Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Trial-Based Calibration for Speaker Recognition in Unseen Conditions

Mitchell McLaren, Aaron Lawson, Luciana Ferrer, Nicolas Scheffer, Yun Lei

Speech Technology and Research Laboratory
SRI International, California, USA
{mitch,aaron,lferrer,scheffer,yunlei}@speech.sri.com

Abstract

This work presents Trial-Based Calibration (TBC), a novel, automated calibration technique robust to both unseen and widely varying conditions. Motivated by the approach taken by forensic experts in speaker recognition, TBC delays estimating calibration parameters until trial-time when acoustic and behavioral conditions of both sides of the trial are known. An audio characterization system is used to select a small subset of candidate calibration audio samples that best match the conditions of the enrollment sample and a subset that resembles the test conditions. Calibration parameters learned from the target and impostor trials generated by pairing up these samples are then used to calibrate the score output from the speaker identification system. Evaluated on a diverse, pooled collection of 5 different databases with 14 distinct conditions, the proposed TBC outperforms traditional calibration methods and obtains calibration performance similar to having an ideally matched calibration set.

1. Introduction

Calibration is an important aspect in the usability of speaker identification (SID) systems. Calibration aims to transform scores to log-likelihood ratios (LLR) so that a single identification score can be meaningfully interpreted. For well-calibrated scores the optimal decision threshold for a certain cost function, given by a linear combination of the probability of miss and probability of false alarm, can be theoretically determined using Bayes decision theory.

In this work, we explore the problem of calibration when the trial conditions are variable. We wish to obtain a set of calibrated scores for which the optimal decision threshold computed for each pair of enrollment and test conditions is independent of these conditions. For example, we want the optimal thresholds to be the same for both telephone channel and microphone channel conditions, or for mixed channel conditions. Even the most accurate SID systems, if left un-calibrated or calibrated without regard for trial conditions, require a variety of condition-specific thresholds if optimal decisions are required within each condition.

Techniques developed to cope with varying conditions, such as calibration with metadata using Discriminative Probabilistic Linear Discriminant Analysis (DPLDA), incorporate information regarding conditions into the calibration parameters [1]. Information such as predicted gender, language and channel can be extracted via Universal Audio Characterization (UAC) [1] and represented as a low-dimensional vector of class posteriors. These dynamic calibration methods attempt to adjust the calibration shift and scale according to conditions observed in the trial. As shown in this work, calibration with metadata fails to adapt well to a score space not well covered during training.

The objective of this work was to investigate methods of calibrating scores such that a single threshold could be defined for multiple trial conditions to optimize the desired operating point. For this initial study, we focused on calibrating scores to minimize the calibration loss, computed as the difference between the actual cost and the minimum cost, across a set of 14 distinct conditions in a data set supplied by the Federal Bureau of Investigation (FBI). For this, we focused on an operating point where misses and false alarms are weighted equally.

2. Universal Audio Characterization

Universal Audio Characterization (UAC) is a technique that attempts to represent the conditions of a speech signal in terms of a small dimension vector of class posteriors. Previous work has shown improved calibration performance when taking into account these class posteriors [1, 2].

We utilize a Gaussian Backend (GB) to extract metadata in which each class of interest is modeled using a single Gaussian. I-vectors are used as input with corresponding output being a vector of likelihoods for each class (in the case of calibration with metadata, these likelihoods are log-normalized to obtain posteriors assuming equal priors for all classes). We use a GB for each category in which discrimination is useful (for instance, language or channel) after which we concatenate UAC vectors for each audio file.

3. Existing Calibration Methods

The process of calibration transforms scores to log-likelihood ratios (LLR). This in turn allows identification scores in isolation to be meaningfully interpreted. Common to all calibration techniques is the need to learn a set of calibration parameters (typically a scale and shift) from a development set. The development set contains both target and impostor scores representative of the conditions expected to be encountered during end use of the system.

