MITLL 2010 Speaker Recognition Evaluation System Description

Technical Contributors in Alphabetical Order:
William Campbell[†], Najim Dehak[‡], Zahi Karam[†*], Alan McCree[†],
Doug Reynolds[†], Fred Richardson[†], Douglas Sturim[†], Pedro Torres-Carrasquillo[†]

[†]MIT Lincoln Laboratory
[‡]MIT CSAIL
[*]DSPG, Research Laboratory of Electronics at MIT

1. Submission Descriptions

1.1. Core systems

Our submissions were built upon 5 core systems and novel norming methods:

- ***IPDF*** – Approximate KL divergence scoring with WNAP compensation
- ***JFA*** – Joint factor analysis with GMM linear scoring
- ***Prosodic*** – Prosodic modeling system
- ***ECS*** – Eigenvoice comparison system
- ***GSV*** – SVM trained using GMM-UBM MAP adapted parameters and NAP compensation
- ***TV*** – Total variability space modeling with multiple compensation methods
- ***ZAT-NORM/SAS-NORM*** – Adaptive norming to optimize the minDCF

Details of the core systems are provided in the section below.

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. We fused at the level of subconditions of the various tasks, since the top were considered too broad. Subconditions were represented using a quadruple hyphenated string, *(num conversations)-(duration)-(style)-(channel)*.

Before fusing, we applied various combinations of adaptive norming with the core systems. In addition to the core systems which used either ZT-norm or S-NORM, we produced the following systems:

- ***SAS-TV*** – TV system with S-norm followed by adaptive S-norm
- ***IZAT*** – IPDF with Z-norm followed by adaptive T-norm
- ***ZAT3*** – JFA with Z-norm followed by adaptive T-norm

Our secondary submission fused only the core systems for each condition.
The final fusions were:
Primary Submissions:

- ***1-long-int-mic/1-long-int-mic*** – GSV + TV + ZAT3
- ***1-long-int-mic/1-short-cnv-4w*** – IPDF + TV + ZAT3
- ***1-long-int-mic/1-short-cnv-mic*** – GSV + TV + ZAT3
- ***1-long-int-mic/1-short-int-mic*** – GSV + TV + ZAT3
- ***1-short-cnv-4w/1-short-cnv-4w*** – SASTV + IPDF + IZAT + PROS + ECS + TV + JFA + ZAT3
- ***1-short-cnv-mic/1-short-cnv-mic*** – GSV + TV + ZAT3
- ***1-short-int-mic/1-long-int-mic*** – GSV + TV + ZAT3
- ***1-short-int-mic/1-short-cnv-4w*** – IPDF + TV + ZAT3
- ***1-short-int-mic/1-short-cnv-mic*** – GSV + TV + ZAT3
- ***1-short-int-mic/1-short-int-mic*** – GSV + TV + ZAT3
- ***8-short-cnv-4w/1-short-cnv-4w*** – GSV + JFA

Secondary Submission:

- ***1-long-int-mic/1-long-int-mic*** – GSV + TV + JFA
- ***1-long-int-mic/1-short-cnv-4w*** – IPDF + TV + JFA
- ***1-long-int-mic/1-short-cnv-mic*** – GSV + TV + JFA
- ***1-long-int-mic/1-short-int-mic*** – GSV + TV + JFA
- ***1-short-cnv-4w/1-short-cnv-4w*** – IPDF + PROS + ECS + TV + JFA
- ***1-short-cnv-mic/1-short-cnv-mic*** – GSV + TV + JFA
- ***1-short-int-mic/1-long-int-mic*** – GSV + TV + JFA
- ***1-short-int-mic/1-short-cnv-4w*** – IPDF + TV + JFA
- ***1-short-int-mic/1-short-cnv-mic*** – GSV + TV + JFA
- ***1-short-int-mic/1-short-int-mic*** – GSV + TV + JFA
- ***8-short-cnv-4w/1-short-cnv-4w*** – GSV + JFA

2. Development Data and Front-End Processing

2.1. Development Trial Lists

The corpora used for NIST SRE 2010 development lists consisted entirely of data from the 2008 NIST speaker evaluation. These lists were used for system tuning, system selection, and backend fusion/calibtration. Several adjustments were made so that the data would be suitable for developing a system designed to perform well in the 2010 NIST SRE:

- **additional background training data**: the NIST SRE 2008 interview microphone data was partitioned into two approximately equal sets one of which was used for subspace, ZT-norm or background model data and the other was used for development train and test data.

