THE SRI NIST 2010 SPEAKER RECOGNITION EVALUATION SYSTEM

Nicolas Scheffer, Luciana Ferrer, Martin Graciarena, Sachin Kajarekar, Elizabeth Shriberg, Andreas Stolcke

SRI International, Menlo Park, CA, USA

ABSTRACT

The SRI speaker recognition system for the 2010 NIST speaker recognition evaluation (SRE) incorporates multiple subsystems with a variety of features and modeling techniques. We describe our strategy for this year's evaluation, from the use of speech recognition and speech segmentation to the individual system descriptions as well as the final combination. Our results show that under most conditions, the cepstral systems tend to perform the best, but that other, non-cepstral systems have the most complementarity. The combination of several subsystems with the use of adequate side information gives a 35% improvement on the standard telephone condition. We also show that a constrained cepstral system based on nasal syllables tends to be more robust to vocal effort variabilities.

Index Terms— speaker recognition, prosody, high-level modeling, system fusion.

1. INTRODUCTION

The NIST SRE 2010 evaluation introduced several challenges compared to earlier SREs. In addition to the variability in speech genre and microphones found in SRE08, the SRE10 core evaluation condition included speech samples of varying lengths, and vocal effort was introduced as a dimension of intrinsic variability. Moreover, the decision cost function (DCF) was redefined to favor a new operating point aimed at lower false alarm (FA) rates. The new cost function (termed "newDCF" here) made each FA 1000 times more costly than a miss error. (We use "oldDCF" to refer to the SRE08 definition for which the cost ratio was only 10 to 1.) In order to achieve stable results at very low false alarm rates, an *extended evaluation set* was defined, containing an order of magnitude more trials than the original set, or about 6 million. The large evaluation set in turn made efficient scoring methods a necessity.

SRI submitted two systems to SRE10. SRI_1 is a static score-level fusion of three cepstral Gaussian mixture model (GMM) systems, one system based on maximum likelihood linear regression (MLLR) transforms, one prosodic system, and one word N-gram system. SRI_2 is an enhanced system that adds a constrained cepstral GMM system, as well as a score combiner that uses signal-to-noise ratio and amount of detected speech as side information. The SRI submissions were among the best-performing systems in SRE10.

2. COMMONALITIES

This section describes aspects of our system that were common to all speaker modeling subsystems, as well as the system fusion approach.

2.1. Design of the development set

A set of 82 interview speakers (48 females and 34 males) from SRE08 (both original and follow-up evaluation) was set aside as additional training data. A development set was created using the remaining SRE08 data. For each original condition from SRE08, an extended set was created by pairing every available model with every available test sample (except where model training and test sample came from the same recording session). No additional models were created, and only samples originally used for testing were used for testing in the extended development set.

Here, we use the following notation for the trial conditions: *trainDuration-testDuration.trainStyle-testStyle.trainChannel-testChannel*, where

- Duration: short (shrt) or long (long)
- Style: telephone (tel) or interview (int)
- Channel: telephone (phn) or alternate microphone (mic)

For the shrt-shrt.tel-tel.phn-phn condition, the target trials were restricted to be the target trials as defined by NIST. Trials for the short-long condition (not found in SRE08) were created by using the long-long condition and replacing the training data with a long sample from the same speaker using the same microphone. Table 1 gives a summary of the created trials by condition, as well as the mapping to SRE10 conditions. This mapping, which in some cases constituted an imperfect match, was used for to training combination parameters for SRE10 based on data from SRE08. Note that we redefined "long" as an utterance of up to eight minutes of interview speech, to match the SRE10 condition, using the first eight minutes of the long samples in SRE08.

2.2. Waveform preprocessing and segmentation

For both telephone and microphone recordings, utterances were segmented into short segments containing mostly speech, using a speech/nonspeech Hidden Markov Model (HMM) decoder and various duration constraints. For interview recordings, we used a more complex algorithm to suppress cross-talk from the interviewer's speech. The algorithm incorporated elements from LPT's 2008 processing [1], using the NIST-provided automatic speech recognition (ASR) output for interviews. The steps were as follows.