Calibration methods considered in this work include simple logistic regression [3] and calibration with metadata which takes into consideration metadata extracted via UAC [1] and calibrates scores using DPLDA [4]. Alternate techniques include Neural Networks and SVMs [5], where their application in the context of heavily-degraded speech has been effective.

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.

3.1. Logistic Regression (Global) Calibration

Perhaps the most commonly used calibration technique in the field of SID is Logistic Regression [3]. A calibration model (shift and scale) is learned using linear logistic regression from a large pool of trials deemed to be representative of the end-use conditions. In many cases, a single model is trained using all development data. This approach optimizes calibration globally for all conditions in the development data. The resulting parameters, though, might not be optimal when performance is computed for each condition separately, especially for conditions not seen during training. To overcome this problem, a separate logistic regression model can be trained for each condition, optimizing calibration for each of them, but this requires prior knowledge of the conditions. Furthermore, conditions not seen during training of the calibration model cannot be handled in this case.

3.2. Calibration with metadata

Global calibration using metadata takes into account meta- or side-information extracted from each side of the trial to determine the best shift for the trial score. In practice, this is achieved using Universal Audio Characterization [1] and subsequent Discriminative Probabilistic Linear Discriminant Analysis (DPLDA) [4]. Specifically, the score s is converted to a log likelihood ratio using,

$$L = \beta + \alpha s + \boldsymbol{m_t}\Lambda\boldsymbol{m_e} + \boldsymbol{m_t}\Gamma\boldsymbol{m_t} + \boldsymbol{m_e}\Gamma\boldsymbol{m_e} + (\boldsymbol{m_e} + \boldsymbol{m_t})\boldsymbol{c}$$

where m_t and m_e are the UAC vectors for test and enrollment, respectively, and the parameters β, α, Λ, Γ and $\boldsymbol{c}$ are jointly are learned by minimizing a cross-entropy objective. Here, it can be seen that the shift required to convert the scores into LLRs depends on the UAC vectors.

Metadata-based calibration attempts to be robust to unseen conditions by assuming such conditions can be made of conditions that were observed during training (i.e., a segment spoken in a noisy and reverberant environment for which no training data was available may appear 80% noisy and 20% reverberant to a system that was trained with noisy and reverberant speech as separate classes). Limitations of side-information and the subsequent DPLDA model include unpredictability when salient acoustic and voice conditions not modeled a priori are present in the trials. Although able to cope with multiple trial conditions, calibration with metadata struggles to accommodate unknown trial conditions (as demonstrated in Section 7).

4. Trial-Based Calibration

In forensic speaker verification, conditions of each side of the trial are first determined by the forensic expert. The expert then attempts to locate calibration data that closely matches the trial conditions from which target and impostor trials can be made[1]. These trials provide probability distributions of target and impostor trials, which are then used to calibrate the score to a log likelihood ratio using Bayes theory.

To address the issue of mismatch between calibration data and evaluation data in an automated manner, we propose an approach motivated by forensic experts in SID. Termed Trial-Based Calibration (TBC), the approach is based on standard calibration training technique used in global calibration (logistic regression), however, the decision regarding the selection of data from which to learn calibration parameters is postponed until trial time. At this point, conditions of both sides of the trial can be extracted via UAC and used to select a small subset of highly relevant calibration data to produce trial-specific calibration parameters.

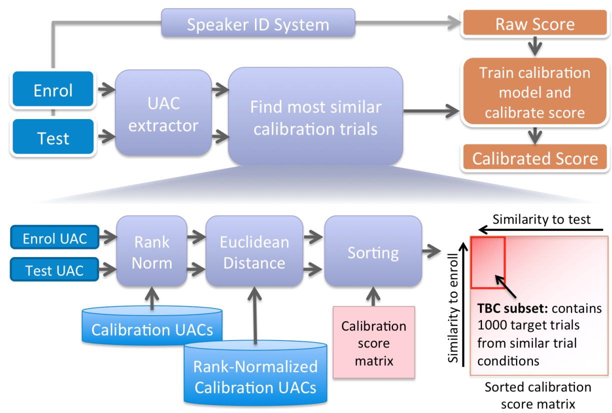


Figure 1: Flow diagram of Trial-based Calibration (TBC).

