Toward Fail-Safe Speaker Recognition: Trial-Based Calibration with a Reject Option

Luciana Ferrer, Mahesh Kumar Nandwana, Mitchell McLaren, Diego Castan, Aaron Lawson

***Abstract*—The output scores of most speaker recognition systems are not directly interpretable as stand-alone values. For this reason, a calibration step is usually performed on the scores to convert them into proper likelihood ratios (LR), which have a clear probabilistic interpretation. The standard calibration approach transforms the system scores using a linear function trained using data selected to closely match the evaluation conditions. This selection, though, is not feasible when the evaluation conditions are unknown. In previous work, we proposed a calibration approach for this scenario called trial-based calibration (TBC). TBC trains a separate calibration model for each test trial using data that is dynamically selected from a candidate training set to match the conditions of the trial. In this work, we extend the TBC method, proposing (1) a new similarity metric for selecting training data that results in significant gains over the one proposed in the original work, (2) a new option that enables the system to reject a trial when not enough matched data is available for training the calibration model, and (3) the use of regularization to improve the robustness of the calibration models trained for each trial. We test the proposed algorithms on a development set composed of several conditions and on the FBI multi-condition speaker recognition dataset, and we demonstrate that the proposed approach reduces calibration loss to values close to 0 for most conditions when matched calibration data is available for selection and that it can reject most trials for which relevant calibration data is unavailable.**

***Index Terms*—Speaker Recognition, Trial-based Calibration, Forensic Voice Comparison**

I. Introduction

Speaker recognition has become a critical tool in a myriad of domains including forensic voice comparison, user authentication, and speaker search. Speaker recognition systems output a score for each trial composed of two sets of recordings: the enrollment or "known-speaker" recordings and the test or "questioned-speaker" recordings. The goal of calibration is converting these scores into proper likelihood ratios, which provide an interpretable value that can be used directly in some applications, like forensic voice comparison, or converted to binary decisions using Bayes rule for other applications, like user authentication. The likelihood ratio (LR) for a speaker recognition trial is given by the likelihood of the two recordings in the trial given the hypothesis that they come from the same speaker divided by the likelihood of the two recordings given the hypothesis that they come from different speakers.

Luciana Ferrer is with Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-Universidad de Buenos Aires, Buenos Aires, Argentina, e-mail: lferrer@dc.uba.ar.

All other authors are with the Speech Technology and Research Laboratory, SRI International, California, USA.

The usual procedure for calibration is to transform the scores using a linear function with trainable parameters. The parameters are trained to optimize an objective function that measures the quality of the resulting LRs using a set of trials with corresponding scores. These trials must be representative of the conditions of the test trials in order to guarantee that the resulting LRs reflect the actual distribution of scores under those conditions.

When test conditions are known ahead of time, the user of the system can attempt to select representative data for those conditions and use it to train a calibration model. If the selected data is indeed a good match to the test data, this procedure works quite well. This method is commonly used in speaker recognition evaluations organized by NIST, where a small set of predefined conditions is considered and clearly described in the evaluation plans. Representative development data is usually available for those conditions. In these cases, developers can train a separate calibration model for each evaluation condition, leading to reasonable calibration performance.

The problem arises when the test conditions are unknown and potentially differ among trials, as will likely occur outside controlled evaluations. In this case, pre-training a matched calibration model is not possible. One approach that can be used for these cases is to train a calibration model using trials from many different conditions, in the hope that the model will generalize to the test conditions. This, though, is not necessarily optimal, since the best calibration model is usually condition dependent. Estimating a single model for all conditions then leads to suboptimal results within each condition, which, in turn, leads to a suboptimal global result.

Several approaches have been proposed to tackle this problem. One family of approaches involves extracting meta or auxiliary information about the two recordings in a trial. This information, which can be discrete or continuous, is then used as input for the calibration process. The idea is that a calibration model that takes into account this meta-information will be able to accommodate all acoustic conditions, or at least those with which the meta-information extractor was trained. In [1], we proposed a method that uses meta-information from the enrollment and test utterances, such as the duration of utterances, estimated channel type, and speaker gender. The meta-information for the enrollment and test conversations from each trial was clustered, and separate combiners were trained for each resulting cluster. Solewicz [2], [3] proposed a very similar method for performing combination using attributes obtained from the utterances. Another similar approach was proposed in [4] where discrete

meta-information is used to condition the parameters of a regularized linear logistic regression model. Finally, in [5], continuous vectors of meta-information are used as input to the calibration process. A bilinear function of the scores and the meta-information vector for each sample in the trial is optimized by using a logistic objective function.

A novel approach for condition-dependent calibration, called trial-based calibration (TBC), was proposed in [6]. The approach is based on training a separate calibration model for every new trial, similar to the way in which a forensic expert would aim to calibrate each trial individually. The model is trained using a subset of the available development data selected to match as closely as possible the characteristics of the test trial. This method, while significantly costlier in terms of computation, offers the advantage of being more flexible than all methods previously described. The model obtained for each trial is targeted to the characteristics of the particular trial. If the test data contained a discrete number of conditions that were well represented in the development data, then the TBC approach could be implemented by pre-training a discrete set of models which would then be retrieved for each trial depending on its condition. In this case, the method coincides with the one described in [4]. Yet, in the general case, each test trial may be slightly different from all other trials and require selecting a different subset of the development data for training the calibration model.

All methods described above were designed to calibrate all possible trials, regardless of whether the trial's conditions are well represented in the development data. This means that if a trial's conditions are not represented in the development data, the system's output will quite likely be poorly calibrated, and the user may not be aware that the scores output by the system, in this case, are not amenable to interpretation. That is, the LR output by the system will not be a proper LR reflecting the distribution of scores for the conditions of the trial. While generating an output for every possible trial might be acceptable or even necessary for some practical scenarios, this is not always the case. For example, in the forensic domain, a large error in the LR value may have a very high cost for the defendant. This scenario must be avoided, even if it means discarding the audio recordings as evidence in the trial, as explained by Morrison [7] and Schwartz [8]. Schwartz states: "In forensic speaker comparison, it is crucial to decide when completion of the examination may not be possible." She makes a crucial distinction between rendering a decision, even if that decision is "result inconclusive", and determining that the conditions of a trial are such that automatic speaker recognition is not feasible. She makes it clear that, in her practice, a substantial number of comparison trials must be rejected a priori. In this paper, we make a first attempt at automating the rejection of trials for which calibration cannot be reliably performed.

The novel contributions of this paper are as follows: We propose a new similarity metric used to select development trials for training each calibration model in TBC. This new similarity metric provides significant gains over the one proposed in the original TBC paper. We also propose using a reject option when not enough data similar to a test trial is found for calibration. Finally, we introduce the use of regularization toward a global model when training the calibration model for each test trial, enabling us to select fewer and, hence, more matched, samples for calibration. We thoroughly test the new contributions on a development set composed of several different data sets and a held-out FBI dataset.

II. Calibration with Linear Logistic Regression

Consider a training set with M samples, $S = \{(x_i, y_i); i = 1, ..., M\}$, where x_i is the score output by the system for trial i and $y_i \in \{-1, +1\}$ is the class corresponding to the trial (-1 for an *impostor* trial, where the two samples correspond to different speakers, and +1 for a *target* trial, where the samples correspond to the same speaker). Our goal is to transform the score x_i into a proper log-likelihood ratio (LLR) $\log(p(x_i|y_i = 1)/p(x_i|y_i = -1))$. In linear logistic regression, the LLR is assumed to be a linear function of the scores. That is, we wish to estimate α and β such that $\alpha x_i + \beta$ is as close to the proper LLR as possible. This is done by minimizing the negative log-likelihood of the data given by

$$\mathcal{L}(\alpha, \beta) = \sum_{i=1}^{M} \frac{p(y_i)}{N(y_i)} \log(1 + e^{-y_i(\alpha x_i + \beta)}) \qquad (1)$$

where we assume equal prior probabilities for both classes to obtain the posterior from the LLR (so that $P(y_i|x_i) = 1/(1 + e^{-y_i(\alpha x_i + \beta))})$ and where we have balanced the effect of the positive and negative samples by weighting the log-likelihood corresponding to each sample by the inverse of the total number of samples for the sample's class, $N(y_i)$, times an effective prior probability $p(y_i)$ (with $p(y_i = 1) = 1 - p(y_i = -1)$), which can be set depending on the operating point of interest. See, for example, [9], for a discussion on the effect this parameter has on the model's performance.

In this work, we also consider a regularized version of linear logistic regression where a term is added to the objective function to penalize the distance from the estimated parameters to a default set of parameters. The use of regularized linear logistic regression for calibration of speaker recognition systems was studied in several works (e.g., [4], [10]). Here we use a version where we regularize toward default parameter values. That is, we maximize the following objective function

$$\mathcal{L}_R(\alpha, \beta) = \mathcal{L}(\alpha, \beta) + \lambda \mathcal{L}_0 \left[\frac{(\alpha - \alpha_0)^2}{\alpha_0^2} + \frac{(\beta - \beta_0)^2}{\beta_0^2} \right] \qquad (2)$$

where $\mathcal{L}_0 = \mathcal{L}(\alpha_0, \beta_0)$ and is used to multiply λ in an effort to make that parameter easier to tune. The default values for the parameters, α_0 and β_0 can be taken to be 1.0 and 0.0, respectively, or the values learned on a separate dataset. The value of λ is chosen empirically to optimize calibration performance.