1. Segment the interviewee channel into speech segments according to the NIST ASR output.
2. Segment the interviewee channel with speech/nonspeech models trained on distant-microphone meeting speech (from our NIST RT-07 evaluation system), and remove regions that have ASR output for the interviewer.
3. Intersect the segments from steps 1 and 2.
4. Choose segmentation from step 3 if it comprises at least 40% of the original waveform duration; or, use output from step 1

S. Kajarekar is now with Cisco Systems, Inc.

Table 1. Development conditions, the number of trials, and the SRE10 conditions used as training data for the combiner.

DEV Condition	# target trials	# impostor trials	SRE10 conditions *(* means any value for that setting)*
long-long.int-int.mic-mic	9,774	319,956	long-long.int-int.mic-mic
long-shrt.int-int.mic-mic	32,248	1,054,592	long-shrt.int-*.mic-mic
long-shrt.int-tel.mic-phn	1,362	754,729	long-shrt.int-tel.mic-phn
shrt-long.int-int.mic-mic	10,234	336,437	shrt-long.int-int.mic-mic
shrt-shrt.int-int.mic-mic	33,743	1,108,882	shrt-shrt.*-*.mic-mic
shrt-shrt.int-tel.mic-phn	1,459	797,812	shrt-shrt.int-tel.mic-phn
shrt-shrt.tel-tel.phn-phn	1,108	1,453,237	shrt-shrt.tel-tel.phn-phn

if it comprises at least 40% of the original waveform; otherwise, use output from step 2.

5. Merge segments separated by no more than 0.5s and pad with 0.04s at the start and end of the merged segments.

2.3. Speech recognition system

Several of the speaker verification models described below relied on word and sub-word recognition hypotheses obtained by ASR. We used a fast version of SRI's conversational telephone recognition system with modifications for the SRE data. With this method, the first recognition pass generated lattices using a bigram Language Model (LM) and acoustic models based on MFCC features with fMPE transforms, augmented with MLP phone posterior features. The lattices were then rescored with a 4-gram LM. A second recognition pass used speaker-adapted fMPE-PLP models, generating N-best lists that were then further rescored with pronunciation and duration models. 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 word error rate on the transcribed portions of the Mixer corpus was 23.0% for native speakers and 36.1% for nonnatives. Non-telephone (microphone) data was preprocessed with the ICSI/Qualcomm Aurora Wiener filter implementation, and then recognized with the telephone ASR system. The word error rate measured on SRE06 alternate microphone data (transcribed at ICSI) was 28.8.

2.4. Scoring mechanism

To deal with the extremely large trial set in SRE10, as well as with the development set of similar size, all the systems described in Section 3 used a dot-product-like approach for computing verification scores. We found that regardless of the number of training and testing utterances, performing the full matrix scoring (i.e., scores from all models against all test utterances) was always faster. The use of optimized linear algebra libraries, such as BLAS, was critical to that end. All systems based on the JFA paradigm used the fast likelihood computation described in [2]. The support vector machine (SVM) systems all used a linear kernel, which can be evaluated as a dot product by appending the model hyperplane offset to the hyperplane vector, and a constant 1 to the test feature vector.

2.5. System combination

The combination of systems was performed using linear logistic regression separately for each condition, as given in Table 1. In addition, the SRI_2 system use the method proposed in [3] to use side-information, specifically to compensate for score biases as a function amount of speech and SNR of the signals. SNR was computed on each session and thresholded at 15 dB to generate two categories: "low" and "high" SNR. Similarly, the number of words in the session was obtained by ASR and thresholded at 200 to generate two categories: "short" and "long" sessions. Finally, the category for each trial was define as the cross-product of the word-count and SNR categories for the training and test sessions, creating a total of 16 possible categories. A regularization parameter encouraged small category-dependent weights. After combination, the scores were assumed to be calibrated likelihood ratios. Therefore, a fixed threshold, given by the theoretically optimal value for the target DCF, was used to make hard detection decisions.

3. INDIVIDUAL SYSTEM DESCRIPTIONS

In this section we describe the component subsystems and their associated speaker modeling approaches.