Selection of calibration training data first involves extracting UAC vectors, $\boldsymbol{m_t}$ and $\boldsymbol{m_e}$ from each side of the trial. These vectors are then rank-normalized into $\hat{\boldsymbol{m}}_{\boldsymbol{t}}$ and $\hat{\boldsymbol{m}}_{\boldsymbol{e}}$ using the UAC vectors $\boldsymbol{M_c}$ extracted from the entire candidate calibration set. Rank normalization replaces each value in a vector with their relative position when sorted against the same vector component in the normalization set. The rank values are then scaled to the range $[0, 1]$. The candidate calibration set UAC vectors $\boldsymbol{M_c}$ are also rank normalized against themselves to obtain $\hat{\boldsymbol{M}}_{\boldsymbol{c}}$. The Euclidean distance between a rank-normalized trial UAC vector and each vector in $\hat{\boldsymbol{M}}_{\boldsymbol{c}}$ is used as a measure of relevance between candidate calibration segments and the corresponding trial side. This sorting or ranking of calibration segments can be performed independently for both $\hat{\boldsymbol{m}}_{\boldsymbol{t}}$ and $\hat{\boldsymbol{m}}_{\boldsymbol{e}}$. Finally, for the purpose of learning the trial-dependent calibration model, target and impostor trial scores are selected from a pre-computed score matrix made up of the exhaustive set of trials from the candidate calibration set. Selection from this matrix is performed by sorting rows (or columns) according to the relevance measures of $\hat{\boldsymbol{M}}_{\boldsymbol{c}}$ with respect to $\hat{\boldsymbol{m}}_{\boldsymbol{t}}$ and the alternate axis sorted by relevance to $\hat{\boldsymbol{m}}_{\boldsymbol{e}}$. This results in the most representative scores in the upper-left corner of the matrix. In this work we select for calibration training the $N \times M$ matrix of scores from the upper-left corner such that at least 1000 target trials exist. In this selection, segments pertaining only to impostor trials are discarded to prevent easy cross-database trials from biasing calibration. Calibration parameters are learned through simple logistic regression using this small, highly relevant calibration dataset with which the trial score is transformed (calibrated). Figure 1 depicts the TBC process.

One of the major shortfalls of the metadata-based calibration approach is the reduction in system identification performance after calibration for conditions in which very few or no trials were observed during training. The proposed TBC approach is expected to provide additional robustness in these cases due to the selection of the closest set of segments to the trial sides and enforcing a minimum number of target trial scores from which to learn parameters. In the instance of trial

[1]In reality, finding datasets that closely represent the acoustic conditions of a forensic trial is difficult and often not possible. This differs from most speaker recognition research in which databases of largely homogeneous conditions are in ample supply.

Table 1: 14 Condition Evaluation Corpus sourced from Pan-Arabic (PA), CrossInt (CI) NoTel (NT), LASR and NIST99.

Cond.	Chan(s)	Lang(s)	# Spks	Source Corpora
01	Mic	Arabic	240	PA
02	Mic	Arabic	422	PA, LASR
03	Mic	Cross	179	LASR
04	Tel	English	225	NIST99
05	Tel	English	467	LASR, NT, NIST99
06	Tel	Cross	597	CI, LASR
07	Cell	English	62	NT
08	Cell	Cross	460	CI
09	Mic, Tel	English	645	CI, LASR
10	Mic, Tel	Cross	768	CI, LASR
11	Mic, Cell	English	460	CI, LASR
12	Mic, Cell	Cross	632	CI
13	Mic, Cell	English	51	NT
14	Mic, Cell	Cross	460	CI

conditions being completely absent from the candidate calibration segments, the selection of the TBC calibration set could be seen as an improvement on random (and therefore the standard linear regression model) as the most relevant segments are selected via ranking.