III. Trial-based Calibration

The goal of TBC, first proposed in [6], is to customize the model used to calibrate each test trial to the exact conditions of the enrollment and test samples. To this end, enrollment and test samples are selected from the available calibration data

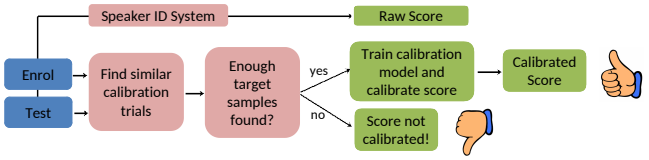

Fig. 1. The proposed TBC approach. Raw scores are computed for the test trial (enrollment and test samples) using a speaker recognition system. Calibration samples are selected based on their similarity to the enrollment and test samples using a minimum similarity threshold. If enough calibration target trials are selected, a calibration model is learned and applied to the trial. Otherwise, the score is not calibrated.

based on their similarity to the enrollment and test samples in the trial. Finally, calibration trials are defined by pairing up all selected enrollment and test samples and a model is trained with the scores for these trials using linear logistic regression. The test trial is then calibrated with the obtained parameters. In the current work we propose a modification to the original procedure proposed in [6] that includes the option to reject a test trial if not enough calibration trials are found that are similar to the test trial. Figure 1 depicts the revised TBC process.

Trial-based calibration shares some aspects with adaptive symmetric normalization (AS-Norm) [11]. This technique is an adaptive version of symmetric normalization (S-Norm) [12] where the statistics used for normalization are obtained from data selected to match the characteristics of the trials' samples. In [13] we study the use of calibration and normalization for different datasets. We found that, in most cases, the best performance is achieved when both normalization and TBC are performed. Hence, we believe that the gains obtained from TBC in this paper would still apply on normalized scores. Since the implementation of both normalization and calibration poses some questions (which normalization method is optimal for each calibration approach, how to split available data for each purpose, etc.) we decided to leave the question on the interaction between normalization and calibration for a later work, focusing here on the main contributions of this paper: the new similarity metric, the rejection option, and the regularization approach.

A. Similarity Metrics

The key step in the TBC approach is the selection of calibration trials. This selection is done based on a similarity metric between the trial's samples and the candidate calibration samples. In this work, we evaluate three different metrics.

I-vector (IV) similarity: The i-vector similarity is given by the cosine similarity between the i-vectors corresponding to the two samples. The cosine similarity is computed as the dot product between two vectors divided by the product of the Euclidean norms of the vectors.

UAC-based (UAC) similarity: This was the metric proposed in the original TBC paper. A set of universal audio characterization (UAC) models [14] are trained to predict different aspects of an audio signal, which could include gender, language, channel, noise type and level, and so on. Each UAC model consists of a Gaussian model for each class to be predicted (for example, female and male for the gender UAC model) with shared covariance matrix. The features modeled by the UAC are the same i-vectors extracted by the speaker recognition system. The output of each UAC model is the posterior probability for each of the corresponding classes. The posteriors from all UAC models are concatenated and each component is normalized by replacing it with the ranking of the value. The ranking of each value is computed as its index in the sorted vector of all the values for that component in the calibration data. Finally, the similarity between two samples is given by the dot product between the rank-normalized vectors of UAC posteriors for the samples. A detailed explanation of this metric is given in [6].

Condition PLDA (CPLDA) similarity: The condition probabilistic linear discriminant analysis (PLDA) similarity proposed in this work is given by the score produced by a PLDA model trained to estimate the log-likelihood ratio of the samples' i-vectors given the hypothesis that the two samples come from the same condition versus the hypothesis that they come from different conditions. The model is trained with data from many different speakers under many different conditions. Conditions are given by the cross-product of the samples' gender, language, channel, noise type, noise dB level, etc.

The UAC and the CPLDA metrics are conceptually similar in that they are trained to be independent of the similarity between the speakers in the two samples, focusing on the similarity between their conditions. The IV metric, on the other hand will be sensitive to both the speakers and the conditions in the two samples. In principle, this is not a desirable characteristic of a metric for our purposes. We wish to select calibration data that is similar to the test trial in terms of condition so that any effect of the condition on the score can be neutralized by the calibration procedure. If the similarity metric is affected by the characteristics of the voices in the test samples, then we will tend to select calibration data that is similar to the test samples in terms of voice. In the extreme, if the voices in the enrollment and test sides of the trial are very different, we would not be able to find enough target calibration trials that are similar to the test trial for enrollment and test sides. As a consequence, we would not be able to calibrate the easier impostor samples (those for which the voices are so different that the system is quite confident on its decision, leading to a very large negative score) for lack of enough target samples to train the model. This may not be an issue in some forensic applications where only trials that are difficult reach the forensic expert [7]. In this case, selecting calibration data that is similar to the test trial both in terms of conditions and voices would be appropriate. In this work, we will not consider this case, because our test data is not restricted to difficult impostor trials. Nevertheless, as we will see, the IV metric performs quite well, indicating, as is well known, that the dot product between i-vectors is highly affected by the conditions in the samples.

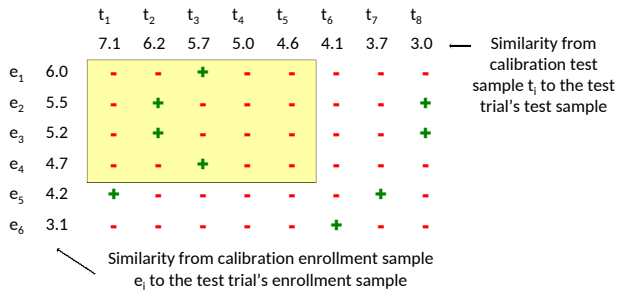

Fig. 2. A toy example of the selection process. The rows of the matrix correspond to the enrollment calibration samples, sorted by their similarity to the enrollment sample in the test trial. The columns correspond to the test calibration samples sorted by their similarity to the test sample in the test trial. The "-" and "+" symbols correspond to impostor and target trials, respectively. The highlighted trials would be selected for a similarity threshold of 4.5 if MaxTgt is larger than 4.

B. Selection Algorithm

The selection of calibration data is governed by three parameters: (1) a similarity threshold (SimThr), (2) a maximum number of target trials (MaxTgt), and (3) a minimum number of target trials (MinTgt). It also takes as input the calibration set, composed of enrollment and test samples. Given the parameters, the calibration set, and a given trial to be scored, composed of an enrollment sample, E, and a test sample, T[1], the selection algorithm proceeds as follows: The similarity values from E to all enrollment samples in the calibration set and from T to all test samples in the calibration set are computed. All calibration enrollment samples with a similarity to E larger than a certain threshold t are selected. Similarly, all calibration test samples with a similarity to T larger than t are selected. The trials obtained by pairing all selected enrollment samples against all selected test samples are used to train a calibration model for the trial to be scored. The threshold t is equal to the maximum between SimThr and the threshold that would result in MaxTgt target trials being selected. Hence, when the largest of these two thresholds is SimThr, fewer than MaxTgt target trials are selected. If the number of selected target samples is smaller than MinTgt then calibration of the trial is not attempted. Figure 2 shows a toy example for the selection process.

Note that our decision to reject a trial when not enough matched calibration data can be selected could be replaced by other strategies. For example, either the global calibration model could be used to calibrate that trial, or the LLR could be set to 0.0. Yet, this would mean that the resulting LLR might be very different from what a well-calibrated LLR would have been. Our goal in this work is to never allow a trial to be wrongly calibrated, even if this is done at the cost of rejecting some trials. Ultimately, we aim to only calibrate trials for which we can achieve close to the matched calibration performance. If the system believes this is not possible, then the trial would not be calibrated.

[1]In this work, we will only consider single-enrollment single-test trials.

C. Training of Calibration Model

Once the calibration trials have been selected for a certain test trial, a model is trained using regularized linear logistic regression. The default parameters for regularization, α_0 and β_0, are given by the parameters trained using all available calibration data. As we will see, the use of regularization enables us to reduce the MinTgt parameter, which in turns allows us to properly calibrate a larger percentage of samples when the amount of matched data to a certain trial is restricted.

D. Duration-Dependent Calibration

The duration of the samples was shown to have a significant effect in calibration performance in [15]. Our preliminary experiments confirmed that this is also the case in our development data, in particular when the samples are relatively clean and short. Speech duration can be directly obtained from the speech activity detection output. Further, by chopping each sample in the calibration data to a predefined set of durations (see Section VI), we can create a multi-duration calibration set. Hence, in our experiments, when the test data has variable speech durations, we use a two-stage calibration approach. First, for each side in a test trial, the calibration data that best matches its duration is selected. Second, we apply global calibration or TBC on this matched-duration subset of the data only.

IV. Experimental Design

We present results for the proposed approaches on two datasets: (1) a large and varied development set where all parameters are tuned, and (2) an FBI dataset composed of several different language and microphone conditions for independent evaluation of the chosen configuration.

A. Development Data

We tune all system's parameters using a collection of conditions where each condition is designed to be highly homogeneous. This homogeneity allows us to define, for each condition, a highly matched calibration set. This, in turn, enables us to compute the matched calibration performance for each condition, which can be used as a baseline to compare the performance of the proposed algorithms.