3.1. Cepstral GMM-JFA systems

3.1.1. Standard cepstral system: cep

The cepstral GMM system used a 300-3300 Hz bandwidth frontend, consisting of 24 Mel filters to compute 20 cepstral coefficients and their delta and double delta coefficients, producing a 60 dimensional feature vector. The resulting features were mean and variance normalized over the utterance. The feature vectors were modeled by a 1024-component, gender-independent GMM. The background GMM was trained using data from the SRE04, SRE05, and SR08 development data. We used a full Joint Factor Analysis model (JFA) in which 600 eigenvoices were trained using SRE data from 2004 and 2005, and the Switchboard-II corpus. By training two subspaces separately on telephone and interview data, 500 eigenchannels were obtained. The diagonal term was trained with the same data as speaker factors. Scores were normalized using gender-dependent ZTnorm.

3.1.2. Nasal syllable constrained cepstral system: nasals

The SRI submission contained a single constrained cepstral system [4] that uses features computed as in 3.1.1 but restricted to frames occurring in syllables that contained the recognized phone [n] or [ng]. Syllables were based on an automatic maximum-onset-based cross-word syllabification of ASR output. The resulting frames comprised about 18% of the total speech-aligned frames used in the standard system. UBM, JFA parameters, and score normalization techniques were the same as for the standard *cep* system.

Table 2. Results of the SRI submission and subsystems on the required core conditions of NIST SRE 2010 extended set. Results are given as newDCF/oldDCF

# Condition	cep	plp	foc	mllr	ngram	nasals	pros	SRI_1	SRI_2
1 Int SameMic	**.43/.09**	.67/.16	.50/.10	.58/.26	1./.93	.67/.27	1./.60	.36/.06	.33/.05
2 Int DiffMic	**.51/.13**	.71/.24	.61/.17	.68/.35	1./.95	.81/.41	1./.71	.44/.10	.42/.10
3 Int Tel	**.47/.14**	.61/.21	.61/.18	.54/.26	1./.95	.84/.38	1./.74	.30/.10	.28/.08
4 Int Mic	**.39/.11**	.51/.17	.46/.13	.50/.23	1./.94	.67/.28	1./.56	.24/.07	.22/.07
5 Tel Tel	.47/.14	**.47/.14**	.57/.16	.47/.18	1./.90	.73/.32	.99/.62	.31/.09	.29/.08
6 Tel High Vocal Effort	.83/.26	**.80/.24**	.86/.30	.99/.32	1./.91	.91/.48	1./.88	.72/.17	.71/.17
7 Mic High Vocal Effort	.90/.24	.88/.33	.87/.28	.88/.35	.99/.87	**.86/.40**	1./.91	.87/.23	.81/.21
8 Tel Low Vocal Effort	**.45/.11**	.53/.13	.62/.15	.65/.17	1./.90	.76/.33	.99/.65	.33/.07	.32/.07
9 Mic Low Vocal Effort	**.27/.06**	.39/.11	.33/.07	.31/.11	1./.89	.67/.19	.95/.38	.17/.05	.17/.04

3.1.3. Class-dependent cepstrum: foc

This second MFCC-based system differed with respect to *cep* by its use of gender-dependent UBM models, eigenvoices, and eigenchannels, and by eliminating the JFA diagonal term. Another difference was that the ZTnorm process was condition-dependent, in the sense that normalization data was matched to the target testing condition. For example, for the long-shrt.int-tel.mic-phn condition, Tnorm used only short telephone data, while Znorm used only interview data. This approach proved useful in all conditions except for the tel-tel condition, for which it seemed that the more data, the better the result. Condition-dependent eigenvoices and eigenchannels gave no gains.

3.1.4. Class-dependent PLP-SAT cepstrum: plp

This system used the exact same setup as *foc*. However, the input features were generated by the PLP frontend of the ASR system. After 13 PLP feature were computed, the first, second, and third derivatives were appended, and the following normalizations and transformations were applied: vocal tract normalization; mean and variance normalization; LDA; MLLT (from 52 dimensions to 39); and a feature transform estimated by constrained MLLR, as used in speaker-adaptive training (SAT). These feature normalizations used gender-dependent reference models and transformations. The frontend was optimized for telephone ASR.

