# 2010 NIST Speaker Recognition Evaluation: SRI System Description

**Luciana Ferrer, Martin Graciarena, Sachin Kajarekar, Nicolas Scheffer, Elizabeth Shriberg, Andreas Stolcke**
**(SRI International, CA, USA)**

# 1 Overview of the submission

SRI submitted two systems to SRE10:

**SRI_1** is a static score-level fusion of three cepstral GMM/JFA systems, an MLLR transform SVM system, and a word N-gram SVM system (the latter two being based on ASR). The originally submitted version of SRI_1 had a bug in the processing of development and training data that caused degradation of microphone and interview data and associated ASR. This has been corrected in the resubmission.

**SRI_2** is an enhanced system that adds a constrained cepstral GMM/JFA system, and uses signal-to-noise ratio and amount of speech detected as side information in the score combiner.

# 2 Commonalities

## 2.1 Development set based on 2008 NIST SRE

A set of 82 interview speakers (48 females and 34 males) was held out from the SRE08 trial definition to be used for training the systems. These speakers were all speakers from the interview conditions, some from the original set, others from the follow-up set.

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 against every available test sample (except when the model and the test sample used data from the same original recording session). No additional models were created and only samples originally used for testing were used for testing in the extended set. The follow-up test data was added to all the interview conditions.

In the rest of this document 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: alternate microphone (mic) or telephone channel (phn)

For consistency and completeness, we always use the full definition of the condition, even when some part is redundant (eg, in long-long.int-int.mic-mic, the mic-mic part is redundant in the current data).

In the case of the shrt-shrt.tel-tel.phn-phn condition, the target trials were restricted to be the target trials defined by NIST. This is because, as Niko Brummer observed, additional

target trials seemed to belong to same-telephone number trials. Since the telephone number is not available, we were not able to do any smart filtering.

Trials for the previously nonexistent short-long condition were created by using the long-long condition while replacing the training data by a long sample from the same speaker using the same microphone.

Table 1 shows a summary of the created trials by condition and the mapping used to SRE10 conditions for the cases in which an exact match was not possible. This mapping is used for training combination parameters.

**Table 1: Development conditions, the number of trials and the SRE10 conditions for which they are used as training data for the combiner.**

| Condition | Number target trials | Number impostor trials | Used for 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 |

Note: we define “long” as an utterance of 8 minutes of interview speech. In order to match this duration, we used the first 8 minutes of the long samples from SRE 2008.

## 2.2 Background and score normalization data

Background data for Gaussian mixture model (GMM) systems and impostor data for support vector machine (SVM) systems were obtained from 2004, 2005 and 2006 speaker recognition evaluation (SRE), NIST 2008 interview held-out development data, and NIST-provided 2010 development sample data. Different subsets of this data were used by different systems for Tnorm and Znorm.

## 2.3 Waveform preprocessing and segmentation

Prior to cepstral feature extraction and automatic speech recognition (ASR) processing, phonecalls (for both telephone and microphone channels) were segmented into short

segments containing mostly speech, using a speech/nonspeech 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 incorporates elements from LPT's 2008 processing and makes use of the NIST-provided ASR output for interviews:

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 the NIST RT-07 evaluation training data), 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, otherwise use output from step 1 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.

## 2.4 Speech recognition system

Several of the speaker verification models described below rely on word and sub-word recognition hypotheses obtained by ASR. We used a fast version of SRI's conversational telephone recognition system [2] with modifications for the SRE data. The first recognition pass generates lattices using a bigram LM and acoustic models based on MFCC features with fMPE transforms, augmented with MLP phone posterior features. The lattices are then rescored with a 4-gram LM. A second recognition pass uses speaker-adapted fMPE-PLP models, generating N-best lists that are 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 system runs in real-time on a 4-core 2.6GHz AMD Opteron machine. The word error rate on 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 altmic data (transcribed at ICSI) was 28.8%.

# 3 Individual System Description

Table 2 gives an overview of the individual systems listing the features and the modeling method used for each case.

