THE UMD-JHU 2011 SPEAKER RECOGNITION SYSTEM

D Garcia-Romero[1], *X Zhou*[1], *D Zotkin*[1], *B Srinivasan*[1], *Y Luo*[1], *S Ganapathy*[2], *S Thomas*[2], *S Nemala*[2], *GSVS Sivaram*[2], *M Mirbagheri*[1], *SH Mallidi*[2], *T Janu*[2], *P Rajan*[1], *N Mesgarani*[1,2], *M Elhilali*[2], *H Hermansky*[2], *S Shamma*[1], *R Duraiswami*[1]

[1]University of Maryland, College Park, MD, USA
[2]Johns Hopkins University, MD, USA

ABSTRACT

In recent years, there have been significant advances in the field of speaker recognition that has resulted in very robust recognition systems. The primary focus of many recent developments have shifted to the problem of recognizing speakers in adverse conditions, e.g in the presence of noise/reverberation. In this paper, we present the UMD-JHU speaker recognition system applied on the NIST 2010 SRE task. The novel aspects of our systems are: 1) Improved performance on trials involving different vocal effort via the use of linear-scale features; 2) Expected improved recognition performance in the presence of reverberation and noise via the use of frequency domain perceptual linear predictor and cortical features; 3) A new discriminative kernel partial least squares (KPLS) framework that complements state-of-the-art back-end systems JFA and PLDA to aid in better overall recognition; and 4) Acceleration of JFA, PLDA and KPLS back-ends via distributed computing. The individual components of the system and the fused system are compared against a baseline JFA system and results reported by SRI and MIT-LL on SRE2010.

Index Terms— Speaker recognition, LFCC, FDLP, Cortical, JFA, PLDA, KPLS, NIST SRE 2010

1. INTRODUCTION

Automatic speaker recognition is concerned with designing algorithms that infer the identity of people by their voices. This is a very challenging task since the speech signals are highly variable. Apart from carrying the speaker-specific characteristics, the speech data also encapsulates phonemic content, channel variability and inter-session variability. Also, it is subject to degradations due to noise and reverberation. Over the past decade, the field has made substantial progress in addressing these issues.

State-of-the-art speaker recognition systems have two important commonalities. First, they map a temporal sequence of feature vectors (typically MFCCs) into a fixed-size vector. Second, they assume that this fixed-size vector can be explicitly decomposed into a speaker-specific component and an undesired variability component (i.e., anything that does not capture speaker identity). Moreover, these two processes are performed in a data-driven way, where large amounts of data representative of the problem at hand are used to estimate the optimal parameters involved in each process. Examples of fixed-size representations are: GMM supervectors [1] and i-vectors [2]. Joint Factor Analysis (JFA) [1] applied to supervectors and Probabilistic Linear Discriminant Analysis (PLDA) [12] applied to i-vectors are examples of models that use explicit decompositions.

Another common trend in the filed is that of achieving robustness through diversity of representation, modeling, and classification. Fig. 1 shows the schematic of our system which is a good example of this principle. In particular, the overall architecture uses five complementary features [3, 4, 5, 6] that are transformed into speaker supervectors and i-vectors and used in three different classifiers [11, 12, 13]. The final recognition score is obtained by fusing the scores produced by each classifier via logistic regression [9]. Special attention has been placed in using feature extraction techniques that exhibit some inherent robustness to reverberation and noise [5, 6].

The paper is organized as follows. Sec 2 describes the voice activity detection (VAD) and Sec 3 introduces the different features. Sec 4 addresses the computation of GMM supervectors and i-vectors; Sec 5 describes the modeling and scoring techniques used in our back-end. The complete performance is evaluated on the SRE 2010 core-extended data in Sec 6 and the robustness of some features is briefly analyzed. Sec 7 concludes the paper.