3.2. MLLR-SVM system: *mllr*

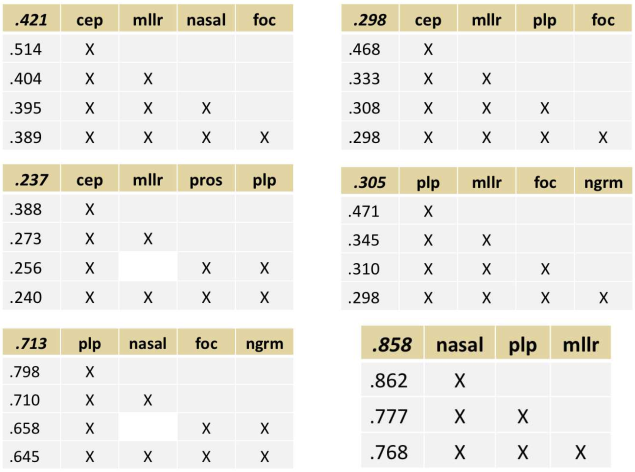
The MLLR-SVM system used the speaker adaptation transforms from the speech recognition system as features for speaker verification. A total of 16 affine 39x40 transforms were used to map the Gaussian mean vectors from speaker-independent to speaker-dependent speech models; eight transforms each were estimated relative to the male and female recognition models, respectively. The within-speaker variance was estimated on SRE04 telephone data, SRE05 microphone data, SRE08 and SRE10 sample data, and SRE08 speakers designated for training. The impostor (background) data for SVM training was from SRE06 telephone and microphone sessions, as well as from SRE08 data designated for training. For more details on MLLR-SVM modeling, see [5].

3.3. Word N-gram SVM system: *ngram*

This system used the relative frequencies of word N-grams as a sparse feature vector, and SVMs as speaker models. The impostor/background data was drawn from SRE04 and SRE05, plus SRE08 data reserved for training. The 9000 most frequent word bigrams and trigrams from the training data were included as features. No score normalization was applied.

3.4. Prosodic system: *pros*

The prosodic system was composed of a total of 10 individual systems combined at the score level with fixed weights. All individual systems used the same type of feature: the coefficients of the Legendre polynomial approximation of order 5 of the pitch and energy signals over a certain region, plus the duration of the region [6]. The region definition varied across systems. Additionally, some systems modeled sequences of two consecutive feature vectors [7]. Each individual system was modeled in a gender-dependent way using JFA, with 50 channel factors and 100 speaker factors. The three region definitions are: (1) Energy-valley regions: defined by the valley in the energy profile restricted to voiced regions; (2) Syllable regions: defined by automatic syllabification of the phone alignments produced by our speech recognizer; (3) Uniform regions: defined over speech regions to shift by a fixed amount of frames (15) and be of a fixed frame length (30). The Uniform regions definition was inspired by the work in [8]. For the first two regions, four systems were created: System (1) for the features over the nonpause regions (unigrams); System (2) for the concatenated features of two, consecutive, nonpause regions (ff bigrams); System (3) for the duration of a pause concatenated with the features of the following nonpause region (pf bigrams); and System (4) for the features of a nonpause region concatenated with the duration of the following pause (fp bigrams). For the Uniform regions definition, we found that Systems (3) and (4) added nothing to the overall combination; hence, only Systems (1) and (2) were used for these regions. The 10 systems were combined by giving a weight of 1.0 to the unigrams and the ff bigrams, and a weight of 0.5 to the pf and fp bigrams. The weights for the syllable region scores were set to half of those weights. These weights were obtained by first training a combiner using logistic regression and then performing a very rough exploration of manual weights that led to similar results. Pitch and energy features signals for each conversation side were obtained using the `get_f0` code from the freely available Snack toolkit [9]. The waveforms were preprocessed with a bandpass filter (250-3500 Hz) to make the spectral content of all channels similar to that of the telephone bandwidth.



.421	cep	mllr	nasal	foc
.514	X			
.404	X	X		
.395	X	X	X	
.389	X	X	X	X

.298	cep	mllr	plp	foc
.468	X			
.333	X	X		
.308	X	X	X	
.298	X	X	X	X

.237	cep	mllr	pros	plp
.388	X			
.273	X	X		
.256	X		X	X
.240	X	X	X	X

.305	plp	mllr	foc	ngrm
.471	X			
.345	X	X		
.310	X	X	X	
.298	X	X	X	X

.713	plp	nasal	foc	ngrm
.798	X			
.710	X	X		
.658	X		X	X
.645	X	X	X	X

.858	nasal	plp	mllr
.862	X		
.777	X	X	
.768	X	X	X

Fig. 1. N-best combination system results on Condition (from left to right, top to bottom, 2 3 4 5 6 7. The overall newDCF (with 7 subsystems) appears in the upper left corner of each table. While not appearing as the best subsystem individually, high-level systems like *mllr* and *nasal* are key to the overall performance.

