimal choices
- In addition, language labels were a moving target
- New SRE06 data appears much harder for ASR (nonnative speakers, noise), affecting many of our systems
- Nonnative speakers present interesting challenges for SID
  - Add to training data
  - Score distributions suggest separate modeling
- Current post-eval results show our DCF for 1s CC is reduced by:
  - 19% relative to "corrected submission" (XSRI)
  - 28% relative to buggy submission (SRI)
- Expect further gains from improved ASR, and session variability normalization in all relevant systems.