The conditions used for development are all restricted to male speakers and the waveforms are chopped to contain approximately 20 seconds of speech as determined by our speech activity detection system. The conditions are:

SWCELLP1: Switchboard Cellular Part 1, consisting of cellphone conversations [16].

SWPH2: Switchboard 2 Phase 2 samples, consisting of telephone conversations [17].

FVC-int: Interviews from Australian English speakers from the forensic voice comparison dataset [18].

COD: Test samples in this set were obtained by transcoding the FVC-int test samples through a GSM codec. The enrollment samples were not degraded.

REV: Test samples in this set were obtained by adding reverberation to the FVC-int test samples. The reverberation

impulse response corresponded to a large room (the Wangenheim Rare Books Room at the San Diego's Central Public Library) and was added using the FCONV tool [19]. The enrollment samples were not degraded.

NOI: Test samples in this set were obtained by adding babble noise to the FVC-int test samples using the FANT tool [20]. The same noise signal was added to all samples. The enrollment samples were not degraded.

RATS-G: Subset of the RATS speaker verification data [21] including only Pashto language and retransmitted channel G.

Each of these sets is divided equally into calibration and test splits keeping approximately half of the speakers in each split. The number of speakers in each set and each split ranges from 126 to 340. The number of trials in each split ranges between 300 and 1000 target trials and 34,000 and 65,000 impostor trials.

B. Evaluation Data

We evaluate a small number of final configurations for the calibration system using the FBI multi-condition dataset first described in [6], calibrated with matched data from the same set or with mismatched data from the large variability dataset (LVD). Both sets are described below.

FBI: The FBI evaluation corpus was supplied by the Federal Bureau of Investigation (FBI) and consists of 14 distinct conditions including same/cross-channel and same/cross-language trials. One of the languages in the cross-language trials is always English while the other language varies. The data is sourced from several different corpora: LASR [22] (LA), PanArabic [23] (PA), Nist99 [24] (N99), NoTel [25] (NT), CrossInt (CI), Cavis (CA), ABSpanish (AB), and CHArabic (CH). The CHArabic dataset contains interviews in Arabic with a studio quality cardioid microphone about 1 meter from the target speaker. The ABSpanish set contains Spanish and English data also collected with a cardioid studio microphone. Cavis is a conversational dataset in English collected with both a studio microphone and a telephone microphone. CrossInt contains speech from 3000 speakers across three conditions (landline telephone, cellular and live room microphone), with two different speaking styles (interview and spontaneous conversation) using bilingual participants in India to allow for cross-language and same language trials for all the conditions. Languages were English for the first session of the collect and the participants native language (Hindi, Gujarati, Bengali, Marathi, Tamil, Kannada and Telugu) for the second session. The NoTel corpus contains telephone recordings from naturally noisy locations in Indian accented English. The PanArabic set contains speech from 100 subjects in five different Arabic dialects, recorded in a studio using a lapel microphone. The NIST99 set corresponds to the data used in the 1999 speaker recognition evaluation organized by the National Institute of Standards and Technology (NIST) and contains telephone conversations in English from approximately 600 subjects. Finally, the LASR corpus is composed of data from 100 bilingual speakers from each of three languages: Arabic, Korean and Spanish. Each speaker is asked to perform a series of tasks in two sessions recorded on different days using several devices.

TABLE I

CHANNELS, LANGUAGES, NUMBER OF MALE AND FEMALE SPEAKERS AND SOURCES INCLUDED IN EACH OF THE FBI CONDITIONS USED IN THIS WORK. THE CHANNELS ARE STUDIO MICROPHONE (MIC), TELEPHONE (TEL) OR CELLPHONE (CELL). CROSS INDICATES CROSS-LANGUAGE TRIALS FOR WHICH THE LANGUAGE OF ONE OF THE TWO SIDES OF EACH TRIAL IS ENGLISH AND THE OTHER LANGUAGE DEPENDS ON THE SOURCES INCLUDED IN THE CONDITION.

Cond	Chan(s)	Lang(s)	#Male	#Fem	Source Corpora
02	Mic	Arabic	422	280	PA, LA, CH
03	Mic	Cross	179	193	LA, AB
05	Tel	English	467	519	LA, NT, N99, CA
06	Tel	Cross	597	264	CI, LA
08	Cell	Cross	460	97	CI
09	Mic, Tel	English	645	264	CI, LA, CA
10	Mic, Tel	Cross	768	281	CI, LA
11	Mic, Cell	English	460	97	CI
12	Mic, Cell	Cross	632	114	CI
14	Tel, Cell	Cross	460	97	CI

This set of corpora were selected by the FBI for calibration research to represent a very wide range of different conditions, collection sources, environments, languages, and channels. See Table I for a detail of the conditions and the corpora from which they were sourced. We exclude conditions 1 and 4 for being subsets of conditions 2 and 5, respectively, and conditions 7 and 13 for having too few speakers.

The FBI data is split into calibration and evaluation sets, choosing one third of the speakers in each condition for calibration and the rest for evaluation. This calibration data is used for some experiments on the FBI data, while other experiments use the LVD data described next. The mean speech duration in the FBI samples is 85 seconds, with only 2% of the samples having less than 30 seconds of speech. To enable duration-dependent calibration, each calibration sample is cut to obtain new samples with approximately 5, 10, 20, 40, 80 and 160 seconds of speech, up to the maximum duration available in the sample. Test samples are not cut.

LVD: For calibration of the FBI data we also create a multi-condition set sourced from datasets not included in testing, in order to generate a challenging scenario for the calibration algorithms. Data was obtained from telephone and microphone sources of the NIST 2004-2008 SRE corpora, and clean telephone data from the non-English DARPA RATS SID task [21]. Speakers were not overlapping with system training data as we have empirically found that avoiding overlap between the calibration data and data used to train the speaker recognition system improves calibration results. This data consists of 1259 waveforms from 144 speakers, roughly balanced across the SRE and RATS data sources. Distribution of conditions in the waveforms included 4% microphone data and 96% telephone data, a 55% female 45% male split, and representation of over 27 languages with the top 5 languages (Farsi, Levantine Arabic, English, Pashto, and Urdu) each having a share of 8-12%. These samples are cut to create new samples that contain 5, 10, 20, 40 and 60 or 80 seconds of speech, up to the maximum duration available for each sample, to enable duration-dependent calibration.

C. Performance Metrics

Throughout this report, we measure performance using three metrics: the cost of likelihood ratio (Cllr), the Cllr loss (Closs), and the percent of rejected trials (%Rej). The Cllr [26] measures the quality of the scores as LLRs using a logarithmic cost function. This metric is affected both by the discrimination and calibration performance of the system. In this work, we are not aiming to improve discrimination, though we may do so as a side-effect since TBC is not a global transformation and might end up better aligning the scores from different trials which may, in turn, improve discrimination. Yet, our main goal is to improve calibration of the scores. That is, we want to make sure that the system's scores are interpretable as proper LLRs, regardless of the system's discrimination performance. To this end, we use the Closs as our main metric.

We define the Closs as the relative difference between the Cllr of a certain calibration procedure (TBC, or global calibration), Cllr_T, and the Cllr obtained when training the calibration model with calibration data well-matched (same condition from the same dataset) to the test set, Cllr_M:

$$\text{Closs} = (\text{Cllr}_T - \text{Cllr}_M)/\text{Cllr}_M \tag{3}$$

Both Cllr values involved in the computation of the Closs are calculated using only the trials that are calibrated (i.e., not rejected) by the system under study. Figure 3 depicts the process used to obtain the Closs values.

Note that we could have defined Closs in a more traditional way as the relative difference (or absolute difference) between the actual Cllr on a certain test set and the minimum Cllr obtained using some optimal calibration procedure like the pool adjacent violators algorithm (PAV) [27] or perhaps by linear logistic regression on the test data itself. Yet, we believe a minimum Cllr calculated in this way would result in an inadequate baseline since such Cllr would not be achievable in practice by any calibration procedure because it is optimized on the test data itself. For this reason, we believe that using matched held-out data for computing the minimum Cllr in the Closs computation is a better approach for the purpose of this paper. Further, this Closs has a nice characteristic: if a TBC algorithm happens to select for each trial exactly the matched data for that trial, then the Closs will be 0.0. Note that we could also, in fact, get negative Closs values. This would mean that the calibration method under evaluation is better than the matched calibration model for that test set. As we will see, this happens in our experiments for a few conditions.

Our proposed TBC algorithm has the option of refusing to calibrate a trial if no sufficient matched data can be selected for calibration. Hence, another performance metric of interest is the percent of trials that were calibrated by the system. As mentioned above in all cases the Closs is computed only over the trials that were calibrated by the system under study.

Finally, for some of the experiments, we want to summarize metrics over all the test sets within a certain group. For this purpose, we use a weighted average Closs where the weight is given by the fraction of trials that are calibrated by the system for each test set. That is

$$\text{WEIGHTED\_AVE}(\text{Closs}) = \frac{\sum_i w_i \text{Closs}_i}{\sum_i w_i} \tag{4}$$

where i runs over the test sets under consideration and $w_i = 1 - \%\text{Rej}_i/100$ is the fraction of trials calibrated by the system for test set i. This coincides with the standard average when the system calibrates all trials as is the case for global calibration and TBC when no similarity threshold is used.

