MITLL 2012 Speaker Recognition Evaluation System Description

Technical Contributors in Alphabetical Order:
Jonas Borgstrom[†], William Campbell[†], Najim Dehak[‡], Reda Dehak[**], Daniel Garcia-Romero[*],
Kara Greenfield[†], Alan McCree[*], Doug Reynolds[†], Fred Richardson[†], Elliot Singer[†],
Douglas Sturim[†], Pedro Torres-Carrasquillo[†]

[†]MIT Lincoln Laboratory
[‡]MIT CSAIL
[*]JHU Human Language Technology Center of Excellence
[**]LRDE

1. Submission Descriptions

1.1. Core systems

Our submissions were built upon 5 core systems:

- **ivec-MITLL/JHU HLT COE** – two i-vector systems from MITLL and JHU using the same code base
- **ivec-MIT CSAIL/LRDE** – i-vector system from MIT and LRDE
- **SVM IPDF** – MITLL SVM system with IPDF KL kernel
- **Content Graph** – MITLL content graph system using IPDF kernel

Details of the core systems are provided in the sections that follow.

1.2. Submitted Systems

We examined various combinations of the core systems based on dev fusion results for the different conditions represented in the evaluation. Our final system was a fusion of calibrated individual systems.

The systems submitted were:

- **(Primary Submission) MITLL_01 Fuse** – Fusion of the 3 i-vector systems and the SVM IPDF system
- **MITLL_02** – SVM IPDF system
- **MITLL_03** – MITLL i-vector system
- **MITLL_04** – MIT CSAIL and LRDE i-vector system
- **MITLL_05** – JHU HLT COE i-vector system

2. Development Data

2.1. Development Trial Lists

Speaker training lists were obtained from the NIST distribution. Training lists were filtered to eliminate any nominally short durations (10s) and 2-wire data. Lists were augmented with some additional speaker data found in keys. Some redundant cuts (exact duplicates and same sessions) were included in the lists. Additionally, inconsistent key speaker pins for various key releases resulted in some changes in lists.

The resulting training lists were split into development lists:

- **dev-trn** A training set consisting of approximately 36k sides from all of the target speakers.
- **dev-tst** A test set consisting of approximately 2k telephone sides and 6k microphone sides from known target speakers.

By side, we mean one channel of a two-channel telephone or microphone file.

Additional data preparation was used to construct a full development set. For interview microphone data in the development set, a new set of files was constructed to better match the NIST SRE 2012 plan. The process used was:

- NIST SRE 08 interview data. All 13 microphone types were covered. When needed, new files were constructed with the interviewee on side "a" and the corresponding interviewer on a lapel microphone on side "b." Only nominal length 3 minute data was used. Data was in its original 8 kHz format.
- NIST SRE 10 interview data. Interview and interviewee data was paired as side a/b using 16 kHz data. The speech was then resampled using an FIR filter to 8 kHz. Both 3 minute and 8 minute nominal duration data was available.

Both phonecall microphone and 4-wire data were left in their original form. Note that mismatch occurs across the microphone data. NIST SRE 08 data was ulaw'd and resampled by LDC/NIST; the original 16 kHz data was not available. The NIST SRE 10 interview data was encoded with linear PCM (no ulaw) and resampled to 8 kHz by MIT LL. Additionally, interviewer data was not "clean", but included the original noise added during the evaluation in 2008 and 2010.

Several modifications to the files generated for **dev-tst** were constructed to address the conditions specified in the NIST evaluation plan:

- Utterances of varying duration. Using SAD, durations of 5, 10, 15, 20, 25, 30, 35, 40, 50, and 60 seconds were extracted from the full sides.
- Utterances with added noise. Noise was added at SNRs of 6dB and 15dB. Noise types used were babble and hvac.

For development trials, all models were scored against all **dev-tst** utterances including the modified test. Approximately 212 million trials were available for development.