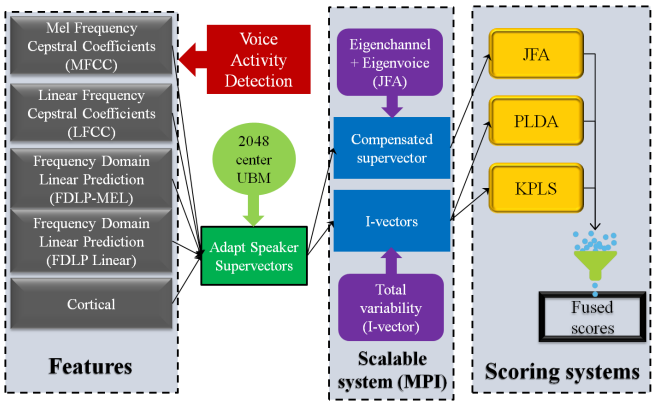

Fig. 1. [color] *Complete Schematic of the UMD-JHU system*

2. VOICE ACTIVITY DETECTION (VAD)

The VAD block of our recognition system determines the speech frames in a given utterance. It is based on phoneme posteriors derived from a multi-layer perceptron (MLP). The MLP is trained using modulation spectral features, where long temporal segments of the speech signal are analyzed in critical bands. In each sub-band, temporal envelopes are derived using the autoregressive modeling technique called frequency domain linear prediction (FDLP). The robustness of the sub-band envelopes is achieved by a minimum

This research was partially funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) through the Army Research Laboratory (ARL), Office of Naval Research (ONR), NSF award 0403313 and NVIDIA support for the Chimera cluster at the CUDA Center of Excellence at UMIACS. 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 ONR, ODNI, IARPA, NSF, NVIDIA or the U. S. Government.

mean square envelope estimation technique. The speech features are input to the trained MLP to estimate phoneme posterior probabilities. For the VAD, all the speech phoneme probabilities are merged to one speech class to derive speech/non-speech decisions. For more details refer to [10].

3. FEATURES EXPLORED

Mel-frequency cepstral coefficients (MFCC) [3] are the most widely used features in speaker recognition. Apart from the standard MFCC features, our system also consists of several novel feature representation to improve recognition in adverse conditions. All the representations described below produce 19 base components every 10ms, that, along with their deltas and double deltas, result in a 57-dimensional feature space.

Linear Frequency Cepstral Coefficients (LFCC) [4]: MFCCs have been dominantly used in speaker recognition as well as in speech recognition. MFCC uses a mel-warped frequency scale to mimic how the human ear processes sound. Its spectral resolution becomes lower as the frequency increases. Therefore, the information in the higher frequency region is down-sampled by the mel scale. However, based on theory in speech production, some speaker characteristics associated with the structure of the vocal tract, particularly the vocal tract length, are reflected more in the high frequency region of speech. Therefore, the mel-scaling of MFCC is replaced with a linear scale in this feature representation [4].

Frequency Domain Linear Predictors (FDLP) [5]: The performance of a typical speaker verification system degrades significantly in reverberant environments. This degradation is partly due to the conventional feature extraction/compensation techniques that use analysis windows which are much shorter than typical room impulse responses. To address this, we use a feature extraction technique that estimates long-term envelopes of speech in narrow sub-bands using frequency domain linear prediction (FDLP). When speech is corrupted by reverberation, the long-term sub-band envelopes are convolved in time with those of the room impulse response function. To a first order approximation, gain normalization of these envelopes in the FDLP model attenuates the room reverberation artifacts [7]. Here too, features based on both the linear and mel-scales are used independently.

Cortical representation: Humans exhibit a remarkable ability to reliably classify sound sources in their environment even in the presence of high levels of noise, while most engineering systems suffer from a drastic drop in performance with channel or background distortions. Our brains are equipped with elaborate machinery for speech analysis and feature extraction, which holds great lessons for improving front-end schemes used in automatic systems. One of the intriguing processes taking place in the central auditory system involves an ensemble of neurons with variable tuning to spectral profiles of acoustic signals. In addition to the frequency (tonotopic) organization emerging as early as the cochlea, neurons in the central auditory system (in the midbrain, more prominently in the auditory cortex) exhibit tuning to a variety of filter bandwidths and shapes. This elegant neural architecture provides a detailed multi-resolution analysis of the spectral sound profile, which is presumably relevant to speech and speaker recognition. The cortical representation in our system is simple, effective, computationally-efficient and is carefully optimized to be particularly sensitive to the information-rich spectro-temporal attributes of the signal while maintaining robustness to unseen noise distortions. The choice of model parameters builds on the current knowledge of psychophysical principles of speech perception in noise [8] complemented with a statistical analysis of the dependencies between spectral details of the message and speaker information. More details are available in [6].