**Table 2: Individual Systems**

| Feature | Statistical Modeling Framework |
|---|---|
| Cepstrum | GMM-JFA |
| Constrained Cepstrum | |
| Cepstrum | GMM-JFA, class-dependent |
| PLP-SAT Cepstrum | |
| MLLR Transform from English ASR | SVM |
| Word Ngrams | |
| Prosodic | GMM-JFA + Score-level combination |

## 3.1 Cepstral GMM-JFA systems

### 3.1.1 Standard cepstral system

The cepstral GMM system uses a 300-3300 Hz bandwidth front end consisting of 24 Mel filters to compute 20 cepstral coefficients with cepstral mean subtraction, and their delta, double delta coefficients, producing a 60 dimensional feature vector. The resulting features are mean and variance normalized over the utterance. The feature vectors are modeled by a 1024-component gender independent GMM. The background GMM is trained using data from the 2004 and 2005 SRE and 2008 interview development data. Joint factor analysis (JFA) is performed on mean supervector with speaker, channel and diagonal factors. Speaker factors are trained with 2004 and 2005 SRE data with Switchboard-II corpus. Channel factors are obtained separately for telephone (phonecall and mic) data and interview data. The two factors are combined to form single channel factor matrix. The diagonal term is trained with the same data as speaker factors. Scores are generated using asymmetrical scoring of subspace adapted mean supervectors. The resulting scores are normalized using gender dependent ZTnorm.

### 3.1.2 Constrained cepstral system

The SRI_2 (but not the SRI_1) system includes a single constrained cepstral system that uses features computed as in SRI's MFCC-based cepstral GMM-UBM system, but restricted to frames occurring in syllables that contain the recognized phone [n] or [ng]. Syllables are based on an automatic maximum-onset-based cross-word syllabification of ASR output. The resulting frames comprise about 18% of the total speech-aligned frames used in the standard system. UBM and JFA parameters are the same as for the standard cepstral GMM system, and the same dot-product scoring with ZT-normalization is employed.

### 3.1.3 Class-dependent cepstrum

This second MFCC-based system is similar to 3.1.1. It differs by using gender-dependent UBM models, eigenvoices and eigenchannels. The ZT-normalization process is also condition-dependent. The system also differs from the basic cepstral system in that it does not use the diagonal term in the JFA framework. When computing the speaker verification score, the statistics on the test utterance are normalized by a value derived from the speaker eigenvoice matrix. This normalization is giving an effect similar to Torm (veneer Tnorm).

Eigenvoices and eigenchannels are not condition-dependent since that did not prove to be useful. Eigenchannels are a concatenation of 4 subspaces: 32% telephone channel data, 32% telephone data over microphone, 32% interview data, 4% intrinsic variation data from the NIST SRE10 development set.

The ZTnorm process is condition-dependent, in the sense that normalization data is matched to the target testing condition. For example, for the `long-shrt.int-tel.mic-phn` condition, Tnorm is using short telephone data only, while Znorm is using long interview data only. We generalized this approach to all other conditions.

### 3.1.4. Class-dependent PLP-SAT cepstrum

This system uses the exact same setup as 3.1.3. However, the input features are generated by the PLP front end of the ASR system. After PLP-13 feature computation, first, second, and third differences are appended, and the following normalizations and transformations are applied: vocal tract normalization (VTLN), 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 use gender-dependent reference models and transformations.

## 3.2 MLLR SVM system

The MLLR-SVM system uses speaker adaptation transforms used in the speech recognition system as features for speaker verification. A total of 16 affine 39x40 transforms are used to map the Gaussian mean vectors from speaker-independent to speaker-dependent speech models; 8 transforms each are estimated relative to male and female recognition models, respectively. The transforms are estimated using maximum-likelihood linear regression (MLLR) [2], and can be viewed as a text-independent encapsulation of the speaker's acoustic properties. The transform coefficients form a 24,960-dimensional feature space. Each feature dimension is rank-normalized by replacing the value with its rank in the background data, and scaling ranks to lie in the interval [0, 1]. Nuisance attribute projection (NAP) is then applied. 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 resulting normalized feature vectors are then modeled by SVMs using a linear kernel, as described

in more detail in Section 3.4. The impostor (background) data for SVM training comes from SRE06 telephone and microphone sessions, as well as SRE08 data designated for training. For more details on MLLR SVM modeling see [3, 4].