Additional data was also available for hyperparameter construction. Data was aggregated from NIST SRE evaluation 2004,

2005, 2006, 2008, and 2010 as well as the Fisher data set to provide additional training. Different uses of this data are highlighted in the system descriptions.

3. Front-End Processing

3.1. Features

Two forms of preprocessing before feature extraction were performed on all the NIST data (including microphone and telephone): 1) steady tone removal and 2) wideband noise reduction. The steady tone suppression method used a very long analysis window, 8 seconds, to exploit the coherent integration of the Fourier transform. The wideband noise reduction algorithm used an adaptive Wiener-filter approach directed toward preserving the dynamic components of a speech signal while effectively reducing noise. Details can be found in [1, 2].

Speech activity detection (SAD) was performed with multiple methods based on the system. SAD was available to all systems in three forms: a SAD system based on GMMs trained on 4-wire data, a standard energy based detector based on one-channel data, and a two-channel multi-frequency band SAD designed by Borgstrom, see [2], was used on microphone data. Additional alternate SAD methods were used on a per-site basis—see the system descriptions.

MFCC features were extracted from the speech signal every 10ms using a 20ms window to produce a 20-dimensional mel-cepstral vector. The mel-cepstral vector is computed using a simulated triangular filterbank on the DFT spectrum. All frequency bands are kept from 0Hz-4kHz, and cepstral coefficients are computed via a DCT transform. The MFCC C0 coefficient was also included as an energy estimator. Delta cepstra are then computed over a +-2 frame span and appended to the cepstra vector. Double delta cepstral coefficients are formed on top of these, producing a 60 dimensional feature vector. Finally, the cep+dcep+ddcep features are normalized using mean/variance normalization and optionally RASTA on speech-only frames.

For LPCC features, pre-emphasis with a coefficient of 0.97 and a Hamming window are applied to a 30ms window every 10ms to obtain 18 LP coefficents. These LP coefficients are converted to 18 LPCCs and energy is appended to form a 19 dimensional vector. Both delta- and acceleration coefficients are found to form a 57 dimensional feature vector. Feature normalization and RASTA were applied to speech only frames.

Individual systems used different subsets of features based on tuning.

4. Detailed System Descriptions

4.1. MITLL and JHU HLT COE i-vector systems

Both the MIT LL and JHU HLTCOE i-vector systems use Bayesian model adaptation with an additive Gaussian noise model [3]. More specifically, we assume that speakers are normally distributed with mean $\boldsymbol{\theta}$ and across-class covariance $\boldsymbol{\Phi}_s$, and observed i-vectors are degraded by an additive channel component with within-class covariance $\boldsymbol{\Phi}_c$. This results in a scoring formula for test i-vector $\mathbf{w}_t$ given by

$$\mathcal{L}(\mathbf{w}_t|\mathcal{D}) = \log \frac{\mathcal{N}\left(\mathbf{w}_t; \boldsymbol{\mu}_{\mathcal{D}}, \boldsymbol{\Phi}_c + \boldsymbol{\Phi}_{\mathcal{D}}\right)}{\mathcal{N}\left(\mathbf{w}_t; \boldsymbol{\theta}, \boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right)} \tag{1}$$

where

$$\boldsymbol{\Phi}_{\mathcal{D}} = \frac{1}{N}\boldsymbol{\Phi}_s \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \boldsymbol{\Phi}_c, \tag{2}$$

and

$$\begin{aligned}\boldsymbol{\mu}_{\mathcal{D}} =& \frac{1}{N}\boldsymbol{\Phi}_s \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \sum_{i=1}^{N} \mathbf{w}_i \\ &+ \frac{1}{N}\boldsymbol{\Phi}_c \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \boldsymbol{\theta}.\end{aligned} \tag{3}$$

This is equivalent to full-rank Gaussian PLDA scoring, but this form is simpler to evaluate with multiple enrollment cuts. However, due to the high redundancy of training data in this evaluation, we found it more effective to assume only one enrollment cut, i.e. to represent each enrollment set by its mean i-vector.