Train				Test		
#Enrl	Dur	Style	Chan	Dur	Style	Chan
1	long	int	mic	long	int	mic
1	long	int	mic	short	conv	tel
1	**long**	**int**	**mic**	**short**	**conv**	**mic**
1	long	int	mic	short	int	mic
1	short	conv	tel	short	conv	tel
1	short	conv	mic	short	conv	mic
1	short	int	mic	short	conv	tel
1	short	int	mic	long	int	mic
1	**short**	**int**	**mic**	**short**	**conv**	**mic**
1	short	int	mic	short	int	mic
8	short	conv	tel	short	conv	tel

Table 1: List of train and test conditions for SRE 2010.

- **increased non-targets**: an exhaustive set of non-target trials was created for each development test data set in order to match the much lower target prior in the 2010 NIST SRE.

The complete set of train and test conditions used for the LL/MIT SRE 2010 submission are given in Table 1. No development data was available for the two conditions that use interview data for training and conversational microphone data for testing. These two conditions will be discussed further in Section 3.8.

2.2. Features

Two forms of preprocessing before feature extraction were performed on the NIST data. For telephone speech, standard echo cancellation (EC) was applied using the ISIP tools. For microphone recorded data, noise reduction techniques were applied—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. Greater detail can be found in [1].

Speech activity detection (SAD) was performed with two distinct methods based on the channel type. For 4-wire and conversational microphone data, non-speech was eliminated using a feature-based GMM SAD detector. These speech/non-speech marks were further refined with an energy-only based detector. For interview microphone data, SAD was based on the NIST-supplied ASR transcripts.

MFCC features are 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. Delta cepstral 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 feature warping 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 warping is applied to speech only frames.

All systems used the common features excepted as noted in the description.

3. Detailed System Descriptions

3.1. Inner Product Discriminant Function System (IPDF)

Inner product discriminant functions (IPDFs) are described in [2]. We use a comparison function from the IPDF framework based on approximations to the KL divergence between two GMMs [3, 2], C_{GM}.

For a sequence of feature vectors from a speaker i, we adapt a GMM UBM by using standard relevance MAP [4] on the means and an ML estimate of the mixture weights. The adaptation yields new parameters which we stack into a parameter vector, $\mathbf{a}_i$, where

$$\mathbf{a}_i = \begin{bmatrix} \lambda_{i,1} & \cdots & \lambda_{i,N_m} & \mathbf{m}_{i,1}^t & \cdots & \mathbf{m}_{1,N_m}^t \end{bmatrix}^t \quad (1)$$

where $\lambda_{i,j}$ are the mixture weights, $\mathbf{m}_{i,j}$ are the means, and N_m is the number of mixtures. A relevance factor of 0.01 was used for MAP adaptation for the IPDF. The UBM used was gender independent and had 512 mixtures.

The inner product C_{GM} is given by

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

In equation (2), $\mathbf{m}$ 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 from (1).

For compensation, weighted NAP (WNAP) [5] was used. Weighted NAP optimizes the criterion,

$$\min_U \sum_j W_j \| Q_{U,D} \delta_j \|_D^2 \quad (3)$$

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

$$D = (\boldsymbol{\lambda}^{1/2} \otimes I_n) \Sigma^{-1/2}, \quad (4)$$

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

To obtain scores, we applied gender independent WNAP to both enroll and verification mean parameter vectors. The WNAP corank was fixed at 128. We then scored using the C_{GM} kernel. Both Z- and T-Norm were applied.

Depending on the task, different Z- and T-norm sets were used as well as different training sets for the WNAP subspaces. IPDF was run on 3 subconditions of the core task:

- 4w/4w: The WNAP training set was speakers from NIST SRE Eval04, SWB2 p1, p4, and p5. Z- and T-norm speakers were taken from Eval04, Eval05, and Eval06 4w conditions.
- short-interview-mic/4w: This condition was run in swapped mode where the role of enroll/verify was swapped. The WNAP training set was speakers from Eval05/Eval06 cross 4w/microphone conditions and a background development subset of cross 4w/interview microphone data from Eval08. T-Norm speaker were the Eval04/05/06 4w telephone set mentioned above. Z-norm utterances were from the Eval05, Eval06 and Eval08 microphone conditions.

- long interview-mic/4w: This condition was also swapped. Lists for the subspace, Z- and T-norm were the same as the short interview mic/4w condition.