4. FIXED-SIZE REPRESENTATIONS

Our system uses both GMM supervectors [1] and i-vectors [2] to capture speaker specific-information. For each of the five feature sets described above we train a collection of gender-dependent GMM Universal Background Models (GMM-UBM) based on a development set comprising data from: NIST SRE 2004, 2005, 2006, 2008, Switchboard phases 2 and 3, Switchboard-Cellular parts 1 and 2 and Fisher (total of $17,319$ male and $22,256$ female utterances). The details about the specific size and structure of the UBMs follow.

GMM supervectors: Given a sequence of d-dimensional feature vectors from an utterance, the N-component GMM-UBM is used to collect Baum-Welch sufficient statistics. A GMM supervector is created by appending together the first order statistics into a vector of dimension dN. Two different sets of supervectors were created. One for JFA modeling and another for i-vector extraction. For the first set, a diagonal covariance UBM with 2048 mixtures was used, except for the cortical features where using 4096 was found beneficial. The supervectors used for i-vector extraction were computed using a 2048 mixture full-covariance UBM.

i-Vectors: In order to compute an i-vector, a GMM supervector is projected into a lower-dimensional subspace (400 dimensions in our system). This is accomplished by using a factor analysis model for the supervector s as [2]:

$$s = m + Tw, \tag{1}$$

where m is a global offset (usually a supervector from the means of the UBM), T is a low-rank $Nd \times 400$ dimensional matrix that spans the subspace where most of the speaker-specific information lives (along with channel variability), and w is a normally-distributed latent variable. The MAP point estimate of the vector w is called the *i-vector*. In our system, the T matrix was estimated in a gender dependent fashion using the same data as the UBM.

5. MODELING AND SCORING

Our system uses two probabilistic generative models, namely, JFA and PLDA along with a discriminative system based on kernel PLS to produce verification scores. In the following we provide details about each one of them.

Joint Factor Analysis: JFA provides an explicit mechanism to model the undesired variability in the speech signal. It decomposes the speaker supervector as

$$s = m + Ux + Vy + Dz, \tag{2}$$

where $\{m, U, V, D\}$ are the hyper-parameters of the JFA model, which are estimated via Expectation Maximization (EM). The key idea in the JFA technique is to find two subspaces (V and U) that best capture the speaker and the channel variabilities in supervector space. In our system, we use the training and inference algorithms for JFA as described in [11]. Gender-dependent U and V matrices are learned with 300 and 100 dimensions respectively from a subset of the UBM training data. The final scores are produced by linear scoring [14] and normalized by channel and gender dependent ZT-norm.

Probabilistic LDA [12]: The PLDA model can be considered as a particularization of the JFA formulation for a single Gaussian applied in i-vector space. However, due to the sufficiently low dimensional nature of the i-vector space (e.g., 400), it is common practice (see [12]) to use a single full-covariance term to model both the eigenchannel and residual variability terms. In this way an i-vector w is modeled as:

$$w = \mu + \Phi\beta + \epsilon. \tag{3}$$

In particular, μ is a global offset; the columns of Φ provide a basis for the speaker-specific subspace (eigenvoices); β is a latent identity vector having a standard normal distribution; and ϵ is a noise term assumed to be Gaussian with zero mean and full covariance Σ. For our system, maximum likelihood point estimates of the model parameters $\{\mu, \Phi, \Sigma\}$ where obtained from the same data as the UBM but removing the Fisher subset.