The total variability space of dimension 600 was trained using principal component analysis (PCA). Linear discriminant analysis (LDA) was applied to further reduce dimension to 200. The within-class and across-class parameters were initially estimated by sample covariance matrices (as in LDA), and then refined using discriminative training. We used the two-class maximum mutual information (MMI) technique from [3]. To match the adverse acoustic conditions introduced in the SRE12, we used multicondition data, which included both additive noise and short duration speech.

The HLTCOE i-vector system also included MMI training of the mean vector for each speaker, $\boldsymbol{\mu}_{\mathcal{D}}$, while keeping the Bayesian covariance $\boldsymbol{\Phi}_{\mathcal{D}}$. This was done for all speakers simultaneously using multiclass MMI to optimize the closed-set identification performance. Again, multicondition training data was used including clean and noisy speech cuts.

To leverage information provided by the known speakers of the SRE12, we normalized the LLR scores using a back-end based on Bayes' rule. If $\mathcal{L}_{i,j}$ denotes the log-likelihood ratio (LLR) obtained when scoring test cut i against model j, the normalized LLR was determined as:

$$\hat{\mathcal{L}}_{i,j} = \log \frac{\frac{1}{M}\exp\left(\mathcal{L}_{i,j}\right)}{\frac{1}{M}\sum_{l\neq j}\exp\left(\mathcal{L}_{i,l}\right) + \frac{Poos}{1-Poos}} \tag{4}$$

Here, $Poos$ refers to the out-of-set probability, and M is the total number of known speakers.

Specific system parameters for both systems were:

- Identical processing for telephone and microphone speech signals
- Completely gender-independent parameters
- Tonal and wideband noise reduction
- MFCCs and deltas
- Short-term mean and variance normalization
- GMM and 2-channel frequency dependent energy SAD
- 2048 mixture GMM-UBM trained on clean target cuts
- 600-dimensional i-vector extractor trained using clean target cuts
- 200-dimensional LDA estimated from multicondition target cuts

4.2. MIT/LRDE i-vector system

This system is based on a gender independent i-vector representation [4] of dimension of 800 trained on all switchboard II and previous NIST SRE datasets as well as some noisy from NIST SRE data. The i-vector was trained using a gender independent universal background model comprised by 2048 Gaussians using the same data as the i-vector training. This system operates on cepstral features and log energy with delta and double-delta to produce a 60 dimensional feature vector (as described in the feature extraction section). Feature warping was performed using a 3s sliding window.

The silence was removed based on two different VADs. The first VAD, which is an MLP based system, was provided by Brno University of Technology (BUT), this VAD was applied only for telephone data. We would like to thanks BUT for agreeing to share their system with us. The second VAD was applied for both microphone and interview data. For microphone data, we used an iterative algorithm based on GMM models. At each iteration, we use the previous Viterbi segmentation to estimate two different GMMs comprised by 16 Gaussians each and which correspond to the noise and speech frames. These two GMMs were respectively initialized with frames that have the 10% highest and lowest energy. For the interview data, similar steps were applied as in microphone data for both channels, which corresponds to the interviewer and interviewee. After obtaining both segmentations for the interviewer and interviewee, we select the speech labeled frames from the interviewee side that have higher energy or have been labeled as noise in the corresponding frames in the interviewer.

Our channel compensation approach consists on a cascade of gender dependent Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) [5]. We estimated two different LDA matrices on target and no-target data. In the first LDA, the between speaker covariance was estimated on telephone data and within speaker covariance matrix was estimated on both telephone and microphone dataset. However for the second LDA, we estimated both between and within speaker covariance matrices on telephone and microphone data. The last within class covariance normalization matrix was trained on the target speaker telephone and microphone data.

