THE SRI NIST 2008 SPEAKER RECOGNITION EVALUATION SYSTEM

Sachin S Kajarekar[1]*, Nicolas Scheffer*[1]*, Martin Graciarena*[1]*, Elizabeth Shriberg*[1,2]*, Andreas Stolcke*[1,2]*, Luciana Ferrer*[1,3]*, Tobias Bocklet*[1,4]

[1]SRI International, Menlo Park, CA, USA
[2]ICSI, Berkeley, CA, USA
[3]Stanford University, CA, USA
[4]University of Erlangen-Nürnberg, Germany

ABSTRACT

The SRI speaker recognition system for the 2008 NIST speaker recognition evaluation (SRE) incorporates a variety of models and features, both cepstral and stylistic. We highlight the improvements made to specific subsystems and analyze the performance of various subsystem combinations in different data conditions. We show the importance of language and nativeness conditioning, as well as the role of ASR for speaker verification.

Index Terms— speaker recognition, prosody, speech recognition

1. INTRODUCTION

The NIST SRE 2008 evaluation was both more complex and more data-intensive than in prior years. Three different types of speech signals were used in the required common condition: conversational telephone speech, conversational speech over auxiliary microphones, and interview speech over auxiliary microphones. This paper describes the SRI submission to NIST SRE-08. We describe the main improvements to the individual subsystems from SRE-06, our combination strategy, the SRI automatic speech recognition (ASR) system, and the new nativeness classifier. Overall results show that a combination of four systems results in the same performance as the submission on SRE-08 data. The results also show significant improvements in performance with language compensation. The SRI submission was one of the best-performing in conditions involving telephone data.

2. COMMONALITIES

2.1. Development Data

Background data for Gaussian mixture model (GMM) systems, impostor data for support vector machine (SVM) systems, and the data for score normalization were drawn from the SRE-04 and SRE-05 altmic data sets.

Most of the systems also used the SRE-04 and SRE-05 alternate microphone (altmic) data for estimating directions for within-speaker variations or channel factors. SVM systems (except those using supervector features) use nuisance attribute projection (NAP) [1] to estimate within-speaker variation, and the resulting eigenvectors are estimated using SVDLIBC. GMM systems use factor analysis [2] to estimate channel factors. System-specific details are described below.

2.2. Automatic Speech Recognition System

The ASR system is similar to that used in SRE-06, with modifications for the SRE-08 data. The acoustic models were trained on Switchboard and Fisher Phase 1 data (with additional text and web data for language model training). Extra weight was given to nonnative Fisher training data to achieve more balanced performance on nonnative speakers. The system ran in real time on a 4-core 2.6 GHz AMD Opteron machine. The word error rate (WER) on transcribed portions of the Mixer corpus was 23.0% for native speakers and 36.1% for nonnatives. Nontelephone (microphone) data was preprocessed with the ICSI/Qualcomm Aurora Wiener filter implementation, and then recognized with the telephone ASR system. The WER measured on SRE-06 altmic data (transcribed at International Computer Science Institute (ICSI)) was 28.8%.

3. SYSTEM INNOVATIONS

Table 1 lists the systems used in our submission. Systems on shaded rows did not use any information from ASR. Systems marked with (*) are either new or were redesigned since SRE-06. These systems are described further in this section. See [3, 4] for descriptions of systems that were unchanged from prior years.

Table 1 Individual Systems

FEATURE	Model	Use ASR
MFCC (Standard)	GMM-LLR	No
Constrained GMM*	GMM-LLR	Yes
Polynomial Cepstrum, MFCC	SVM	No
Polynomial Cepstrum, PLP	SVM	No
Supervector Cepstrum, MFCC*	GMM-SVM	No
Supervector Cepstrum, PLP*	GMM-SVM	No
MLLR Transform Phoneloop*	SVM	No
MLLR Transform English ASR	SVM	Yes
Word N-Grams	SVM	Yes
State-In-Phone Durations	GMM-LLR	Yes
Phone-In-Word Durations	GMM-LLR	Yes
Supervector prosodic polynomial*	GMM-SVM	No
GMM weight prosodic polynomial *	SVM	No
GMM weights SNERFs*	GMM-SVM	Yes

3.1. Cepstral GMM Systems

3.1.1. Standard system

This system differs from the system used in SRE-06 in only two respects. It uses13 cepstral coefficients (C0-C12), and the scores are normalized using gender-dependent TZnorm.