3.2. JFA System

The base system for our Joint Factor Analysis (JFA) work was the MITLL GMM-UBM speaker detection system, fully described in [4]. Our JFA setup is based on the work of [6], where the mean supervector is decomposed as:

$$M = m + Vy + Dz + Ux, \tag{5}$$

where m is the speaker-independent mean supervector of GMM means, U defines the within-class (session/channel) variability subspace, V defines the across-class (speaker) variability subspace, and D is a diagonal matrix describing the remaining speaker variability.

We used gender-dependent UBMs with 1024 mixtures. 300 eigenvoices were trained using a variation of PCA of the across-class variability covariance matrix. To reduce over-estimation bias of the eigenvalues, a cross-validation approach was used where the eigenvectors were estimated from one partition of the training data and the eigenvalues were estimated as the energy in these directions over the other partition. We found that using this approach, the diagonal matrix could be estimated from the same data. In a similar way, 100 eigenchannels were estimated from the within-class covariance matrix. Two of these estimates were generated, one for telephone channels and the other for microphone conditions, and stacked together into a combined 200-dimensional matrix.

Enrollment of speakers in this system consists of estimating $Vy + Dz$ in the presence of Ux, and is done by stacking all the parameters together and extracting the speaker model. Testing is done by removing Ux from the test utterance. To speed up the Gaussian scoring, only the linear (inner product) term is calculated as in [7]. ZT-norm was applied to these output scores.

The following data was used to train this system:

- GMM background model – Trained from Switchboard II and SRE 2004,5,6 corpora.
- Across-class (speaker) matrices – Trained from NIST SRE SRE 2004,5,6 and switchboard II using data from speakers with 8 or more enrollment sessions.
- Within-class (session) matrix, telephone – Trained from NIST SRE 2004,5,6 using data from speakers with 8 or more enrollment sessions.
- Within-class (session) matrix, microphone – Trained from NIST SRE07 interview microphone data.
- Z-norm test utterances – NIST SRE SRE 2004,5.
- T-norm speakers – NIST SRE SRE 2004,5,6.

Two JFA systems were used, one based on MFCC features and the other using LPCC. The scores from these systems were fused to produce a single score as input for the final system fusion.

3.3. Prosodic System

A more extensive overview of the prosodic system and its features is detailed in [8]. These features are extracted at the pseudo-syllabic level and correspond to a Legendre polynomial approximation of the pitch and energy contours. We used six Legendre polynomial coefficients each for pitch and energy, as well as the duration of the pseudo-syllable to obtain a feature vector of 13-dimensions. We used a gender-dependent Universal Background Model (UBM) composed of 512 Gaussians per gender and gender-dependent total variability matrices of 200 eigenvectors trained only on telephone speech [9]. LDA was used to reduce the dimension to 75, while WCCN normalized the cosine scoring. We used cosine scoring and applied zt-norm to normalize the final decision scores.

3.4. Eigenvoice Comparison System (ECS)

For this evaluation we also implemented an eigenvoice comparison system (ECS). This is a speaker comparison system based on the following two key ideas:

- a speaker model lies entirely in the eigenvoice (speaker factor) subspace
- the within-class variability in this subspace is Gaussian.

In this system, speaker model enrollment consists simply of generating eigenvoice coeffients (speaker factors) for the enrollment utterance $\mathbf{s}_{train}$, without any session variability modeling or compensation.

For the test utterance, a test speaker model $\mathbf{s}_{test}$ is generated in the same way. Under the assumptions that the enrollment model is correct and the within-class variability is Gaussian, the likelihood that the test speaker model comes from the training speaker is found by evaluating $\mathbf{s}_{test}$ with mean $\mathbf{s}_{train}$ and covariance $\mathbf{\Sigma}_{wc}$. We have found good results using the non-target hypothesis of a random speaker, with zero mean and covariance $\mathbf{I} + \mathbf{\Sigma}_{wc}$. This covariance represents the sum of the speaker and session variation, where speaker variation in the speaker space is identity. One nice aspect of this system is that it does not require any form of score normalization.