The speaker verification decision was based on cosine scoring [4, 6] and the score were s-normalized directly in the i-vector space as described in [6].

4.3. SVM IPDF System

The SVM IPDF supervector system is based on a combination of SVMs using GMM supervectors [7] and an approximate KL kernel using utterance-dependent mixture weights and MAP mean adaptation [8]. Extensive experiments were conducted to determine the best combination of SVM training, features, channel compensation and score normalization needed to optimize the NIST scoring criterion.

GMM supervectors were derived using MAP adaptation of means with a relevance factor of 0.01 on a per utterance basis and ML estimation of mixture weights. The SVM inner product C_{GM} is given by

$$C_{GM}(\mathbf{a}_i, \mathbf{a}_j) = (\mathbf{m}_i - \mathbf{m})^t (\boldsymbol{\lambda}_i^{1/2} \otimes I_n) \Sigma^{-1} (\boldsymbol{\lambda}_j^{1/2} \otimes I_n)(\mathbf{m}_j - \mathbf{m}). \quad (5)$$

In equation (5), $\mathbf{m}$ is the vector of stacked UBM means, Σ is the block diagonal matrix of UBM covariances, $\otimes$ is the Kronecker product, I_n is the identity matrix of size n, and $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ are diagonal matrices of mixture weights.

For compensation, weighted NAP (WNAP) [9] was used. Weighted NAP optimizes the criterion,

$$\min_U \sum_j W_j \|Q_{U,D}\delta_j\|_D^2 \quad (6)$$

where U is the nuisance subspace, $Q_{U,D}$ is the WNAP projection, D is the metric induced by the UBM,

$$D = (\boldsymbol{\lambda}^{1/2} \otimes I_n)\Sigma^{-1/2}, \quad (7)$$

δ_j is the training set, and W_j is set to the number of frames of speech. WNAP used a fixed matrix multiply.

All target speaker GMM supervectors for training were pooled into one data set. A kernel matrix was computed in parallel. Then, SVM training was applied using the standard 1-vs-rest technique.

System specifics include:

- Gender independent GMM UBM with 512 mixture components
- Pooled WNAP model. We pooled all microphone and telephone data from the target speakers and constructed a NAP projection that was gender independent. A corank of 64 was used for the projection.
- Z-norm. After training speaker models, z-norm parameters were computed using approximately 1k utterances per gender from NIST Eval05 telephone data.
- T-norm. T-norm was applied gender dependent across all target models. Bayes rule was applied as a score normalizer using (4) after T-norm to the scores.
- SAD. A cascade of GMMSAD and energy-based 1-channel SAD was used for 4w data. For microphone data, 2-channel energy based SAD followed by 1-channel energy-based SAD was used.
- Multiple features. The same SVM IPDF system (no configuration changes) was used with MFCCs and LPCCs. The output scores were averaged.

4.4. Content Graph System

For this evaluation, we also implemented a content graph based system. A content graph is a sparse representation of the speaker space in a graphical model. Each vertex corresponds to an utterance and vertices are connected via weighted edges if they are likely to be from the same speaker with the weight corresponding to the strength of this likelihood. For the evaluation, we created a content graph from the development data using the ForceClique link prediction algorithm developed in [10]. Test utterances were added to the graph using incremental k nearest neighbors instead of incremental force-clique link prediction in order to reduce $P(miss)$.

We explored two methods of performing speaker verification on content graphs. The first exploited local features of the test utterance node and and the set of nodes comprising the target speaker. We computed scores by comparing the number of connections between the test node and the target nodes, the number of connections from one target node to another, and the number of connections form either the test node or a target node to some node not in either of those sets. This method performed well in areas

of the graph with high edge prediction recall, but suffered a higher miss rate when edge prediction recall was low. Our second method consisted of perfoming a random walk originating from the test node and computing a score based on the likellihood that the random walk ended on a node from the target speaker. This method had improved performance in areas of the graph with lower edge prediction recall. The best performance was achieved by combining the two methods, using the local features method when there was at least one connection between the test node and the set of target nodes and using the random walk method otherwise.