## 3.3 Word N-gram SVM system

This system uses Word N-grams relative frequencies as a sparse feature vector, and SVMs as speaker models. The impostor/background data is drawn from SRE04 and SRE05, plus SRE08 data set aside for training. The 9000 most frequent word bigrams and trigrams from the training data are include as features. Relative N-gram frequencies are rank-normalized. Speaker models are obtained by training an SVM, using a linear kernel function. No score normalization is applied.

## 3.4 Prosodic system

The prosodic system is composed of a total of 10 individual systems combined at the score level with fixed weights. All individual systems use 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 [7]. The region definition varies across systems. Additionally, some systems model sequences of two consecutive feature vectors, with the features being either the actual features or the duration of the region if the region corresponds to a pause [14].

The three region definitions used are:

- **Energy-valley regions**: defined by the valley in the energy profile restricted to voiced region
- **Syllable regions**: defined by automatic syllabification of the phone alignments produced by our automatic speech recognizer
- **Uniform regions**: defined over speech regions to shift by a fixed amount of frames (15) and be of a fixed frame length (30).

The last region definition was inspired by the work in [13]. For the first two regions, 4 systems are created, one for the features over the nonpause regions (unigrams), one for the concatenated features of two consecutive nonpause regions (ff bigrams), one for the duration of a pause concatenated with the features of the following nonpause region (pf bigrams), and one for the features of nonpause region concatenated with the duration of the following pause (fp bigrams). For the third region definition we found that the last two cases did not add anything to the overall combination; hence, only two systems are used for that region definition.

Each individual system is modeled in a gender-dependent way using JFA, with 50 channel factors and 100 speaker factors. The data for JFA is taken from all the data described in Section 2.2 with the addition of Switchboard data.

The 10 systems are 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 are given by 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 are obtained using the get_f0 code from the freely available Snack [6] toolkit. The waveforms were preprocessed with a bandpass filter (250-3500) to make the spectral contents of all channels similar to that of the telephone bandwidth. These signals are used to extract prosodic features for the systems described in this section.

# 4 System Combination

The combination of systems is performed using linear logistic regression by condition as given in Table 1. The method proposed in [10] is used to compensate for biases produced by differences in word content and SNR of the signals. For this, SNR is computed on each session and thresholded at 15db to generate two categories: low and high SNR. Similarly, the number of words in the session is obtained and thresholded at 200 to generate two categories: short and long sessions. Finally, the category for each trial is created as a concatenation of the word-count and SNR categories for the train and test sessions creating a total of 16 possible categories. The method allows for a regularization parameter that encourages category-dependent weights to be small.

The side-information combiner was only used in the SRI_2 submission due to lack of time before the deadline. Furthermore, for the original SRI_1 submission only six systems were combined: all systems in Table 2 except the constraint cepstrum. For the SRE_2 submission, the constraint system was added to the pool.

After combination the scores are assumed to be calibrated. Therefore, a fixed threshold, given by the theoretically optimal value for the DCF of interest, is used to make the final decisions.

# References

[1] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," presented at Interspeech, Brighton, 2009.

[2] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of HMMs," *Computer Speech and Language*, vol. 9, pp. 171-186, 1995.

[3] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," presented at Eurospeech, Lisbon, Portugal, 2005.

[4] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-transform-based speaker recognition," presented at IEEE Odyssey 2006 Speaker and Language Recognition Workshop, San Juan, Puerto Rico, 2006.

[5] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," presented at European Conference on Machine Learning, 1998.

[6] K. Sjölander, "The Snack Sound Toolkit, www.speech.kth.se/snack."

[7] 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.

[8] 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.

[9] 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.

[10] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System combination using auxiliary information for speaker verification," presented at ICASSP, Las Vegas, 2008.

[11] 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.

[12] T. Bocklet and E. Shriberg,. "Speaker Recognition Using Syllable-Based Constraints for Cepstral Frame Selection" , *Proc. ICASSP*, Taipei, Taiwan, 2009.

[13] M. Kockmann, L. Burget and J. Cernocky , "Investigations into Prosodic Syllable Contour Features for Speaker Recognition", *Proc. ICASSP*, Dallas, 2010.

[14] L. Ferrer, N. Scheffer, E. Shriberg, "A Comparison of Approaches for Modeling Prosodic Features in Speaker Recognition" , *Proc. ICASSP*, Dallas, 2010.