5. Data Sets

5.1. Evaluation Data

The 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 from 5856 unique segments. Table 1 details these conditions, the corpora from which they were sourced and the number of speakers involved. The total number of trials exceeds 2.8 million. Note that when reporting results, comparisons for conditions 07 and 13 will have limited statistical significance due to the limited speaker count. The source data in Table 1 contains both matched and cross language trials. The second language of all cross-language trials is English. First language (L1) trials for the LASR corpus [6], are in Korean, Spanish and Arabic. For the CrossInt corpus, first language trials are in several of the languages of India, depending on what the first language of the participant was, including Hindi, Gujarati, Bengali, Marathi, Tamil, Kannada and Telugu. The NoTel corpus [7] contains telephone recordings from naturally noisy locations in Indian accented English. Both the NoTel and CrossInt corpora were collected by Appen for speaker recognition research, while the LASR corpus was collected by BAE Systems. This set of corpora were selected for calibration research in order to represent a very wide range of different conditions, collection sources, environments, languages and channels.

5.2. Calibration Data

Matched Data: A small collection of 1503 segments from the NIST and Fisher corpora of speech data was assembled as an initial held-out dataset.

Data was chosen with the goal of matching or approximating conditions in the FBI provided corpus, although it was not possible to represent certain trial conditions (cross-language) and languages. Both telephone and microphone channels were represented with speakers in most languages offering cross-channel trials. Table 2 details the characteristics of this data. The segments provided 10736 target trials and 2.1 million impostor trials from which calibration parameters could be learned.

Table 2: Characteristics of the Matched Dataset.

Lang.	#Seg	#Spkr	Chan(s)	Source Corpora
Arabic	214	32	tel, mic	SRE
English	753	336	tel, mic	SRE
Hindi	165	52	tel	SRE
Korean	69	17	tel	SRE
Spanish	302	63	tel, mic	SRE

Table 3: Characteristics of the Large Variability Dataset.

Lang.	#Seg	#Spkr	Chan(s)	Source Corpora
Arabic	45	6	tel	SRE,RATS
Chinese	48	8	tel	SRE
Dari	74	13	tel	RATS
Pashto	160	25	tel	RATS
Urdu	136	21	tel	RATS
Other	73	20	tel	SRE,RATS
English	902	89	tel, clean / noisy / reverb mic	SRE

Large Variability Data: This dataset was collected to have a wide range of conditions and sources without regard for the conditions and channels in the FBI dataset. The intention here was to develop a general calibration set suitable for use in a deployed system where data variability will be higher than in research corpora. This will provide a means of measuring robustness to imperfectly matched conditions.

Data was sourced from NIST SRE corpora, clean data from the DARPA RATS SID task [8] (trimmed to 120 seconds of audio), and artificially reverberated and noisy NIST SRE data. This data was split into a calibration set of 1468 segments and the remaining 7k segments used for training the UAC extractor (disjoint sets were found to work better for all calibration techniques). There were 12930 target trials and 2.1 million impostor trials from the calibration set. Table 3 details the various conditions and languages in these datasets. Languages listed as other include Farsi, Hindi, Russian, Spanish, Thai, and Vietnamese.

6. Experimental Protocol

The system evaluated in this work was a gender-independent MFCC system consisting of fast, noise-robust voice activity detection [2], a 2048 Gaussian Universal Background Model (UBM), 600 dimensional i-vector subspace and 200D reduction of i-vectors via Linear Discriminant Analysis (LDA) before scoring with a gender-independent Probabilistic Linear Discriminative Analysis (PLDA) model [9]. Clean speech training data was sourced from the PRISM data set [10] with PLDA being additionally exposed to noisy and reverberant data [11].