For the final set of results, we also show the equal error rate (EER) and the minimum Cllr on each condition. The EER is defined as the false alarm rate at the operating point in which this rate is equal to the miss rate. The minimum Cllr is obtained using the PAV algorithm which gives the smallest Cllr value that can be obtained by mapping the test scores with a monotonic non-parametric transformation.

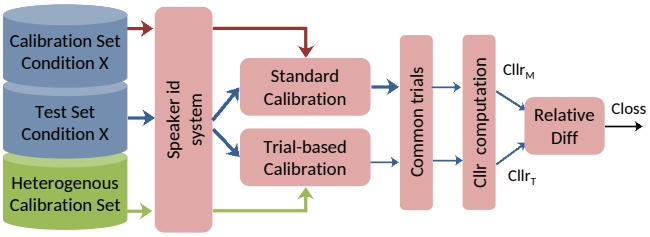

Fig. 3. Computation of the Cllr loss (Closs) performance metric. The speaker recognition scores for three sets are used: a test set for a certain condition X, a matched calibration set for that condition and a large heterogeneous set. Standard linear logistic regression calibration is performed using the matched calibration set. An alternative calibration method like TBC is also applied to the test trial's scores using the heterogeneous calibration set. Different number of trials may come out of each calibration method (indicated by the different arrow widths). Finally, the scores calibrated by both systems are used to compute their respective Cllr. The Closs is given by the relative difference between these two Cllrs.

D. MFCC i-vector/PLDA Speaker Recognition System

All experiments in this work, except otherwise indicated, use a standard mel-frequency cepstral coefficient (MFCC)-based i-vector/PLDA speaker recognition system. The MFCC acoustic features are based on 20-dimensional MFCCs including C0, spanning the frequency range of 200-3300 Hz using 24 filter banks, a window of 25ms, and a step size of 10ms. These features are used as input for speech activity detection (SAD) and for the speaker recognition system.

For SAD we used a deep neural network (DNN)-based model trained on telephone and microphone data from a subset of the Mixer data [28] used in SRE 2008, consisting of more than 19,000 samples. The MFCC features are mean and variance normalized using a sliding window of two seconds, and are concatenated over a window of 31 frames. The resulting 620-dimensional feature vector forms the input to a DNN that consists of two hidden layers of sizes 500 and 100. The output layer of the DNN consists of two nodes trained to predict the posteriors for the speech and non-speech classes. These posteriors are converted into likelihood ratios using Bayes rule (assuming a prior of 0.5), and a threshold of 0.5 is applied to obtain the final speech regions.

We trained a gender-independent UBM model with 2048 Gaussians followed by a 400-dimensional i-vector extractor [29]. For i-vector extraction, the MFCC features are contextualized with deltas and double deltas prior to utterance-level mean and variance normalization over the speech frames as determined by the SAD system. A PLDA model is used to generate the scores for the i-vectors in each test set. I-vectors are length-normalized and reduced to a dimension of 200 using LDA before PLDA. These scores are the input to the calibration stage. All system components are trained using only the SRE 2008 subset of the Mixer data, excluding speakers used in the LVD set described in IV-B. There is no overlap between this system's training speakers and the calibration or test speakers.

E. Hybrid I-vector/PLDA Speaker Recognition System

For the final evaluation of the proposed approaches, we also use a DNN-based speaker recognition system, which we call the hybrid system. The hybrid-alignment framework [30] provides competitive speaker recognition performance across mixed conditions. This system leverages a DNN trained to predict 3450 tied tri-phone states to extract 80-dimensional bottleneck features. These phonetically rich bottleneck features are used to train a UBM of 2048 Gaussians, which is later used to generate frame occupancies or alignments for input audio. The alignments are used to generate zero-order statistics and combined with 20-dimensional MFCCs appended with deltas and double-deltas to calculate first-order statistics. The statistics are used in the training of an i-vector subspace of 400 dimensions, from which i-vectors are extracted for our PLDA experiments. Training data for the DNN included Fisher, Switchboard and Callhome data (more details on the DNN can be found in [31]), the UBM and i-vector extractor for this system are trained with the non-degraded subset of the PRISM training dataset [32], while LDA and PLDA are trained using the full PRISM dataset including over 73,000 files from 3300 unique speakers. This system is only used to evaluate on the FBI data, with which there is no speaker overlap. Care was also taken to avoid overlap between the training speakers for this system and the speakers included in the LVD set.

F. TBC Systems

For the UAC similarity metric, three models are used which predict gender, style/channel, and language. These models are trained using the same Mixer data used for all system components in the MFCC i-vector/PLDA speaker recognition system (Section IV-D). The data includes 19,298 files from 965 speakers. The input to each of the models are the same i-vectors used for the speaker recognition system and the output classes are given by (1) female vs male for the gender model, (2) interview microphone speech vs telephone speech over microphone channel vs telephone speech over telephone channel for the style/channel model, and (3) English vs non-English speech for the language model.

For the CPLDA similarity metric, the models use the same i-vector extractor as the speaker recognition system to be calibrated. We compute three models using different subsets of the PRISM collection [32]. The PRISM collection includes telephone and microphone data from Switchboard, Fisher and Mixer collections, and simulated noisy and reverberated data. Based on our work in [33], we also added transcoded data using a large variety of codecs including AAC, AMR-NB, CODEC2, MP3, OPUS, and SPEEX, among others, each of them at different sampling rates for a total of 32 different codecs.

For the development experiments, we compare two CPLDA models, one trained with the small Mixer list used for training the UAC models and another one trained with a subset of the PRISM set after discarding Switchboard data and all samples used for calibration of the development set (Section IV-A). This latter set includes 125,077 files from 15,248 speakers. The condition of each file, used as label when training the CPLDA model, is determined by combining the gender, microphone type, degradation type, language spoken, vocal effort and collection name. The smaller training list contains 83 distinct conditions, while the larger list contains 428 conditions.

For the evaluation experiments we use a subset of the PRISM set after discarding all samples in the LVD calibration set (Section IV-B) as well as any files used to train the i-vector extractors for both speaker recognition systems. We found that discarding the data used to train the i-vector extractors gave a modest but consistent improvement in TBC performance. This training list contains 72,975 files from 1400 speakers and 396 distinct conditions.

All calibration models are trained using linear logistic regression. TBC calibration models are regularized toward the global calibration parameters. In all cases, we use an effective prior of 0.01.

V. Development Results

We initially tune the system parameters using the weighted average Closs as well as the worst Closs over all development sets. If a system setup is similar or even slightly better to another setup in terms of average but significantly worse in terms of the worst Closs, then we choose the one where the worst Closs is lower, since we wish to design a robust system that can handle a wide variety of conditions without failing. After finding the system's parameters in this way, we show results for a few chosen configurations compared to the baseline on all individual development sets. All results in this section use the MFCC speaker recognition system described in Section IV-D.

Except for some of the experiments in Section V-C, the calibration data for all other experiments in this section is obtained by merging the calibration data from each of the development sets. In this way, matched calibration data is available within the calibration set for every set. A good similarity metric should be able to retrieve that matched data for each trial, resulting in a Closs close to 0.

A. Similarity Metric Comparison

Figure 4 shows the average and worst-case Closs for global calibration and four TBC systems using the UAC, IV metrics,

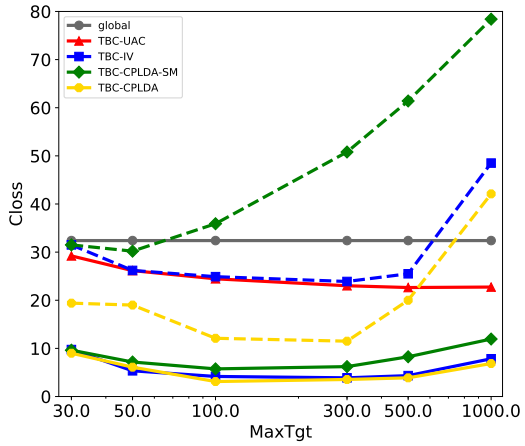

Fig. 4. Comparison of similarity metrics on the development set as a function of the MaxTgt parameter. The global calibration system is independent of MaxTgt and, hence, is a flat line in this plot. Solid lines correspond to the average Closs over all the development sets, while dashed lines correspond to the worst Closs (the color of the dashed lines indicate the corresponding system). Two dashed lines are missing for being too far from the range of the other lines. The worst Closs for the global calibration performance is 151.6, while the worst Closs for TBC-UAC is between 127.6 and 131.7 depending on MaxTgt.

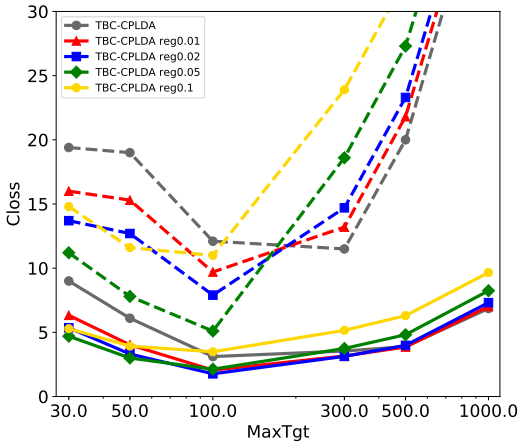

Fig. 5. Comparison of performance on the development set for the TBC-CPLDA method using different regularization weights and no regularization (equivalent to a regularization weight of 0). Solid lines correspond to the average Closs over the test sets and dashed lines correspond to worst-case Closs.

and the CPLDA metric for two different training sets. CPLDA-SM refers to a model trained with the small training set also used to train the UAC models and CPLDA refers to the model trained with the full training set (both sets are described in Section IV-F). Results are shown as a function of the MaxTgt parameter. The TBC systems do not use a similarity threshold for these results. They select for each trial the best-matching data based on the corresponding similarity metric that results in MaxTgt target samples being selected.