Given two i-vectors w_1 and w_2, PLDA defines two hypotheses $\mathcal{H}_s$ and $\mathcal{H}_d$ indicating that they belong to the same speaker or to different speakers respectively. The score is then defined as the log-likelihood ratio between two Gaussian distributions $\log \frac{p(w_1,w_2|\mathcal{H}_s)}{p(w_1,w_2|\mathcal{H}_d)}$ whose mean and covariance are defined by the PLDA hyperparameters $\{\mu, \Phi, \Sigma\}$. From our experiments, we found that using a Φ of rank 200 produced the best results.

Kernel partial least squares: Partial least squares (PLS) is a subspace based learning technique that has been used for dimensionality reduction as well as a regression and is popular due to its ability to handle learning where the data has a very low rank. In our system we used a kernelized PLS version in i-vector space [13]. We use a cosine kernel for the kernel PLS (KPLS) classifier and this makes the speaker scoring equivalent to a linear combination of the cosine scores between the testing data i-vector and the combination of target speaker and development data i-vectors. For each target speaker in the dataset, we learn a specific linear combination during training. More details can be found in [13].

Parallelization: The resources necessary for training the subspace matrix T or JFA eigen-voices V and eigen-channel U pose a challenge for both memory and computational scalability. The EM algorithm requires the zero and first order sufficient statistics (Baum-Welch) for each utterance along with the inverse covariance matrices of the UBM centres. The number of training utterances varies from 10^4 to 10^6 with typically 10^3 to 2×10^4 speakers. Storing and processing the data in memory on a single node is infeasible leading to the need for a distributed approach. For our system we use a tiered model: at the highest level, the message passing interface (MPI) provided coarse-grain parallelism that decomposed the speaker data across multiple nodes. On each node, we parallelized over the assigned speaker data using OpenMP under a shared memory architecture. At the lowest level, we obtained fine-grain parallelism using optimized multi-threaded linear algebra routines via the Intel Math Kernel Library (MKL).

6. EXPERIMENTS

All the experiments were conducted on the core NIST-SRE10 evaluation dataset using the extended-trial list for all 9 conditions. The main goal was to analyze the performance of each feature set with respect to all the classifiers as well as to evaluate their potential to complement each other by performing score fusion. Also, we paid special attention to using feature extraction techniques that exhibited some inherent robustness to additive noise and reverberation.

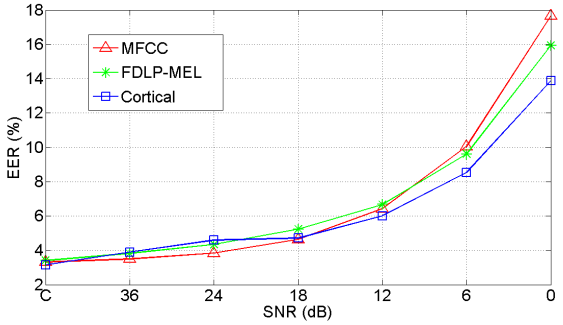

Fig. 2. [color] *Effects of white noise on condition 5 of SRE10 for JFA system with 3 different features*

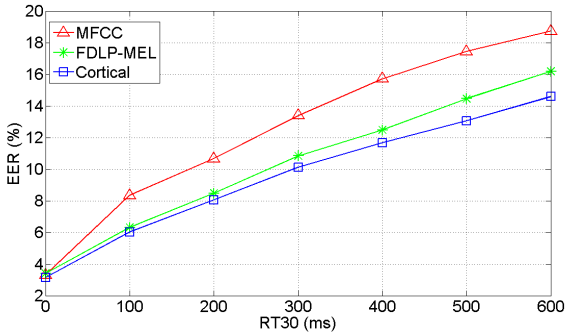

Fig. 3. [color] *Effects of reverberation on condition 5 of SRE10 for JFA system with 3 different features*

Fig 2 shows the performance of the JFA system on condition 5 of SRE10 when the test segments were corrupted by white noise at different levels of SNR. FDLP-MEL as well as the cortical features were compared to the standard MFCCs. While the performance in clean conditions is very similar, the robustness of FDPL and cortical representation is mostly noticed for SNRs below 12 dB. Also, Fig 3 shows the results of the same setup when the test segments are corrupted by reverberation. The corruption was simulated by convolving the original signals with synthetic room impulse responses. Both, FDLP-MEL and cortical features show significant robustness to reverberation compared to MFCCs.