3.1.2. Constrained system

A new, "constrained" cepstral GMM system makes use of automatic syllabification of phone alignments from ASR. The constrained system combines scores from eight subsystems, each of which uses features computed as for the standard system, but restricted to only those frames that satisfy a specific constraint. The eight constraint specifications are (1) syllable nuclei, (2) syllable onsets, (3) syllable codas, (4) syllables containing the phone (/n/), (5) containing the phone (/t/), (6) containing any the phones /b/, /p/, /v/, or /f/, (7) one-syllable words, and (8) syllables following pauses. Constraints are chosen based on results for SRE-06 1conv training data. Background models were trained on SRE-04 English telephone data (for lack of time, no altmic data was used). A 512-component GMM was used in every subsystem except the constrained subsystems (5) and (8), which used 1024 Gaussians. Eigenchannel matrices for each constrained subsystem were trained using data from SRE-04 and SRE-05 altmic data. The rank of these matrices was set to 50; the number of EM iterations was 5. GMM likelihood ratio scores were ZT-normalized using data from SRE-04. The resulting scores were combined using logistic regression by training on SRE-05 1conv4w telephone trials. The constrained system was used for 1-conversation training experiments only.

3.2. Cepstral SVM Systems

SRI submitted two different cepstral GMM-supervector (SV) systems, differing only in their front-end processing: an MFCC-based system and a PLP-based system.

Both GMM-SV systems were gender dependent. Both systems use two gender-dependent 1024 GMM models and employed the factor analysis framework to compensate for intersession (and/or interspeaker) variability [2]. Scores were normalized using ZT-norm. Universal background models (UBMs) were trained on SRE-04 and SRE-05 altmic data. Data for T-norm and Z-norm, unless noted are from the same pool of data. SVM training and classification is performed using the LIBSVM toolkit. The background data is very similar to the UBM training data and has approximately 2000 impostor examples per gender. To increase robustness of the system to different types of data, channel eigenvectors estimated on different databases were stacked.

The Mel frequency cepstral coefficient (MFCC) front end was similar to that of the GMM-LLR system. However, the feature dimension was reduced using a combined linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT) in the same way as for the MLLR (maximum likelihood linear regression)-SVM cepstral features. The rank of the channel factor loading matrix was 150. It was composed of a mixture of variances estimated on SRE-04 data (50), SRE-05 altmic data (50), and Switchboard-II Phases 2, 3 and 5 (50).

The perceptual linear prediction (PLP) system used the same front end as the MLLR-SVM system. The rank of the channel factor loading matrix was 160. It was composed of a mixture of variances estimated on SRE-04 data (80) and SRE-05 altmic data (80). Here, score normalization data omitted SRE-05 altmic to differentiate the system from the MFCC-based version.

3.3. MLLR SVM Systems

The MLLR-SVM systems used speaker adaptation transforms as features for speaker verification . The MLLR reference models used 52-dimensional PLP features normalized and reduced to 39 dimensions with VTLN, LDA+MLLT, and a speaker-adaptive constrained MLLR (SAT) transform. A total of 16 MLLR affine 39x40 transforms maped the Gaussian mean vectors from speaker-independent to speaker-dependent speech models; eight transforms each were estimated relative to male and female reference models,. Each feature dimension was rank normalized and is then subjected to NAP estimated on SRE-04 and SRE-05 altmic data, using 32 nuisance dimensions. The projected feature vectors were then modeled by SVMs using a linear kernel. We use the SVM^{light} toolkit to learn SVMs and classify instances. The impostor set for SVM training comes from SRE-04. No score normalization was applied.