We can see that all TBC methods outperform the global calibration performance. The UAC metric, though, fails on one of the test sets (FVC-int), leading to a worst-case Closs of 127%, close to the worst-case Closs of the global calibration on that same set, which is 151%. The other three metrics give similar results to each other for all values of MaxTgt in terms of average Closs, with the CPLDA metric giving a significantly better worst-case Closs than the other two metrics. A detailed comparison of the three metrics over all development conditions can be found in Section V-D.

Results show that large MaxTgt values result in a significant degradation in the worst Closs. This is due to the fact that the calibration dataset does not contain enough matched-condition target samples. Hence, increasing the required number of target samples implies that a larger number of mismatched samples are being selected, which in turn degrades calibration performance. The optimum value of MaxTgt for our development set is somewhere between 100 and 300. If the calibration set was smaller and only provided a small number of matched target samples, the optimum would probably be smaller.

B. Effect of the Regularization Weight

Figure 5 shows a comparison of results when using no regularization (as in Figure 4) and when using different values for the regularization weight on the TBC-CPLDA method. We can see that regularization offers an improvement in performance, specially for the worst-case Closs when MaxTgt is relatively small. Values between 0.02 and 0.05 lead to the best performance for the smaller values of MaxTgt. Since we aim to design a system that will work well for smaller calibration databases for which MaxTgt might have to be set to a relatively small value, we will set the regularization weight for the rest of the experiments to 0.05, which is optimal for MaxTgt in the range of 30 to 100, specially in terms of worst-case Closs.

C. Effect of the Similarity Threshold

For the experiments in this section, we set the MaxTgt to 100 and the regularization weight to 0.05 and compare the effect of varying the similarity threshold SimThr both on the Closs and the %Rej trials for the TBC-CPLDA method.

Figure 6 shows the Closs and %Rej as a function of SimThr and the MinTgt parameters when using all available calibration data for all test sets. We can see that neither parameter has a large effect on the Closs, except for MinTgt=10 which degrades the worst-case Closs by a factor of two for one specific case of SimThr. This means that training a calibration model with this number of target trials is not sufficiently robust, even when using regularization. Yet, values of MinTgt above 20 already give stable performance.

The fact that the SimThr does not significantly affect Closs is expected since the threshold is only a required minimum. When more than MaxTgt calibration target trials have a similarity to the trial's side above the SimThr value then the SimThr parameter has no effect on the selected trials. As we increase SimThr, the selected trials get downselected to only those with similarity above the threshold. Yet, this does not have a big effect on performance either since we are only selecting from an already well-matched set of trials.

On the other hand, the SimThr and MinTgt parameters have a big effect on the percentage of trials that can be calibrated

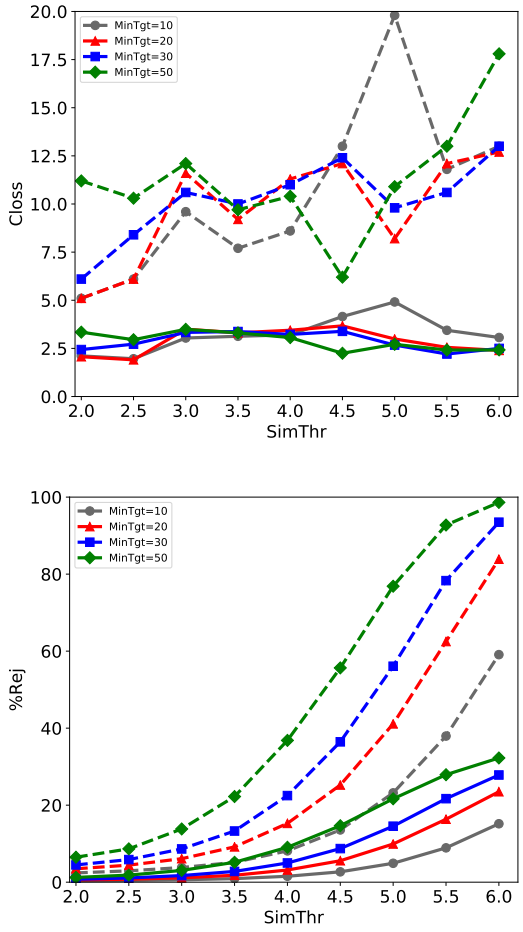

Fig. 6. Closs and %Rej as a function of the similarity threshold (SimThr) and the minimum number of target samples found (MinTgt) for the TBC-CPLDA method when using all available calibration data for all the test sets. Solid lines correspond to average over all the test sets, while dashed lines correspond to the largest value.

by the system. Since for these plots we are selecting samples from the full calibration set, matched data should be available for most test trials. Hence, in this case, we would like to see a small percentage of rejected trials. Indeed, this is what we see for the lower values of SimThr and MinTgt. Once those parameters start to increase, the percentage of rejected trials also increases to undesirable levels.

The selection procedure with a rejection option was designed to deal with cases in which the calibration does not contain matched data for certain trials. To test how the algorithm is working for such cases, we create a new calibration set for each test condition by collecting all calibration sets that are significantly mismatched to the test set condition. A calibration set is considered as significantly mismatched to a certain test set when a calibration model trained on that set results in a Closs larger than 50% on the corresponding test set. Hence, a different calibration set is used for each test where no well-matched data should be available for selection. Figure 7 shows the %Rej for these experiments. The Closs for the global calibration method is 85% (compared to 32% when matched calibration data is available). In this

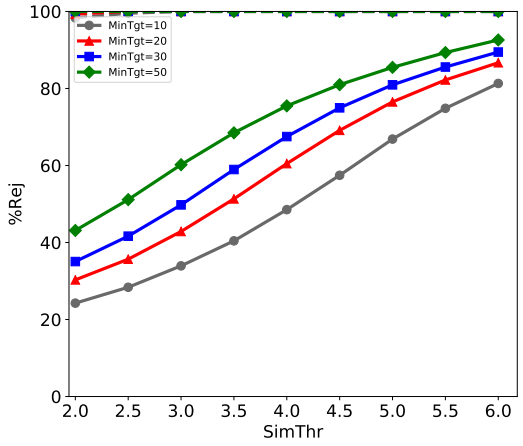

Fig. 7. %Rej as a function of the similarity threshold (SimThr) and the minimum number of target samples found (MinTgt) for the TBC-CPLDA method when using mismatched calibration data for each test set. Solid lines correspond to average %Rej over all test sets, while dashed lines correspond to the largest %Rej. Note that, in this case, all dashed lines overlap each other at the top of the plot (%Rej≈100).

case, we do not show the Closs because too few trials are calibrated in most of the test sets, which results in a noisy estimation of the Closs. Our goal in this scenario should be to reject most trials since none (or very few) of them should have matched calibration data available. As we can see from the figure, a large percentage of trials is indeed rejected on average for the larger values of SimThr and MinTgt.

The optimal system configuration is one that results in a small percentage of trials being rejected for the matched calibration case and a large percentage rejected for the mismatched calibration case. One such configuration would be, for example, SimThr=5.5 and MinTgt=20, which results in 16% of trials rejected for the matched case and 82% of trials rejected for the mismatched case.

D. Final Comparison on Development Data

Figure 8 shows a final comparison of results over all the individual test sets for the global calibration model and five selected TBC configurations. The Cllr values when using matched calibration models are indicated below the names in the bottom plot. These are the Cllr_M values used to compute the Closs (Equation (3)) values in both plots in that figure.