System performance or accuracy is measured in terms of EER. Calibration performance is measured in terms of cost of

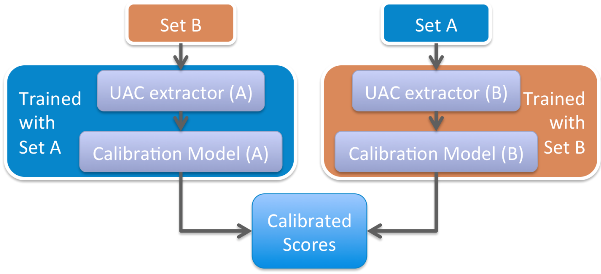


Figure 2: Cross-validation approach used in Sections 7.1 and 7.2 to highlight the benefits of Metadata calibration in the ideal scenario of matched evaluation and calibration data.

the likelihood ratio (Cllr), and calibration loss (Closs). Cllr [12] provides an indication of how well scores are calibrated across all operating points along a Detection Error Tradeoff (DET) curve [13]. In contrast, Closs provides an indication of how miscalibrated the system was at a particular operating point. For this work, the calibration goal is equal costs for miss and false alarm errors. Closs is calculated as the difference between the minimum cost (assuming perfect calibration) and actual cost. Note that Cllr is a more stringent metric and is not always correlated with Closs. For all metrics, a low value is desired.

7. Results

This section details the evaluation of different calibration approaches and draws conclusions on how effectively they accomplish the goals of this work. Throughout this section, the calibration data varies from closely matched to unmatched with regard to the evaluation data. The purpose of this is to contrast those calibration techniques that are dependent on knowing the evaluation conditions and those that are robust to unseen and highly varied conditions; the latter being the goal of this work.

7.1. On the use of metadata for calibration

This section commences with an ideal scenario in which evaluation and calibration training data come from the same source. To this end, we run cross-validation experiments. Specifically, the evaluation data was split into two subsets based on speaker label. Cross-evaluations were conducted in which subset A was used to train the calibration models for the evaluation of subset B and vice versa before calibrated scores were pooled and metrics were evaluated. This process is depicted in Figure 2.

The aim of using matched data in this section is to illustrate the full potential of calibration with metadata by ensuring accurate extraction of class posteriors and allowing the effectiveness of each UAC class to be determined prior to evaluation on mismatched calibration data in the following sections.

Initial experiments were aimed at determining how useful UAC and DPLDA were in mitigating miscalibration in the evaluation dataset. Table 4 details the different classes of UAC evaluated and the relative improvement in Cllr that each provided over global calibration when averaged across the 14 conditions. The combination of both language and channel classes was found to be most effective. The last row in the table indicates the gain that could be achieved with perfect UAC by calibrating each condition using data specific to that condition (i.e., 14 calibration models were used). These results indicate that, in the matched scenario, calibration with metadata is very effective.

Table 4: Comparing the relative Closs improvement of calibration with metadata over global calibration with different UAC classes when using evaluation data for calibration via cross-validation.

UAC Class(es)	Rel. Closs Improvement
SNR	1%
Channel	13%
Language	40%
Language+Channel	44%
Language+Channel+SNR	32%
Condition-Specific Calibration	42%

Table 5: Comparing the use of eval or matched data (cross-validated) for UAC model and calibration model training.

Cal. type	UAC set	Cal. set	Avg. Cllr	Avg. Closs	Avg. EER
Global	-	eval	.243	.048	4.49%
Metadata	eval	eval	.225	.027	4.85%
Global	-	matched	.511	.205	4.60%
Metadata	matched	eval	.243	.041	4.55%
Metadata	eval	matched	.502	.080	10.59%
Metadata	matched	matched	8.17	.101	26.51%

7.2. Calibration using unseen but matched data

The matched data was collected to include similar conditions to the evaluation dataset based on language and channel labels. In this section, we analyze the impact of using this external data instead of the highly matched evaluation data for training the UAC extractor, the calibration model or both components. Table 5 details results from these comparisons. It can be observed that using SRE data for the UAC extractor still benefits metadata-based calibration as long as the calibration model training data is well matched (evaluation data). Whenever matched data is used for calibration model training, calibration performance improves. This is particularly the case for calibration when matched data is used for training of both components.