However, we found the best fusion with the other systems with a simplified version of ECS. First, to simplify computation only the inner product term of the Gaussian likelihood is preserved as in JFA. Second, a small further improvement results from normalizing this inner product to both model magnitudes as in the cosine similarity measure used in total variability space, resulting in the score formula:

$$LL(\mathbf{s}_{test}|\mathbf{s}_{train}) = \frac{\mathbf{s}_{test}^T \mathbf{\Sigma}_{wc}^{-1} \mathbf{s}_{train}}{\sqrt{(\mathbf{s}_{test}^T \mathbf{\Sigma}_{wc}^{-1} \mathbf{s}_{test})(\mathbf{s}_{train}^T \mathbf{\Sigma}_{wc}^{-1} \mathbf{s}_{train})}} \tag{6}$$

As in the JFA system, ZT-norm is applied to these log likelihood scores. Note that this simplified ECS systems is essentially the same as the speaker factor cosine distance with WCCN approach proposed by Dehak et al at the JHU 08 Workshop.

The parameters needed for this modeling are the speaker subspace matrix in the full supervector space and the within-class covariance in the speaker subspace. For the first of these, we use the same matrix as in the JFA system. For the second, we compute a full covariance matrix in this 300-dimensional space using the same training list as the session variability training for JFA. Note that this system only used the MFCC features.

The performance of this new experimental system was quite good, but not as good as our best acoustic systems. However, it did provide a fusion gain for the telephone train and test condition, so was used there.

3.5. SVM GMM Supervector System (SVM GSV)

The SVM GMM supervector system is based on [3] GMM supervectors were derived using MAP adaptation of means only with a relevance factor of 4 on a per utterance basis. The kernel inner product used was

$$\begin{aligned} K(g_a, g_b) &= \sum_{i=1}^{N} \lambda_i \mathbf{m}_{a,i}^t \mathbf{\Sigma}_i^{-1} \mathbf{m}_{b,i} \\ &= \sum_{i=1}^{N} \left(\sqrt{\lambda_i} \mathbf{\Sigma}_i^{-\frac{1}{2}} \mathbf{m}_{a,i} \right)^t \left(\sqrt{\lambda_i} \mathbf{\Sigma}_i^{-\frac{1}{2}} \mathbf{m}_{b,i} \right) \end{aligned} \quad (7)$$

as in prior work. In equation (7), $\mathbf{m}_{*,i}$ are the adapted means, λ_i are the mixture weight of the UBM, and Σ_i are the UBM covariances. SVMs were trained using SVMTorch.

Data and strategy used for the SVM background, NAP and ZT-norm varied depending on the task.

For trials involving microphone training and testing, the following configuration was used:

- GMM UBM with 2048 mixture components.
- SVM background – 4000 Fisher corpus utterances (including non-English)
- NAP projection and T-Norm set–NIST SRE Eval05/06 speakers from the auxiliary microphone task and the NIST SRE Eval08 interview microphone background set. The NAP corank was 64.

Also, for microphone conditions, better performance was achieved by eliminating the acceleration coefficients.

The SVM GSV was also used on the 8 conversation 4w train, 1 conversation 4w test task. The following configuration was used:

- GMM UBM with 512 mixture components.
- SVM background – 4000 Fisher corpus utterances (including non-English)
- NAP projection–Eval04 speakers plus Switchboard 2 part 1 speakers, approximately 8000 utterances. The projection was trained using WNAP [5]. The WNAP corank was 64.
- ZT-Norm–Eval05/06 speakers with 8 conversation training.

3.6. Total Variability System

The total variability system is composed of two subsystems, one exclusively for telephone speech and another for microphone or interview data. The parameters for the first subsystem were trained on telephone data. We used a gender-dependent Universal Background Model (UBM) of 2048 Gaussians and gender-dependent total variability matrices consisting of 600 eigenvectors trained on telephone speech [9]. The use of Linear Discriminant Analysis (LDA) reduces our dimensionality to 250, and Within Class Covariance Normalization (WCCN) carries out the channel compensation in the total variability space [10]. Table 2 shows the list of corpora and their respective roles in the creation of our system. Similar to [9], we use cosine scoring and zt-normalization to make the final decision. As with everything else so far, the impostors for zt-norm were entirely selected from telephone speech data.

The second subsystem is used when we have microphone and interview data in training or in testing. This system is based on the total variability space and its 600 total factors estimated on telephone speech and an additional 200 total factors trained in microphone and interview data. We then use Probabilistic LDA [11] to project all microphone and telephone total factors of dimension 800 into speaker space of dimension 600. The PLDA consists of a mean vector of dimension 800 estimated from telephone data, an eigenvoice matrix of dimension 800x600 trained on telephone speech, an eigenchannels matrix of dimension 800x200 trained exclusively on microphone and interview speech, and a full covariance matrix trained from telephone speech. After the projection with PLDA, we used LDA to reduce the 600 dimensions to 250 and WCCN to normalize the cosine kernel. These channel compensation matrices are estimated using telephone, microphone and interview data all together. And as before, the decision score is computed using cosine scoring, but the final scores are normalized using s-norm [11]. The impostors used for s-norm are taken from NIST 2005, 2006 SRE telephone and microphone data, as well as some interview data from NIST 2008 SRE.