The top plot shows the results when the full calibration set is used for all test sets. We can see that (with the exception of SWPH2, for which the global calibration model is better than the matched model) performance on all test sets greatly improves with the proposed methods. Further, we can see that the two newly proposed similarity metrics (IV and CPLDA) significantly outperform the original UAC metric. Comparing the IV and the CPLDA metrics, we can see, as was also observed in Figure 4, that the IV metric has a significantly worse worst case compared to CPLDA with a Closs for SWCELLP1 of 19% compared to 4% for CPLDA. On average, the TBC method using either of the two new metrics reduces the Closs from 32% using the global calibration model trained on all data to 2% or less, making the calibrated performance

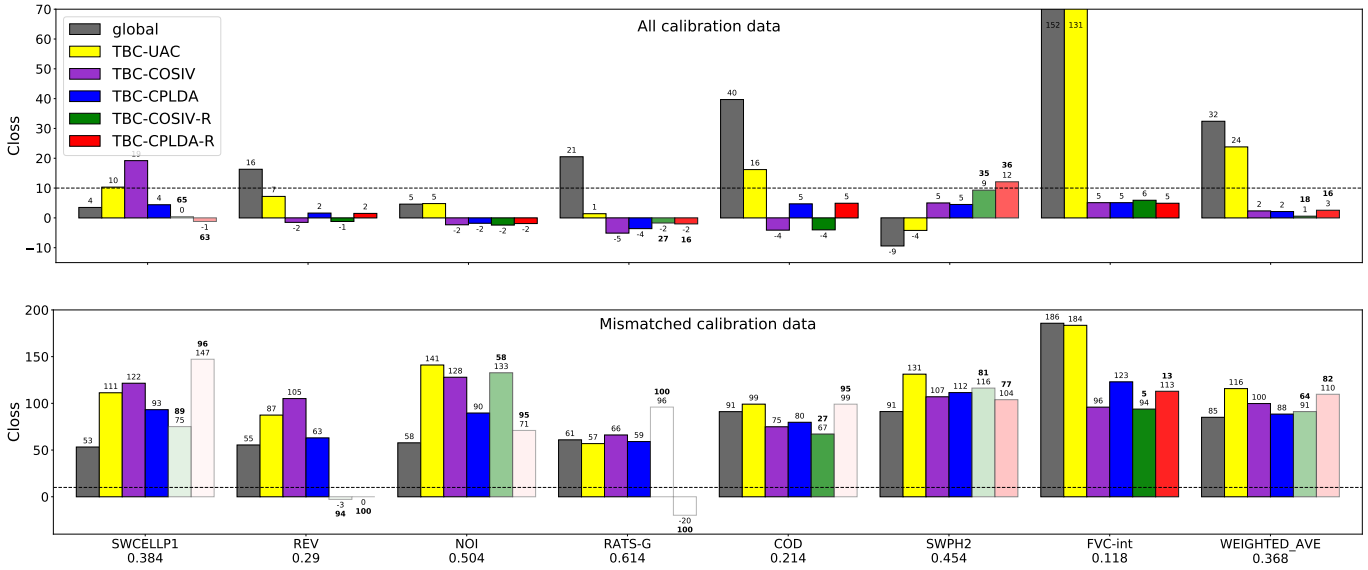

Fig. 8. Closs for each development test set using all available calibration data in the development set (top) and using the mismatched calibration set for each test set (bottom) using six calibration methods: global calibration, and five TBC methods. All TBC methods use MaxTgt of 100 and a regularization weight of 0.05. The TBC-UAC, TBC-IV and TBC-CPLDA methods do not use a similarity threshold, while TBC-IV-R uses a threshold of 0.29 and TBC-CPLDA-R uses a threshold of 5.5. For the two TBC methods with a reject threshold, a MinTgt of 20 is used. The numbers on top of the bars correspond to the height of the bar and the percent of rejected trials when this number is not 0 (in bold). The transparency of the bars is given by the percent of rejected trials. The numbers in the x-axis under the dataset names for the bottom plot are the Cllr values for matched calibration models. Individual sets are sorted based on their global calibration performance when using mismatched calibration data. The dashed horizontal line corresponds to a Closs of 10% and it is included for reference.

almost equal to that of the matched calibration model. In this case, where matched calibration data is available, introducing the rejection option (TBC-IV-R and TBC-CPLDA-R methods) does make a significant difference in the results.

The bottom plot in Figure 8 shows the same plot as in the top of the figure but changes the calibration set for each test set by discarding all matched calibration sets as in Figure 7. In this case, we can see that the TBC method without a rejection option fails to improve performance over the global calibration method in terms of weighted average Closs. This is because we are forcing the system to calibrate all trials even though no matched data is available to train the models. On the other hand, the two TBC methods with a rejection option successfully reject most trials in most conditions (the percentage of rejected trials can be found in bold at the top of each bar), which is the expected behavior given that no matched calibration data should be available for most trials in this scenario. Further, for the sets for which most trials are not rejected (FVC-int), the Closs is much lower on average than the Closs for global calibration. Yet, we consider this case to be a failure of the similarity metric. We would want any set of trials with a Closs as high as that obtained on FVC-int to be rejected. The fact that our CPLDA metric is not able to reject 87% of the trials resulting in a Closs over 100% (for the TBC-CPLDA-R configuration) means that the metric still requires some improvement.

Note that the Closs obtained with the TBC-CPLDA methods in the bottom plot of Figure 8 should be disregarded, since they are computed using very few trials (except for the FVC-int set).

The similarity thresholds for the TBC-IV and TBC-CPLDA methods with a reject option are set to 0.29 and 5.5, respectively. These values were chosen to lead to a similar percentage of rejected trials when matched calibration is available (18% vs. 16%). We can see that for these thresholds, the TBC-CPLDA method can reject a larger percent of trials when no matched calibration data is available (82% vs. 64%). This indicates that the CPLDA metric is better for rejecting trials with no matched calibration data.

VI. Evaluation Results

In this section, we show results on the FBI dataset for the TBC-CPLDA methods with and without a rejection option. Note that for this dataset the Closs is not strictly the calibration loss. The conditions are not perfectly homogeneous to ensure that the matched calibration performance is indeed the best we can do for each condition. Yet, we can still use this "matched" calibration performance as a target performance in calculating the Closs, keeping in mind that we might be able to improve it, getting negative values of Closs.

The FBI data is quite variable in terms of duration. For this reason we use the procedure described in Section III-D where each calibration sample is first chunked to obtain several shorter versions of each calibration sample. During testing, the calibration samples with a duration that matches those of the trial being calibrated are selected for global calibration or as candidates for further selection in case of TBC. The durations of two samples are considered to match if they fall within the same bin, with bins determined by the following thresholds: 7.5, 15.0, 30.0 and 50.0 seconds. In some cases, for the matched calibration models, some bins do not have enough calibration trials to result in robust models. In these

cases, the bin with too few samples is merged with the bin to its left until enough samples are available for calibration.

The top plot in Figure 9 shows the results on FBI conditions when using FBI data for calibration. Results show that our proposed methods succeed in greatly reducing the Closs in the two cases where the global calibration model has a large Closs, conditions 10 and 2, reducing the Closs to close to 0 for condition 10 and below 50% (a 16-fold reduction in Closs) for condition 2. On average, the Closs goes from 92% for global calibration to less than 10% for the two TBC methods. For some conditions, the Closs degrades with respect to that of the global calibration model (conditions 11, 14 and 5). The degradation is small compared to the gain obtained on other conditions. Yet, this again indicates that the similarity metric can still be improved.

The bottom plot in Figure 9 shows the results on FBI conditions when using LVD data for calibration. In this case, global calibration is significantly worse for all conditions, except condition 10. The benefit of using TBC is very clear in this case, significantly reducing Closs in most cases. The only case where the TBC methods do not behave as expected is condition 2, where the Closs is reduced, though still remaining extremely large and only rejecting 17% of the trials. Again, this points to a weakness in the similarity metric that is finding calibration samples that are not a good representation of the test samples, resulting in bad calibration performance.

Interestingly, condition 2 and the FVC-int development condition, which are the two conditions where TBC with relatively mismatched data is failing to reduce Closs to usable levels while also failing to reject a large percent of the trials, share something in common: they have the lowest Cllr in their corresponding group. The corresponding EER for those conditions is 2.4% for FVC-int and 0.7% for FBI's condition 2, when using the MFCC i-vector system. We believe that the fact that our CPLDA metric does not behave as expected on these conditions may be due to a shortcoming in the training data for the CPLDA model, which is dominated by degraded data, perhaps neglecting to learn from the cleaner conditions from close-talking microphones. This is something we plan to explore in the near future.

Finally, we repeat the above experiments using a DNN-based system, described in Section IV-E. Figure 10 shows these results. The DNN-based system is significantly better than the one used for all other experiments in this paper, as can be seen by comparing the Cllr numbers located under the condition names in Figures 9 and 10. Yet, we can see that TBC gives similar gains on both systems (with the only exception of condition 9 where TBC is not able to reduce the Closs significantly) indicating that the approach and chosen configuration generalize reasonably well to this new system.

Note that the weighted average results in the bottom plots of Figures 9 and 10 are mostly determined by the results for condition 2 divided by the number of conditions (since all other conditions have much smaller values) and, hence, should not be taken as an indication of the systems' performance across conditions.

For completeness, Figure 11 shows the absolute values of Cllr and minimum Cllr (minCllr) corresponding to the bottom plot in Figure 10. We do not include the system with a rejection option in this figure because those results would not be comparable to the ones from other systems, given that they are computed on a subset of the trials for each condition. We show the global and the TBC-CPLDA results, as well as the matched calibration results, which are used to compute the Closs in Figure 10. The figure shows that the minCllr is not significantly affected by the calibration approach. On the other hand, the Cllr is greatly improved when using TBC compared to global calibration, reducing the gap between global and matched calibration by approximately 50%. As in the case of minCllr, the EER is basically unaffected by the calibration approach. For this reason, and to avoid cluttering the figure, we only show the EER for the matched calibration approach (under the x-axis labels).

The TBC results in Figures 10 and 11 are obtained using a CPLDA-based similarity metric that uses i-vectors obtained with the hybrid system. Nevertheless, we note that using a CPLDA model based on i-vectors obtained with the MFCC system leads to very similar results (results not shown), indicating that matching the i-vectors used for metric computation to those used in the speaker recognition system does not appear to be essential. This may also mean that the CPLDA metric could be used even with speaker recognition systems not based on i-vectors, though this is a hypothesis that will have to be tested empirically in the future.