The performance of the experimental content graph system was quite good, but not as good as our best systems. In the future, we will be working to improve it by applying techniques such as combining results across content graphs generated with different parameters and better fusion of the local features and random walk methods.

4.5. Fusion

Fusion was accomplished using a two-stage method. In the first stage, individual systems were calibrated using scores and meta-data. In the second stage, individual systems were fused and a calibrated output was produced using a logistic regression method.

First stage. Calibration of individual systems was performed using system scores and a variety of meta-data similar to [11]. A main feature of our methods was adaptation of calibration to the varying means and variances of the target and non-target distributions across different operating conditions (duration, SNR, channel).

For the SVM IPDF system, an MLP with 3 hidden nodes and inputs of score, duration, channel (mic/tel–from NIST), and gender (male/female–from NIST) were used. Full multilayer MLP training was done using scaled conjugate gradients and the Netlab MLP tool using a cross-entropy optimization criterion. For the i-vector systems, a fixed input layer was used which was hand-tuned and had inputs of SNR, duration, estimated gender, estimated language, score, and estimated reverberation. The output layer was trained using logistic regression.

Second stage. Fusion was accomplished by applying logistic regression to the individual calibrated log likelihood ratio scores produced by the four classifiers and was implemented using a single-layer perceptron available in the LNKnet pattern classification software package [12]. The perceptron configuration had four input nodes, two output nodes (target and nontarget), and no hidden layers. Input scores were normalized to zero mean and unit standard deviation using parameters derived from the training data. The perceptron weights were trained from the development data with a mean squared error criterion. Development testing indicated no difference in performance using either the squared error or cross entropy criterion. Both training and testing used uniform priors. The posterior probability was computed from the average of the target and (1-nontarget) scores and was converted to the final log-likelihood ratio using the logit function.

5. References

[1] D. E. Sturim, W. M. Campbell, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Robust speaker recognition with cross-channel data: MIT-LL results on the 2006 NIST SRE auxiliary microphone task," in *Proceedings of ICASSP*, 2007, pp. IV–49–IV–52.

[2] W. M. Campbell, D. Sturim, B. J. Borgstrom, R. Dunn, A. McCree, T. F. Quatieri, and D. A. Reynolds, "Exploring the impact of advanced front-end processing on nist speaker recognition microphone tasks," in *Proc. Speaker Odyssey Workshop*, 2012.

[3] Bengt J. Borgstrom and Alan McCree, "Discriminatively trained bayesian speaker comparison of i-vectors," in *Proc. ICASSP, submitted*, 2013.

[4] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] N. Dehak, Z. Karam, D. Reynolds, R. Dehak, W. Campbell, and J. Glass, "A channel-blind system for speaker verification," in *Proceedings of ICASSP*, 2011, pp. 4536–4539.

[6] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. IEEE Odyssey Workshop*, 2010.

[7] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proceedings of ICASSP*, 2006, pp. I–97–I–100.

[8] W. Campbell and Z. Karam, "Simple and efficient speaker comparison using approximate KL divergence," in *Proceedings of Interspeech*, 2010.

[9] W. M. Campbell, "Weighted nuisance attribute projection," in *IEEE Odyssey*, 2010.

[10] K. Greenfield and W. Campbell, "Link prediction methods for generating speaker content graphs," in *Submitted to Proceedings of ICASSP*, 2013.

[11] WM Campbell, DA Reynolds, JP Campbell, and KJ Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *Proc. ICASSP*, 2005, vol. 2005, pp. 717–720.

[12] R. P. Lippmann, L. C. Kukolich, and E. Singer, "LNKnet: Neural network, machine-learning, and statistical software for pattern classification," *Lincoln Laboratory Journal*, pp. 249–268, 1993.