For English-language conversations, the MLLR estimation used word hypotheses from the first ASR pass as the speech model. For non-English conversation sides, MLLR was performed using a phoneloop speech model; however, the phoneloop still used English phones only. Unlike in past submissions, the cepstral features and number of transforms used by the phoneloop MLLR system were the same as for the ASR-based MLLR system. This change (from MFCC to PLP, and from 2 to 8 phone classes) reduced the minimum decision cost function (DCF) on SRE-06 English data from 1.7 to 1.2 times that of the ASR-based MLLR-SVM. The number of NAP nuisance dimensions for the phoneloop MLLR system is 64.

3.4. Prosodic Systems

Pitch and energy feature signals for each conversation side were obtained using the get_f0 code from the Snack toolkit. The waveforms were preprocessed with a bandpass filter (250-3500 Hz) to make the spectral contents of all channels similar to that of the telephone channel. These signals were used to extract prosodic features in a variety of ways.

3.4.1. ASR-independent prosodic features

This ASR-independent prosodic system used the features described in [5]. Pseudo-syllable regions were defined as the regions between consecutive local minima in the energy signal over voiced regions. For each region, polynomial approximations of order 0 to 5 were obtained for the energy and the pitch signals. The resulting coefficients along with the duration of the region were used as features. The following two modeling techniques are used:

(a) GMM Supervector modeling:

The Joint Factor Analysis (JFA} framework was used (as in [5]) for the prosodic features. The JFA algorithm here was used as a feature extractor, as the same process was applied to both training and test utterances. The SVM kernel used the speaker component (eigenvoice and diagonal model) as an input feature. Gender-dependent 256-mixture GMM models were used to model the prosodic feature distribution. A mean-variance normalization process was applied beforehand as a preprocessing step. The speaker and channel factor loading matrices had ranks of 70 and 50, respectively, and were trained on Fisher Phase 2, SRE-04, and SRE-05 altmic data. Scores were normalized using Tnorm.

(b) Weight modeling:

The method described in [6] is used to transform the features into a single 12,630-dimensional vector. Transformed vectors are rank normalized [7] and 16 NAP directions subtracted before training an SVM for regression on the class labels. Signed distances to the resulting hyperplanes are used as scores, which are further normalized using TZnorm.

3.4.2. ASR-dependent prosodic features

For English trials, two more sets of features were computed: all-syllable features and grammar-constrained wordlist features. In the first set, SNERFs (syllable NERFs, nonuniform extraction region features) were extracted from all syllables, regardless of word identity. Syllables were automatically obtained from ASR phone-level output. Features reflect characteristics about the pitch, energy, and duration patterns inside the syllable (for details on the features, see [8]). The second set of features, GNERFS (grammar-constrained NERFs), used the same prosodic features at the syllable level but constrains extraction location to 16 wordlists.

We provide the resulting features from both sets, along with the polynomial features described above, to an SVM, after performing rank normalization and subtracting 32 NAP directions. The scores obtained from the SVM were normalized using TZnorm.

4. SRE-2008 EVALUATION DATA

The NIST SRE-2008 evaluation protocol included six training data conditions and four test conditions. One of the combinations of these conditions is called the *common evaluation condition* and all participants had to submit scores for this condition. This condition used about 2.5 minutes of data each for training and testing and had eight different subsets for scoring purposes. Table 3 shows how we mapped SRE-06 data conditions to SRE-08 for development purposes; no interview development data was used.

SRI participated in two conditions – short2-short3 and 8conv-short3. The first condition included both telephone and interview data where each recording was about 2.5 minutes. The second condition included 8 telephone conversation sides for training and one 2.5-minute telephone or interview session for testing.

5. COMBINATION PROCEDURE

Table 1 summarizes the systems used for combination. The combination of systems was performed using the method described in [9]. The auxiliary information used here (for English data only) is given by a nonnativeness classifier [10] for English speakers.