VII. Practical Considerations

In this section, we provide general information on how to construct a reasonable CPLDA training list and calibration set for TBC, keeping in mind that the there is still a lot to be discovered in terms of how different parts of the TBC mechanism interact and the following information should be taken as suggestions rather than a hardened and definite recipe.

In general, there are two different pools of data in the system; data for training the system models (UBM, i-vector extractor, PLDA, CPLDA, etc.) and data for calibration. It is recommended that no overlap exist between these two sets. It was heuristically found that including partial overlap between these datasets resulted in a degradation in calibration performance, both for global calibration and for TBC. Hence, the calibration speakers should be, ideally, completely unseen during training of any of the other system components.

The conditions of the data used for calibration should, ideally, cover the expected range of conditions in normal use of the system. If possible, one should aim to include a few dozen speakers for each condition, as well as several conditions per speaker to allow for cross-condition calibration trials. If a relatively large amount of matched data is available, one might wish to use some of that data for system training, holding out only part of it for calibration. This might improve the discrimination power of the system on the test data without significantly hurting calibration performance. In our experiments, to simplify the analysis, we keep the system training data fixed and only change the calibration data. We expect that the effect of adding a relatively small number of files (in the order of hundreds or few thousands) would only

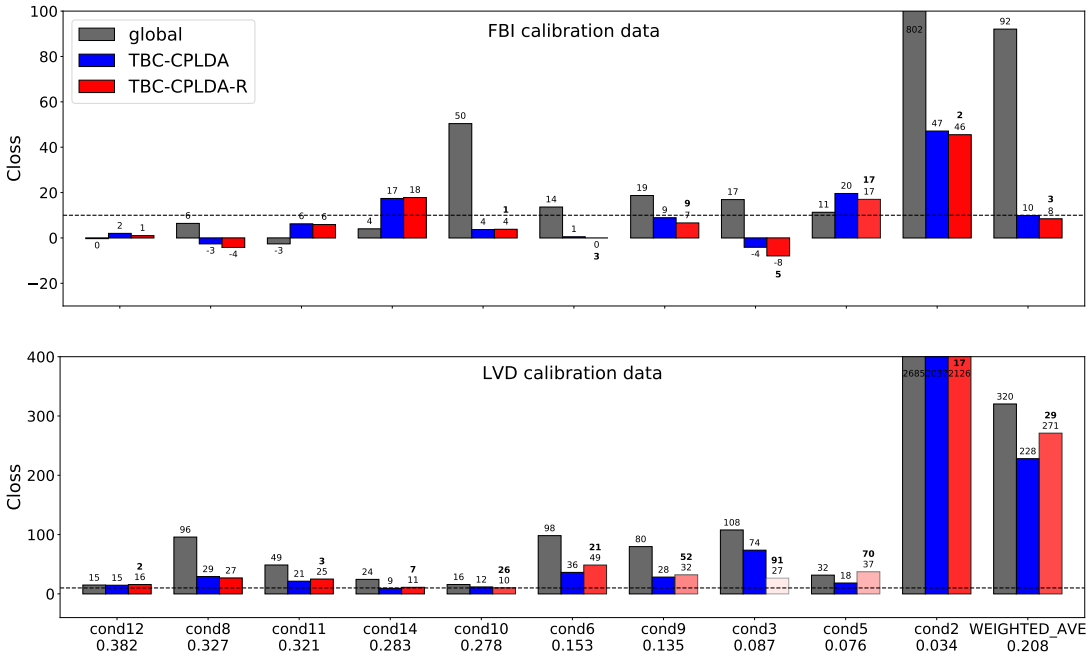

Fig. 9. Same as Figure 8 but on the FBI test sets using FBI calibration data (top) and LVD calibration data (bottom). Conditions are sorted by decreasing Cllr (values under the x-axis labels in the bottom plot).

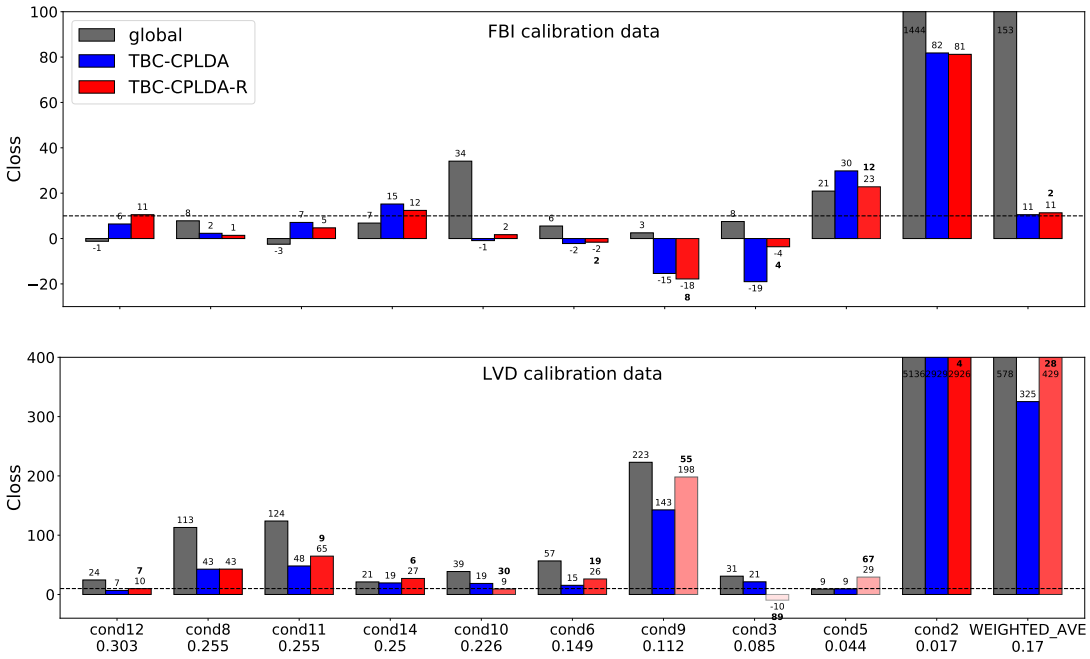

Fig. 10. Same as Figure 9 but using the hybrid speaker recognition system. Conditions are ordered in the same way as in Figure 9.

have a negligible effect given the large amount of data already used to train the system. Further, it is not our intention to adapt fully to a certain test condition (e.g., by fully retraining the PLDA model on data matched to the test data in our experiments), since we aim to design a general purpose system that performs well across a wide range of conditions.

The data for training CPLDA is also an important factor in the performance of the TBC approach. We have found that this data should not overlap with the calibration data or, to a lesser extent, with the data used to train the i-vector extractor. Ideally, the data used to train this model should include a large variety of conditions and speakers to allow the model to learn how these conditions affect the i-vectors independently of the speaker present in the signal. Nevertheless, if test data is restricted to a certain known set of conditions, then it might be preferable to restrict the CPLDA training data to only these known conditions.

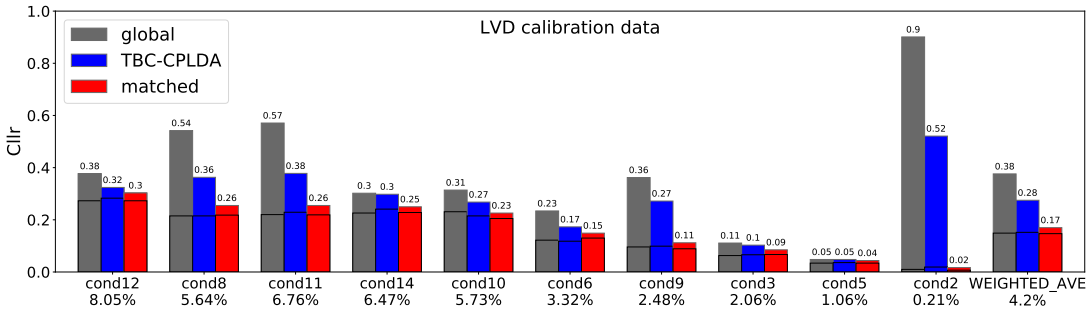

Fig. 11. Cllr values for three calibration approaches when using the hybrid speaker recognition system and the LVD calibration set. The transparent blocks with black edges inside each bar indicate the minCllr. The EER percentage for the matched calibration approach for each condition is indicated under the x-axis labels.

VIII. Conclusion

We presented a method for calibrating speaker recognition scores when the test conditions are unknown and potentially heterogeneous. The approach consists of training a separate calibration model for each test trial using a subset of the available training data that is similar to the trial's conditions. We compare three ways of measuring this similarity, showing that a metric based on a condition PLDA model significantly outperforms the other two metrics based on cosine similarity between i-vectors and on Gaussian condition classifiers. Further, we propose using regularization to train each individual calibration model. This enables us to select fewer samples for training the models, which, in turn, allows for a better match between the selected data and the trial, which is particularly useful when limited matched data is available for each condition.

Our goal is to design a calibration approach that guarantees that no trial is ever calibrated using mismatched calibration data, which would result in a likelihood ratio that does not reflect the score distributions for the conditions in the trial. To prevent such cases, we introduce an option to reject a trial when the system does not find enough similar data to train a calibration model for that trial.

We measure calibration performance using calibration loss (Closs) defined as the relative difference between the Cllr for a matched calibration model and the method under evaluation. We compare the Closs of our proposed methods with the one obtained using a single calibration model trained with all the available calibration data.