Note that for the telephone data, we used a silence detector provided by Brno University, which corresponds to the Hungarian speech recognizer labels (for more details, please see Brno's 2010 Submission). Additionally, the speech activity detection for microphone data was obtained from CRIM (for more details, please see CRIM 2010 submission).

3.7. Adaptive Norming

Adaptive norming of scores showed promise in our development set for minimizing the new minDCF criterion and was applied to three systems—IPDF, JFA, and TV. Adaptive normalization techniques were applied with inspiration from several sources including cohort normalization [12, 13], T-Norm [14], Z-Norm [4], and adaptive variants [15, 16].

As with classic cohort selection [13] and Z- and T-norm, there are several issues in adaptive methods—cohort selection, cohort normalization function, and whether the model or test score is normalized.

The basic cohort normalization functions were:

- Z-Norm Adaptive T-Norm (ZATnorm): for each trial score, the Z-Norm score was first computed:

$$s_z(x_{mod}, x_{msg}) = \frac{s(x_{mod}, x_{msg}) - \mu_{mod}}{\sigma_{mod}} \quad (8)$$

  where μ_{mod} and σ_{mod} are computed across a large set of Z-norm utterances (gender dependent). The ZAT-normed score is

$$s_{zat}(x_{mod}, x_{msg}) = \frac{s_z(x_{mod}, x_{msg}) - \mu_{msg,coh}}{\sigma_{msg,coh}}, \quad (9)$$

  where μ_{msg} and σ_{msg} are the speaker-dependent mean and standard deviation of the K cohorts chosen from a large T-Norm set applied to the message of interest.

- S-Norm followed by Adaptive Snorm (SASnorm): for each evaluation message/model pair Z-norm and Z-norm on the swapped evaluation message/model pair is performed (known as symmetric or S-Norm [11]),

$$\begin{aligned} s_s(x_1, x_2) &= \frac{1}{2} \frac{s(x_{mod}, x_{msg}) - \mu_{mod}}{\sigma_{mod}} \\ &+ \frac{1}{2} \frac{s(x_{mod}, x_{msg}) - \mu_{msg}}{\sigma_{msg}}, \end{aligned}$$

  where μ_{mod}, σ_{mod}, μ_{msg} and σ_{msg} are the mean and standard deviations of the utterances x_{mod} and x_{msg} scores against all of the Z-norm set.

Table 2: Corpora used for the TV system to estimate the UBM, total variability matrix (T), LDA and WCCN

	UBM	T	LDA	WCCN
Switchboard II, Phases 1, 2 and 3	X	X	X	
switchboard Cellular, Parts 1 and 2	X	X	X	
Fisher English database Part 1 and 2		X		
NIST 2004 SRE	X	X	X	X
NIST 2005 SRE	X	X	X	X
NIST 2006 SRE	X	X	X	X

ASnorm is then applied:

$$s_{sas}(x_{mod}, x_{msg}) = \frac{1}{2}\frac{s_s(x_{mod}, x_{msg}) - \mu_{mod,coh}}{\sigma_{mod,coh}} + \frac{1}{2}\frac{s_s(x_{mod}, x_{msg}) - \mu_{msg,coh}}{\sigma_{msg,coh}},$$

where $\mu_{mod,coh}$, $\sigma_{mod,coh}$, $\mu_{msg,coh}$ and $\sigma_{msg,coh}$ are the z-norm stats for the cohorts for the model and message.

Cohorts selection was accomplished by a simple method. For a given message or model, cohorts were selected as the highest scoring models or messages (respectively) from a large dataset. For telephone only tasks, the cohorts were all 4-wire utterances from Eval04, Eval05 and Eval06. For microphone data, the cohorts were chosen from Eval05 microphone, Eval06 microphone, and the heldout background Eval08 interview microphone data set. For IZAT (IPDF with ZAT-norm), 500 cohorts were chosen. For ZAT3 (JFA with ZAT-norm), 300 cohorts were chosen. SASnorm was applied to the total variability system using 694 top scores for males and 929 top scores for females (about 10%) of the data.