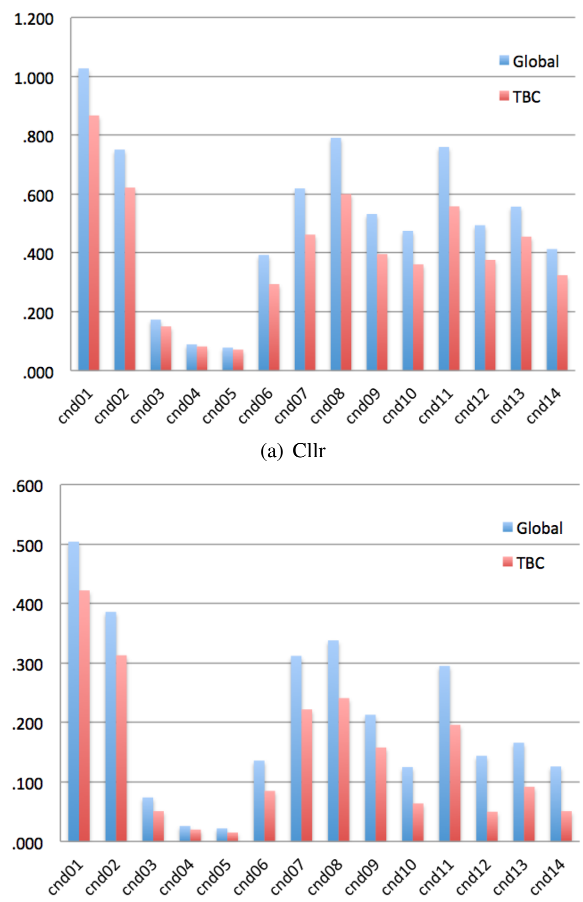
Several conclusions can be drawn from these results. The calibration performance is very sensitive to the data used to train the metadata-based calibration model. Despite the selection of SRE data that contain conditions similar to the evaluation data, the SRE data does not represent non-English cross-channel trials or cross-language trials, and does not contain all of the languages represented in the evaluation data. Given that global calibration with similar data was preferred over the metadata-based alternative when similar data was used for both UAC extractor and model training, this highlights a deficiency in metadata-based calibration to adequately accommodate unseen conditions. At worst, one would expect performance on par with global calibration.

7.3. Trial-Based Calibration (TBC)

The proposed TBC was motivated by the approach taken by forensic experts when performing SID. Specifically, the choice of data used to calibrate a trial score is delayed until conditions of both side of the trial are known. Following the procedure detailed in Section 4, we calibrated all 2.8 million scores from the evaluation set using TBC. Figure 3 compares the Cllr and Closs for global and TB calibration for each of the 14 condi-

Table 6: Comparing global, metadata-based and TBC calibration techniques using the large variability data.

Cond.	**Global Cal.**			**Metadata-based Cal.**			**Trial-based Cal.**		
	Cllr	**Closs**	**EER (%)**	**Cllr**	**Closs**	**EER (%)**	**Cllr**	**Closs**	**EER (%)**
01	.70	.40	1.25	.71	.35	1.71	.19	.09	0.85
02	.50	.29	0.95	.49	.24	1.90	.13	.05	0.95
03	.12	.03	1.68	.15	.02	3.35	.09	.01	1.12
04	.07	.01	1.33	.08	.02	1.33	.08	.00	1.33
05	.06	.01	0.86	.08	.01	1.93	.06	.00	1.29
06	.28	.08	4.36	.28	.01	6.70	.19	.01	3.67
07	.41	.21	4.84	.32	.03	8.36	.21	.01	4.84
08	.54	.23	6.52	.55	.04	13.26	.30	.01	6.74
09	.36	.16	2.97	.22	.03	4.65	.17	.04	3.11
10	.35	.06	7.29	.42	.02	10.55	.29	.01	6.25
11	.52	.19	6.95	.47	.01	12.59	.31	.00	7.38
12	.37	.05	8.26	.58	.05	14.22	.41	.06	8.07
13	.43	.10	9.80	.61	.07	13.73	.44	.09	7.90
14	.32	.06	7.39	.64	.09	12.61	.35	.02	7.39
Avg.	.36	.13	4.60	.40	.07	7.63	.23	.03	4.35