When matched data for all test trials is available for selection within the calibration set, the weighted average Closs over conditions (with weights given by the fraction of trials being calibrated) is reduced from 32% to less than 2% in our development set composed of several conditions including noise, reverberation and transcoded data. For the held-out FBI data, which includes same- and cross-channel and same- and cross-language trials from several different collections, the average weighted Closs is reduced from 92% to 11% or less when held-out data pooled from all FBI conditions is used for calibration.

Finally, when the calibration set is mismatched to the test data, our proposed method succeeds in rejecting over 82% of the development trials. For the FBI data, we use a partially mismatched calibration set composed of a large variety of data from different collections excluding the ones used in the FBI set. When this set is used for calibration, we reject 29% of the trials, while significantly reducing Closs on the trials that are calibrated.

The results above were obtained using a standard MFCC-based i-vector/PLDA speaker recognition system. Similar results were obtained without parameter retuning using a DNN-based system that is significantly better than the standard system in terms of discrimination performance.

For most development and evaluation conditions in our datasets the proposed algorithms behave as desired, significantly reducing Closs when matched data is available or rejecting most trials when that is not the case. Yet, for a small subset of the conditions the algorithms fail, calibrating a significant fraction of the trials with Closs values over 50%. In particular, this happens when data conditions are extremely clean and matched between test and enrollment sides. This indicates that the similarity metric is not yet considering all possible sources of mismatch that cause mis-calibration. In future work, we will focus on different ways of defining the training conditions for the condition PLDA model, which we believe might be the key for getting a more robust similarity metric.

The proposed methods are a first attempt at designing a calibration system that guarantees that no trial is ever calibrated poorly by training an optimal calibration model for each trial when possible, and rejecting a trial when not possible. The ultimate goal of this work is to prevent speaker recognition systems from outputting poorly calibrated scores. We believe we have made significant progress toward this goal, though more work is needed to be able to fully rely on these systems for important applications like forensic speaker recognition.

Acknowledgment

The research by authors at SRI International was funded through a development contract with Sandia National Laboratories (#DE-AC04-94AL85000). The views herein are those of the authors and do not necessarily represent the views of the funding agencies.

REFERENCES

[1] L. Ferrer, K. Sönmez, and S. Kajarekar, "Class-dependent score combination for speaker recognition," in *Proc. Interspeech*, Lisbon, Sep. 2005.

[2] Y. Solewicz and M. Koppel, "Considering speech quality in speaker verification fusion," in *Proc. Interspeech*, Lisbon, Sep. 2005.

[3] ——, "Using post-classifiers to enhance fusion of low- and high-level speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, Sep. 2007.

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

[5] N. Brümmer, "Focal bilinear toolkit," 2008. [Online]. Available: http://niko.brummer.googlepages.com/focalbilinear

[6] M. McLaren, A. Lawson, L. Ferrer, N. Scheffer, and Y. Lei, "Trial-based calibration for speaker recognition in unseen conditions," in *Proc. Odyssey-14*, Joensuu, Finland, Jun. 2014.

[7] G. Morrison and E. Enzinger, "Forensic speech science—review: 2010–2013," in *Proceedings of the 17th International Forensic Science Managers' Symposium*, Lyon, France, 2013, pp. 616–623.

[8] R. Schwartz, "When to punt on speaker comparison?" *The journal of the acoustical society of america*, Oct. 2011.

[9] N. Brümmer, A. Swart, and D. van Leeuwen, "A comparison of linear and non-linear calibrations for speaker recognition," in *Proc. Odyssey-14*, Joensuu, Finland, Jun. 2014.

[10] V. Hautamäki, K. A. Lee, T. Kinnunen, B. Ma, and H. Li, "Regularized logistic regression fusion for speaker verification," in *Proc. Interspeech*, Florence, Italy, Aug. 2011.

[11] D. Sturim and D. Reynolds, "Speaker adaptive cohort selection for T-norm in text-independent speaker verification," in *Proc. ICASSP*, Philadelphia, Mar. 2005.

[12] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in *Proc. Odyssey-10*, Brno, Czech Republic, Jun. 2010.

[13] D. Castan, M. McLaren, L. Ferrer, A. Lawson, and A. Lozano-Diez, "Improving robustness of speaker recognition to new conditions using unlabeled data," in *Proc. Interspeech*, Stockholm, August 2017.

[14] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, "A unified approach for audio characterization and its application to speaker recognition," in *Proc. Odyssey-12*, Singapore, Jun. 2012.

[15] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[16] D. Graff, K. Walker, and D. Miller, "Switchboard cellular part 1 audio LDC2001S13," 2001. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2001S13

[17] D. Graff, K. Walker, and A. Canavan, "Switchboard-2 phase II LDC99S79," 1999. [Online]. Available: https://catalog.ldc.upenn.edu/LDC99S79

[18] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow, "Forensic database of voice recordings of 500+ australian english speakers," 2015. [Online]. Available: http://databases.forensic-voice-comparison.net

[19] S. G. McGovern, "Fast convolution," 2004. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/5110-fast-convolution

[20] G. Hirsch, "Fant," 2005. [Online]. Available: http://dnt.kr.hs-niederrhein.de/download.html

[21] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. Odyssey-12*, Singapore, Jun. 2012.

[22] S. D. Beck, R. Schwartz, and H. Nakasone, "A bilingual multi-modal voice corpus for language and speaker recognition (LASR) services," in *Proc. Odyssey-04*, Toledo, Spain, May 2004.

[23] Y. Lei and J. Hansen, "Dialect classification via text-independent training and testing for arabic, spanish and chinese," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[24] M. Przybocki and A. Martin, "The 1999 NIST speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking," in *Proc. Eurospeech*, Budapest, Sep. 1999.

[25] K. W. Godin, S. O. Sadjadi, and J. H. Hansen, "Impact of noise reduction and spectrum estimation on noise robust speaker identification," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[26] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, apr 2006.

[27] ——, "The PAV algorithm optimizes binary proper scoring rules," 2013. [Online]. Available: https://sites.google.com/site/nikobrummer/pav_optimizes_rbpsr.pdf

[28] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.

[29] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[30] M. McLaren, D. Castan, L. Ferrer, and A. Lawson, "On the issue of calibration in DNN-Based speaker recognition systems," in *Proc. Interspeech*, San Francisco, September 2016.

[31] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP*, Florence, Italy, May 2014.

[32] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and N. Scheffer, "Promoting robustness for speaker modeling in the community: The PRISM evaluation set," in *Proceedings of SRE11 Analysis Workshop*, Atlanta, USA, Dec. 2011.

[33] M. McLaren, V. Abrash, M. Graciarena, Y. Lei, and J. Pesán, "Improving robustness to compressed speech in speaker recognition." in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 3698–3702.

Luciana Ferrer is a researcher at the Computer Science Institute, from the National Scientific and Technical Research Council (CONICET) and the University of Buenos Aires (UBA), Argentina. Prior to her current position, Luciana worked at the Speech Technology and Research Laboratory, SRI International, USA. Her current research interests include speaker and language identification, speech activity detection, and pronunciation scoring for second language learning. Luciana received the B.S. degree from the University of Buenos Aires, Argentina, in 2001, and her Ph.D. degree from Stanford University, USA, in 2009.

Mahesh Nandwana is a computer scientist in SRI International's Speech Technology and Research (STAR) Laboratory. His research interests include machine learning and signal processing for speaker and language identification, speech activity detection, and acoustic event detection. Prior to joining SRI, Nandwana worked as a research assistant at Center for Robust Speech Systems (CRSS), UT Dallas. Earlier, he was a project associate at speech research group, IIT Madras. He earned an M.S. degree in electrical engineering from UT Dallas and holds a B.Tech (Hons.) in electrical and electronics engineering from NIT Jamshedpur, India.

Mitchell McLaren is a senior computer scientist in SRI International's Speech Technology and Research (STAR) Laboratory. His research interests include speaker and language identification, as well as other biometrics such as face recognition. Prior to joining SRI in 2012, Mitchell was a postdoctoral researcher at the University of Nijmegen, The Netherlands, where he focused on speaker and face identification on the Bayesian Biometrics for Forensics (BBfor2) project, funded by Marie Curie Action. His Ph.D. in speaker identification is from the Queensland University of Technology (QUT), Brisbane, Australia.

Diego Castan received the Diploma, the M. Sc. and Ph. D. degrees in electrical engineering and information technology from the University of Zaragoza, Spain, in 2008, 2009, and 2014, respectively. In 2015, he joined the Speech Technology and Research (STAR) Laboratory in SRI International in Menlo Park, California. His research interests include speaker identification, language identification, and acoustic event detection.

Aaron Lawson received the Masters and Ph. D. degree in applied linguistics from Cornell University in 1998 and 2001, respectively. Prior to that he received a bachelor's degree in languages from Siena College in 1992 and a Master's degree in languages and Linguistics from Syracuse University in 1994. In 1999 he began working at TextWise LLC in Syracuse, NY on natural language processing and computation linguistics. In 2003 he joined the Audio Processing Group at the Air Force Research Laboratory in Rome, NY. Since 2011 he has been at the Speech Technology and Research (STAR) Laboratory in SRI International in Menlo Park, California. In 2016 he became assistant laboratory director. His research interests include speaker identification, forensics, language identification, and natural language processing.