3.8. Fusion

Fusion was performed using a logistic regression. The criteria function of the logistic regression, normalized conditional cross-entropy or Cllr, was adjusted to use the new target prior for the 2010 NIST SRE: $P(tar) = 0.001$. Although the criteria function is not the same as NIST performance metric C_{Det}, we found that C_{Det} was generally improved when we optimized using the same effective target prior that matches Baye's optimal decision rule.

A separate logistic regression was trained for nine of the conditions listed in Table 1. The interview microphone logistic regressions with corresponding train and test durations were used for the two unseen conditions mentioned in Section 2 that use interview data for training and conversational microphone data for testing. For each of the IPDF, IZAT, JFA, GSV and ZAT3 systems, the MFCC and LPCC versions of these systems were fused together before the final fusion.

For our secondary submission we attempted to address the "same microphone" and "different microphone" sub-conditions. Our approach was to scale the number of "same microphone" samples to equal the number of "different microphone" samples in each condition where the same style microphone data was used in train and test. On our dev data we found that this approach greatly improved the "same microphone" calibration results with a relative small degradation to the "different microphone" performances.

3.9. Processing Times

Processing times were estimated for single trials of the core task on an Intel Xeon CPU running at 2.00GHz with 4 megabytes of cache and 8 gigabytes of memory. Results are shown in Table 3.9. The overall fused primary system runs about 0.6 times real time.

Table 3: Per System Processing Speed in Real Time Factors for core task estimated from single trials. The last column is the fraction of real time. For systems that fused LPCC and MFCC features (IPDF, JFA, SVM GSV), entries represent sum of both processing times.

Frontend	LPCC Features	0.07
	MFCC Features	0.07
IPDF	Total Enroll	0.026
	Total Verify	0.024
JFA	Total Enroll	0.15
	Total Verify	0.08
Prosodic	Total Enroll	0.035
	Total Verify	0.035
ECS	Total Enroll	0.05
	Total Verify	0.05
SVM GSV	Total Enroll	0.104
	Total Verify	0.096
TV	Total Enroll	0.16
	Total Verify	0.16
Fusion	Total Enroll	0.66
	Total Verify	0.58

Note that the adaptive norming versions (SAS-TV, IZAT, and ZAT3) use the same system outputs as the traditional norm results, so no significant extra system processing is required for these.

Also, these single-trial verification numbers are worst case for many applications. Verification processing of multiple targets on the same utterance requires a negligible increase in processing for most systems, so the per trial complexities would be much less. This savings is approximately linear with the number of targets, so for example verification times should be multiplied by 0.1 for processing 10 targets at a time. This rule does not apply to situations where enroll and verify lists were swapped; in this case the parallel processing savings will occur for enrollment.

4. Acknowledgments

Thanks to Tom Quatieri and Bob Dunn for providing C implementations of their wideband noise reduction and tone removal software. We also thank the SRE community for the interesting and helpful discussions on the SRE google group.

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, Z. N. Karam, and D. E. Sturim, "Inner product discriminant functions," in *Advances in Neural Information Processing Systems 22*, Cambridge, MA, 2009, MIT Press.

[3] 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.

[4] Douglas A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

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

[6] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions On Speech And Audio Processing*, vol. 13, no. 3, pp. 345, May 2005.

[7] Ondrej Glembek, Lukas Burget, Najim Dehak, Niko Brummer, and Patrick Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Proceedings of ICASSP*, 2009.

[8] N. Dehak, P. Kenny, and P. Dumouchel, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2095–2103, Sept. 2007.

[9] 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," in *Proceedings of Interspeech*, 2009.

[10] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front end factor analysis for speaker verification," *submitted to IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[11] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey 2010: The speaker and Language Recognition Workshop*, 2010.

[12] Alan Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, pp. 89–106, 1991.

[13] Aaron E. Rosenberg, Joel DeLong, Chin-Hui Lee, Biing-Hwang Juang, and Frank K. Soong, "The use of cohort normalized scores for speaker verification," in *Proceedings of the International Conference on Spoken Language Processing*, 1992, pp. 599–602.

[14] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[15] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for Tnorm in text-independent speaker verification," in *Proceedings of ICASSP*, 2005.

[16] Yaniv Zigel and Moshe Wasserblat, "How to deal with multiple-targets in speaker identification systems?," in *Proc. Odyssey*, 2006.