A separate combiner was trained for each condition listed in Table 2. For other conditions, a simple logistic regression classifier is used. The primary SRI submission (SRI_1) was the combination of all 14 systems presented in Table 1. This system used ASR for English trials. As a contrastive system, SRI_2 was composed of the 8 ASR-independent systems only.

6. RESULTS

We present the results of the SRI 2008 submission and other combinations for common conditions (CC) 7 and 6. The focus is on the noninterview data conditions, since our development effort was limited to telephone and altmic speech.

6.1 Telephone Conversations in English (CC=7)

This condition is the traditional common condition of previous NIST SRE evaluations and thus matched data was available as a development set (Table 2, Row 1). Table 3 presents the results of SRI's primary and secondary systems SRI_1 and SRI_2. The improvement from using nativeness compensation is also reported. Because of the large number of systems in the submission, we also report the performance of the four best systems for this condition, as well as the four cepstral systems alone. Thus, performance can be compared with that of other sites, where the number of subsystems is typically lower and comprised of cepstral models only.

Table 2 Development and Evaluation datasets (shaded rows represent the English-only datasets)

SRE-06 (development data)			SRE-08 short2-short3	
Train	Test	#Trials	Train	Test
1conv4w	1conv4w	23678	Conv, phn	Conv, phn
1conv4w	1convmic	22715	Conv, phn	Conv, mic
1convmic	1conv4w	19223	Intrv, mic	Conv, phn
1convmic	1convmic	132341	Intrv, mic	Intrv, mic
1conv4w	1convmic	22715	Conv, phn	Intrv, mic
1conv4w	1conv4w	27381	Conv, phn	Conv, phn
1convmic	1conv4w	1703	Conv, phn	Conv, mic
1convmic	1conv4w	1703	Intrv, mic	Intrv, phn
1convmic	1conv4w	1703	Conv, phn	Intrv, mic

Table 3 Results for SRE-08 CC=7 and SRE-06 1conv4w-1conv4w common condition (shaded results use nativeness information)

System/	SRE-06		SRE-08 (CC=7)		
Combination	minDCF	%EER	actDCF	minDCF	%EER
1-BEST	0.072	1.192	0.134	0.132	2.769
4-BEST	0.048	0.921	0.104	0.101	1.954
SRI_1 (14)	0.048	0.867	0.106	0.100	2.117
4-CEP	0.059	1.083	0.106	0.103	2.199
SRI_2 (8)	0.063	1.192	0.113	0.107	2.199
SRI_1 (14)	0.052	0.867	0.108	0.102	2.199

For this condition, the constrained GMM system (1-BEST) presents better performance than all other systems. While an improvement is observed on SRE-06 by combining all 14 systems compared to the four best systems; this improvement is not found on SRE-08. The 4-BEST systems for this condition are Constrained GMM, GMM-LLR, GMM-PLP, and GMM weight prosodic polynomial. It is also clear that the four cepstral systems play an important role in the submission, as the 4-CEP results on SRE-08 are comparable to both SRI primary and secondary systems.

6.2. Telephone Conversations in All Languages (CC=6)

This condition includes trials involving different languages in both training and testing (Table 2, Rows 1 and 6). For the submission, two combiners were trained depending on the language in the trial. We mapped English-only trials onto one category, and the rest onto another (i.e., non-English data in training or testing, as well as mixed trials). Our hypothesis was that the distributions of trials with non-English data will be a unimodal Gaussian distribution that would be different from the distribution of English trials. The results for this condition are presented in Table 4.

After the submitting the results, we realized that the overall distribution of scores is bimodal in a way that is different from what we had hypothesized. The distribution of trials with non-English data on either side completely overlaps with the distribution of English-only trials to form one mode, and the remaining trials form the other mode. We approached this problem by changing the number of combiners trained for each condition. Indeed, instead of considering only two different types of trials (English-only versus the rest), we now consider four different types of trials (English-only, English training only, English testing only, non-English only).