4. SPEAKER DETECTION RESULTS

4.1. Standalone results

Table 2 shows the results of the SRI submission and subsystems on the core conditions in the SRE10 extended set. The reader should refer to the NIST evaluation plan [10] for a more detailed description of the conditions.

First, on the traditional tel-tel condition, both the SRI_1 and SRI_2 combinations outperformed the individual systems by approximately 35%. We also note that SRI_2 outperformed SRI_1 by approximately 5%, which reflects the addition of the SNR and the word count as side information, and the nasals system. The standard cepstral system *cep* performed the best for most conditions. However, the *plp* system has an advantage for the conditions involving telephone data, possibly because its ASR frontend based was optimized and trained for telephone speech. The combination of these two systems resulted in significant gains (not shown here for lack of space). The *nasals* system placed first for Condition 7—outperforming the standard JFA system—and, for Condition 6, gave better results than the MLLR system. While using 18% of the data, the nasal syllable regions seemed relatively robust to high vocal effort. A more detailed study of this phenomena seems in order.

4.2. N-best combination results

The previous section has highlighted the results of individual systems, but the key for good overall performance is identifying models that give complementary information. Figure 1 illustrates results from oracle-based N-best combination to get the best performance, with N from 1 to 4 on Conditions 2 to 7. Here, an interesting trend appears, in which the cepstral systems—while usually showing the best results—were not the second choice in combination. Indeed, the *mllr* system was usually the second best for most conditions. The prosodic system helped most for Condition 4, while the word N-gram system helped for Conditions 5 and 6.

5. CONCLUSIONS

The SRI submissions for NIST SRE 2010 were composed of several subsystems using both low- and high-level features. We showed that the GMM/JFA based systems tended to perform best for most test conditions. A PLP-based system, using a telephone speech ASR frontend, tended to outperform the classic MFCC system for telephone data. The MLLR-SVM system had very good performance in terms of newDCF (relative to oldDCF) and gave excellent gains in combination with the frame-based cepstral systems. The nasal-syllable constrained cepstral system was especially useful for the high-vocal effort conditions, and even outperformed the standard JFA system in one condition. We also found that using SNR and word count as side information (in addition to the evaluation condition) for the combiner yielded gains over a static combiner.

6. ACKNOWLEDGMENTS

This research was funded through a development contract with Sandia National Laboratories (#DE-AC04-94AL85000), as well as by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, the U. S. Government, or Sandia.

7. REFERENCES

[1] E. Dalmasso, F. Castaldo, P. Laface, D. Colibro, and C. Vair, "Loquendo-Politecnico di Torino's 2008 NIST speaker recognition evaluation system", *in Proc. ICASSP*, pp. 4213–4216, Taipei, Apr. 2009.

[2] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis", *in Proc. ICASSP*, pp. 4057–4060, Taipei, Apr. 2009.

[3] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System combination using auxiliary information for speaker verification", *in Proc. ICASSP*, pp. 4853–4857, Las Vegas, Apr. 2008.

[4] T. Bocklet and E. Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection", *in Proc. ICASSP*, pp. 4525–4528, Taipei, Apr. 2009.

[5] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-transform-based speaker recognition", *in Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, pp. 1–6, San Juan, Puerto Rico, June 2006.

[6] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 2095–2103, Sep. 2007.

[7] L. Ferrer, N. Scheffer, and E. Shriberg, "A comparison of approaches for modeling prosodic features in speaker recognition", *in Proc. ICASSP*, pp. 4414–4417, Dallas, Mar. 2010.

[8] M. Kockmann, L. Burget, and J. Cernocky, "Investigations into prosodic syllable contour features for speaker recognition", *in Proc. ICASSP*, pp. 4418–4421, Dallas, Mar. 2010.

[9] K. Sjolander, "The Snack sound toolkit", www.speech.kth.se/snack.

[10] National Institute of Standards and Technolgy, "The NIST year 2010 speaker recognition evaluation plan", http://www.itl.nist.gov/iad/mig//tests/sre/2010/index.html, 2010.