We compare each feature (MFCC, LFCC, FDPLP-Mel, FDPLP-Linear and Cortical) in our systems using each of the classifiers (JFA, PLDA, KPLS). The corresponding EER and the NIST 2010 DCFs are shown in Table 1. It can be seen that the linear scale FDLP and cepstral have significant performance difference in the vocal effort conditions (C6 $-$ 9). FDLP and PLDA have the best performance among the features and classifiers respectively, closely followed by LFCC and KPLS. To asses how features and classifiers complemented each other, we created a two-fold partition of the evaluation data and learned a linear combination of all the systems and features using logistic regression for each partition. The weights learned from partition one were used for partition two and vice versa. It can be seen that all conditions benefited from the fusion. For reference, we also include the best performance reported in [15] and [16], and our fused system outperforms the reported results in all conditions (except Condition 5 of [15]).

7. CONCLUSIONS

This paper presents the speaker recognition system designed for adverse conditions developed at UMD-JHU and reports its perfor-

Systems	C1	C2	C3	C4	C5	C6	C7	C8	C9
#TGT	4,304	15,084	3,989	3,637	7,169	4,137	359	3,821	290
#NTGT	795,995	2,789,534	637,850	756,775	408,950	461,438	82,551	404,848	70,500
[15]	1.89/0.32	3.04/0.15	3.15/0.125	–/–	2.03/0.38	–/–	–/–	–/–	–/–
[16]	–/0.43	–/0.51	–/0.47	–/0.39	–/0.47	–/0.80	–/0.86	–/0.45	–/0.27
JF+MC	2.71/0.50	4.53/0.60	3.84/0.53	3.69/0.61	3.31/0.50	6.10/0.82	7.77/0.93	2.59/0.44	1.99/0.42
PL+MC	1.49/0.23	2.84/0.43	2.71/0.50	2.31/0.36	2.53/**0.41**	**4.38**/0.71	6.23/0.69	2.05/0.41	1.48/0.27
KP+MC	1.51/0.26	3.21/0.48	3.79/0.54	2.30/0.40	3.37/0.45	6.10/0.79	6.82/0.76	2.63/0.49	1.13/0.35
JF+LC	2.47/0.44	3.52/0.50	3.58/0.45	2.88/0.51	3.45/0.48	4.96/0.69	4.56/0.80	2.30/0.44	2.17/0.46
PL+LC	1.68/0.21	2.50/0.35	2.53/0.48	2.22/0.28	2.60/0.44	4.40/**0.68**	**4.38**/0.63	**1.86**/**0.38**	1.48/**0.24**
KP+LC	1.59/0.26	2.66/0.41	3.26/0.52	2.01/0.31	3.35/0.47	5.31/0.71	4.83/0.66	2.61/0.47	1.48/0.28
JF+FL	2.12/0.39	2.62/0.42	2.95/**0.40**	2.33/0.41	2.89/0.50	4.94/0.74	5.34/0.77	2.32/0.44	1.69/0.32
PL+FL	**1.26**/0.20	2.27/**0.29**	**2.13**/0.43	1.84/0.28	**2.43**/0.46	4.62/0.75	4.56/**0.56**	2.25/0.41	1.13/0.26
KP+FL	1.26/**0.18**	**2.10**/0.30	2.86/0.45	**1.56**/**0.26**	3.10/0.48	6.36/0.79	4.83/0.66	2.77/0.52	**0.76**/0.29
JF+FM	2.31/0.46	3.34/0.49	3.89/0.48	3.23/0.53	3.41/0.46	7.01/0.82	7.90/0.88	3.03/0.48	1.82/0.39
PL+FM	1.68/0.25	2.86/0.40	2.76/0.51	2.40/0.37	2.68/0.45	5.32/0.81	6.07/0.69	2.25/0.47	1.48/0.39
KP+FM	1.35/0.22	2.42/0.38	3.56/0.53	1.83/0.34	3.42/0.48	6.94/0.85	5.95/0.73	3.08/0.54	1.13/0.29
JF+CC	3.02/0.54	3.79/0.59	3.38/0.50	3.94/0.63	3.15/0.45	6.68/0.88	9.57/0.97	2.69/0.43	2.51/0.42
PL+CC	1.73/0.26	2.55/0.44	2.91/0.63	2.71/0.44	3.00/0.49	6.38/0.89	7.34/0.91	2.22/0.47	1.48/0.26
KP+CC	1.63/0.26	2.69/0.46	3.84/0.59	2.43/0.42	3.42/0.48	8.32/0.97	7.62/0.97	2.87/0.52	0.89/0.30
Fused	**1.00**/**0.14**	**1.47**/**0.21**	**1.62**/**0.31**	**1.32**/**0.21**	**2.11**/**0.36**	**3.48**/**0.63**	**3.16**/**0.55**	**1.37**/**0.33**	**0.44**/**0.14**