Table 4 Results with SRE-08 CC=6 and SRE-06 1conv4w-1conv4w (shaded results use nativeness compensation)

System/ Combination	SRE-06		SRE-08		
	minDCF	%EER	actDCF	minDCF	%EER
4-CEP	0.140	2.821	0.547	0.408	7.095
SRI_1 (14)	0.124	2.574	0.503	0.372	6.834
SRI_2 (8)	0.137	2.738	0.538	0.397	6.871

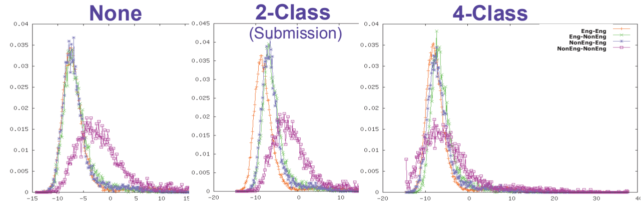

Figure 1 Histograms of scores for different trials. Red (+) is Eng-Eng, magenta (□) is NonEng-NonEng, green (X) is Eng-NonEng and blue (*) NonEng-Eng trials

Table 5 shows that the results are significantly improved after revised language compensation. The table also shows that calibration (difference between actual and minimum detection cost function [DCF]) has significantly improved on SRE-08 data.

Table 5 Results from Table 4 after revised language compensation (shaded results use nativeness compensation)

System/ Combination	SRE-06		SRE-08		
	minDCF	%EER	actDCF	minDCF	%EER
4-CEP	0.116	2.378	0.310	0.276	5.303
SRI_1	0.110	2.015	0.317	0.274	5.302
SRI_2	0.113	2.185	0.309	0.279	5.228

7. CONCLUSIONS

The primary submission by SRI for SRE-08 was composed of 14 systems with eight ASR-independent systems and six ASR-dependent systems. A contrast system used only the ASR-independent subsystems. The main improvements were an new ASR-constrained cepstral system, use of the output of a nativeness detector as side information for the combiner, and the improvement of our ASR system with a significant amount of nonnative English data. In addition, the robustness of the cepstral system on gender and on alternate microphones improved substantially. Results show that the combination of four cepstral systems has about the same performance as a combination of the four best systems and as the overall submission. The language calibration was important for 2008, and performance improved when appropriate language classes were chosen for the trials. Summarizing overall results, the SRI submission was among the best submissions in SRE-08 common condition trials with telephone data. Interview data was used for the first time in this evaluation and data segmentation issues played an important role. Although we used systems developed on telephone data, on interview data our submission was competitive with those of other sites that used the NIST voice activity detector and did not use the small interview development set provided for training.

ACKNOWLEDGMENTS

This work was funded through a development contract with Sandia National Laboratories (#DE-AC04-94AL85000). The views herein are those of the authors and do not necessarily represent the views of the funding agencies.

REFERENCES

[1] A. Solomonoff, C. Quillen, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," presented at ICASSP, Philadelphia, USA, 2005.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," presented at ICASSP, Philadelphia, PA, 2005.

[3] S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST speaker recognition evaluation system," presented at ICASSP, Philadelphia, 2005.

[4] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," presented at Eurospeech, Geneva, 2003.

[5] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 2095-2103, 2007.

[6] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sonmez, "Parameterization of prosodic feature distribution for SVM modeling in speaker recognition," presented at ICASSP, Honolulu, 2007.

[7] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric Feature Normalization for SVM-based Speaker Verification," presented at ICASSP, Las Vegas, 2008.

[8] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication*, vol. 46, pp. 455-472, 2005.

[9] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System Combination using Auxiliary Information for Speaker Verification," presented at ICASSP, Las Vegas, 2008.

[10] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, and A. Stolcke, "Detecting Nonnative Speech Using Speaker Recognition Approaches," presented at Odyssey: A speaker and language recognition workshop, Stellenbosch, South Africa, 2008.