(a) Cllr

(b) Closs

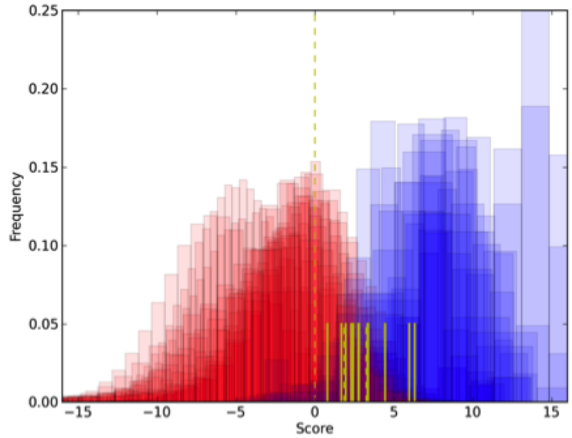
Figure 3: Illustrating Cllr and Closs improvements from TBC over global calibration when using matched data for calibration.

tions. Significant improvements of 20% and 35% in Cllr and Closs, respectively, were observed on average across the conditions. Additionally, the average EER across conditions fell from 4.60% to 4.35%. These results indicate that TBC can more readily adapt to unseen conditions than metadata-based calibration and provide better-calibrated scores for making identification decisions across various conditions.

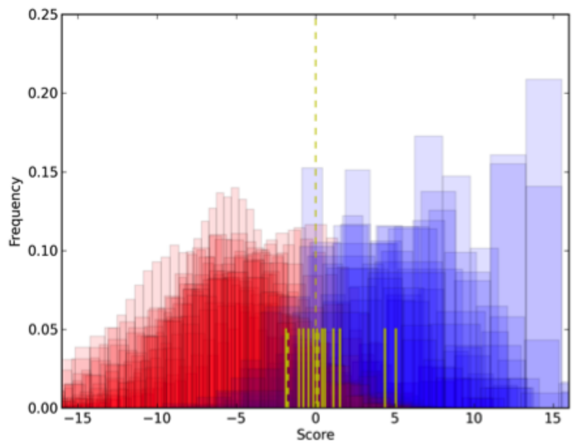
7.4. Unseen, Large Variability Data

The large variability data set was collected without regard to the conditions of the evaluation data with the intention of measuring the robustness of techniques to unseen evaluation conditions. This dataset was used in the evaluation of global calibration, metadata-based calibration and TBC, with the latter two utilizing UAC vectors. The UAC classes in this instance were extracted as language (seven classes in Table 3), channel (tel,mic), and degradation (clean, noisy, reverberated) as these were the most well represented classes in the calibration dataset. Results are detailed in Table 6. Compared to global calibration, calibration with metadata maintains Cllr on par with global calibration for conditions 01–11 but struggles with difficult conditions of 12–14. Closs improves using metadata-based calibration with the average Closs dropping from 0.13 to 0.07. This comes at a cost to accuracy with a doubling of EER in some conditions. This cost can be interpreted as the system being able to calibrate scores well by forcing the distribution of scores into a form that improves calibration, but reduces separability of target and impostor trial scores. TBC on the other hand, significantly improves calibration performance on the straightforward conditions (conditions 01 and 02), and does not reduce performance in very difficult calibration conditions. The Closs through TBC is significantly reduced from 0.13 to 0.03 over global calibration. As found in the previous section, the EER of the system also improved through calibration using TBC.