Table 1. *Equal error rate (EER) and* 2010 *detection cost function (DCF) values (shown as EER/DCF) obtained using Joint Factor Analysis (JF), Probabilistic Linear Discriminant Analysis (PL) and Kernel Partial Least Squares (KP) with MFCC (MC), LFCC (LC), FDLP-linear (FL), FDLP-mel (FM) and Cortical (CC) features for the NIST SRE 2010 extended core data set. For the results reported from Refs. [15] and [16], "–" indicate the unreported conditions.*

mance on the standard SRE 2010 extended core tests. The main focus for obtaining robustness was the use of a diverse set of inherently robust features. A very complete characterization of 5 different features and three different classifiers was presented. The fused system has performance comparable or superior to other systems reported on this task.

8. REFERENCES

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1448 –1460, 2007.

[2] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 19:788 –798, 2011.

[3] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:357 – 366, 1980.

[4] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel- frequency cepstral coefficients for speaker recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[5] S. Ganapathy, S. Thomas, and H. Hermansky, "Front-end for Far-Field Speech Recognition based on Frequency Domain Linear Prediction," in *INTERSPEECH*, 2008.

[6] S.K. Nemala, D.N. Zotkin, R. Duraiswami, and M. Elhilali, "Biomimetic multi-resolution analysis for robust speaker recognition," *EURASIP Journal on Audio, Speech, and Music Processing (submitted)*.

[7] S. Ganapathy, J. Pelecanos, and M.K. Omar, "Feature normalization for speaker verification in room reverberation," in *ICASSP*, 2011.

[8] D.N. Zotkin, T. Chi, S.A. Shamma, and R. Duraiswami, "Neuromimetic sound representation for percept detection and manipulation," *EURASIP Journal of Advanced Signal Processing*, pp. 1350–1364, 2005.

[9] N. Brummer and J. du Preez, "Application-independent evaluation of speaker detection," in *Proc Odyssey Speaker and Language Recognition Workshop*, 2006, vol. 20, pp. 230 – 275.

[10] S. Ganapathy, P. Rajan, and H. Hermansky, "Multi-layer perceptron based speech activity detection for speaker verification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.

[11] D. Garcia-Romero and C. Espy-Wilson, "Joint factor analysis for speaker recognition reinterpreted as signal coding using overcomplete dictionaries," in *Proc Odyssey Speaker and Language Recognition Workshop*, June 2010.

[12] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011.

[13] B.V. Srinivasan, D Garcia-Romero, D.N. Zotkin, and R. Duraiswami, "Kernel partial least squares framework for speaker recognition," in *INTERSPEECH*, 2011.

[14] 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 of ICASSP 2009,* pp. 4057-4060, April 2009.

[15] D. Sturim, W. Campbell, N. Dehak, Z. Karam, A. McCree, D. Reynolds, F. Richardson, P. Torres-Carrasquillo, and S. Shum, "The MIT LL 2010 speaker recognition evaluation system: Scalable language-independent speaker recognition," in *ICASSP*, 2011.

[16] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system," in *ICASSP*, 2011.