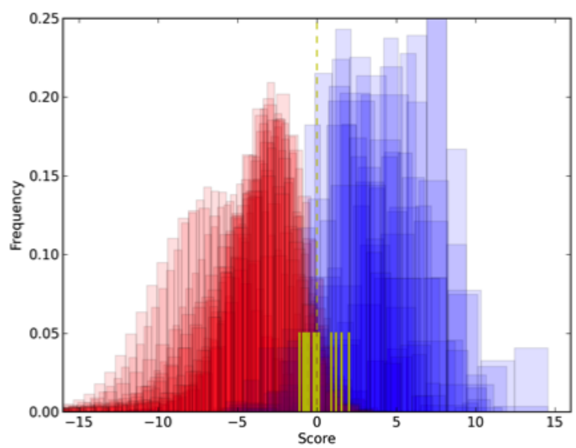
Figure 4 illustrates the threshold stability of the above techniques around the operating point with equal costs (approximate threshold of 0.0). The improved threshold stability offered by TBC over global and metadata-based calibration is clear from these plots.



(a) Global



(b) Metadata-based



(c) TBC

Figure 4: Illustrating the distribution of target and impostor trials (blue and red respectively) and the thresholds (yellow lines) for all 14 conditions at the operating point of equal cost when using the large variability data for UAC and calibration training. The tighter grouping of thresholds around 0.0 of TBC over Global and Metadata-based techniques is indicative of improved threshold stability across conditions.

8. Conclusion

We presented a novel approach to calibration that delays the learning of calibration parameters until conditions of both sides of the trial are known. Basic approaches to calibration were shown to work well with data taken from the same corpus as the evaluation set (cross validation results on the evaluation data), however this scenario is not realistic. Using data from a data set conditioned to the characteristics of the evaluation data showed current state-of-the-art approaches (global and metadata-based calibration) were heavily dependent on having observed conditions during training. We proposed trial-based calibration (TBC) to address these issues. TBC dynamically adapts the data used to produce calibrations dependent on the conditions of the trial. That is, it removes the need to make assumptions about the evaluation scenario by waiting until trial-time where conditions can be properly accommodated. TBC reduced calibration loss and cost functions while maintaining accuracy of the SID system. The use of a diverse calibration set (the large variability dataset) was particularly beneficial in this adaptive technique when dealing with the variety of conditions in the evaluation data.

Calibrating millions of trials using TBC requires considerable computation since each trial takes several seconds to select data and train a corresponding calibration model. Future work will consider methods of the reducing this complexity using closed-form training [14] and prediction of appropriate calibration model parameters based on the UAC vectors. Methods of measuring the quality of selected calibration data for a given trial will also be investigated.

9. References

[1] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, “A unified approach for audio characterization and its application to speaker recognition,” in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[2] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, “A noise-robust system for NIST 2012 speaker recognition evaluation,” in *Proc. Interspeech*, 2013.

[3] N. Brümmer, *FoCal-II: Toolkit for calibration of multi-class recognition scores*, August 2006, software available at http://www.dsp.sun.ac.za/∼nbrummer/focal/index.htm.

[4] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Briimmer, “Discriminatively trained probabilistic linear discriminant analysis for speaker verification,” in *Proc. IEEE ICASSP*, 2011, pp. 4832–4835.

[5] O. Plchot, S. Matsoukas, P. Matejka, N. Dehak, J. Ma, S. Cumani, O. Glembek, H. Hermansky, S. Mallidi, and N. Mesgarani, “Developing a speaker identification system for the darpa rats project,” in *Proc. IEEE ICASSP*, 2013.

[6] S. D. Beck, R. Schwartz, and H. Nakasone, “A bilingual multimodal voice corpus for language and speaker recognition (lasr) services,” in *Odyssey: The Speaker and Language Recognition Workshop*, 2004.

[7] 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*, 2013.

[8] K. Walker and S. Strassel, “The RATS radio traffic collection system,” in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. on Speech and Audio Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[10] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: The PRISM evaluation set," in *Proc. NIST 2011 Workshop*, 2011.

[11] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. IEEE ICASSP*. IEEE, 2012, pp. 4253–4256.

[12] N. Brümmer, "Application-independent evaluation of speaker detection," in *Proc. Odyssey, Speaker and Language recognition workshop*, 2004, pp. 33–40.

[13] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997, pp. 1895–1898.

[14] D. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition." in *Proc. Interspeech*, 2013, pp. 1619–1623.