INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The 2016 Speakers in the Wild Speaker Recognition Evaluation

Mitchell McLaren[1], *Luciana Ferrer*[2], *Diego Castan*[1], *Aaron Lawson*[1]

[1]Speech Technology and Research Laboratory, SRI International, California, USA
[2]Departamento de Computación, FCEN, Universidad de Buenos Aires and CONICET, Argentina
{mitch,dcastan,aaron}@speech.sri.com, lferrer@dc.uba.ar

Abstract

The newly collected Speakers in the Wild (SITW) database was central to a text-independent speaker recognition challenge held as part of a special session at Interspeech 2016. The SITW database is composed of audio recordings from 299 speakers collected from open source media, with an average of 8 sessions per speaker. The recordings contain unconstrained or "wild" acoustic conditions, rarely found in large speaker recognition datasets, and multi-speaker recordings for both speaker enrollment and verification. This article provides details of the SITW speaker recognition challenge and analysis of evaluation results. There were 25 international teams involved in the challenge of which 11 teams participated in an evaluation track. Teams were tasked with applying existing and novel speaker recognition algorithms to the challenges associated with the real world conditions of SITW. We provide an analysis of some of the top performing systems submitted during the evaluation and provide future research directions.

Index Terms: speaker recognition, speakers in the wild database, evaluation

1. Introduction

Evaluations provide a means of assessing the state of a certain technology across a number of groups that are working on a task. They provide the community of researchers in the area with a set of results against which to compare technology. They also motivate research to solve the specific problems posed by the evaluation data. Years after the evaluation is held, groups might still be trying to work on the problems associated with the data. By using a common evaluation dataset, evaluations allow comparison of results across publications, and the progress of performance on that data can be tracked throughout time.

For speaker recognition, the main evaluations that have been guiding a large part of the research on this task for two decades are the ones held by the National Institute of Standards and Technology (NIST) [1]. These evaluations have occurred every one or two years since 1996. They have evolved from using only telephone data to using additional microphone data from a variety of different microphones, telephone conversation and interview speaking style, different induced vocal efforts (low, normal and high), simulated noisy data (created by adding noisy signals to clean signals), and real noisy data collected from noisy environments. Some of these evaluations also included a "summed" condition in which the two channels of a telephone conversation or an interview were added together to create a multi-speaker recording which was then used in testing to determine whether a certain enrolled speaker was present in the recording. See [2] for a review of the NIST speaker recognition evaluation (SRE) series from 1996 to 2014.

While NIST speaker recognition evaluations provide great value to the community, they have focused on relatively controlled data. Although some challenging acoustic conditions have been explored, the dimensions of variability are restricted. This restriction facilitates the understanding of particular strengths or shortcomings of evaluated technology. However, these evaluations provide little insight into the performance of technology when applied to data collected in less constrained scenarios, such as open-source media in which multiple audio degrading artifacts are often convolved.

These observations motivated us to work on the collection and annotation of a new database which could fill in some of the gaps presented by the data used in NIST speaker recognition evaluations. As a result, we created the Speakers in the Wild (SITW) database [3], a new database designed for text-independent speaker recognition. The database consists of audio recordings from open source media and contains a wide variety of acoustic conditions, including real background noise, reverberation, compression artifacts and large intra-speaker variability. Furthermore, the database contains audio segments that include multiple speakers: some in interview or dialog situations, and some in more uncontrolled scenarios where multiple speakers might be involved. Multi-speaker audio is not only used for testing, but also enrollment with the aid of a small annotation.

In 2016, SRI organized a speaker recognition challenge based on the SITW database. A total of 25 international teams from 18 different countries participated in the challenge to evaluate technology on the database. As part of the challenge, an optional evaluation was held in which 11 of the teams participated. These teams submitted a description of their efforts for the evaluation to the challenge organizers (the authors of this article). In this work, we provide a summary of these submissions to draw attention to how current technology fairs on the SITW database and provide future research directions. We anticipate the results and publications that result from the challenge and the corresponding database, which is publicly available for research purposes, will motivate the community to spend time and effort trying to solve some of the challenges that still remain in the speaker recognition task.

2. The SITW Evaluation

The SITW evaluation was based on the SITW database [3]. The SITW database aims to provide a large collection of real world data that exhibits speech from individuals across a wide array of challenging acoustic and environmental conditions. Additionally, SITW includes multi-speaker audio from quiet set interviews, noisy red-carpet interviews, reverberant question and answer sessions in an auditorium, and more casual conversational multi-speaker audio in which backchannel, laughter, and overlapping speech is observed. Each individual also has

http://dx.doi.org/10.21437/Interspeech.2016-1137

raw, unedited camcorder or cellphone footage in which they speak. This footage potentially contains other speakers and (often) spontaneous noises. The audio of the SITW database was extracted as partial excerpts of the audio track from open-source media (videos). The data was not collected under controlled conditions and thus contains real noise, reverberation, intraspeaker variability and compression artifacts.

The evaluation consisted of two enrollment and two test conditions. The enrollment conditions were: (1) **core**, where audio files contain 6 to 180 seconds of contiguous speech from a single speaker; and (2) **assist**, where the audio files contain speech from one or more speakers, including the speaker of interest. In the assist case, the recordings contain anywhere from 6 seconds to more than an hour of speech from the speaker of interest. For this condition, a small annotation, or "seed," is provided to indicate a region where the speaker of interest has been verified to be speaking. This seed is used to assist systems in expanding the amount of data that can be used for enrollment. The two test conditions were: (1) **core**, where the audio files have the same characteristics as the core ones in enrollment; and (2) **multi**, where the audio files contain one or more speakers, one of which might be the speaker of interest. If so, the amount of speech from that speaker can be approximately from 6 seconds to 10 minutes. Note that the **multi** test samples do not coincide with the **assist** enroll samples due to differences in the design criteria for these two sets (see [3] for details).

Four evaluation conditions were created by combining each enrollment with each test condition[1]. These trial conditions are denoted as *enroll-test* (i.e., **core-multi** denotes the **core** enrollment and **multi** test trial condition). Cross-gender trials were included in all conditions. The SITW database was split in two sets for the purpose of the evaluation: a development set and an evaluation set. Sets were disjoint in terms of speakers, with 2,597 target and 335,629 impostor trials from 119 unique speakers in the development set and 3,658 target and 718,130 impostor trials from 180 unique speakers in the evaluation set. The evaluation trial set included approximately 11% female and 45% male same-gender trials, and 44% cross-gender trials.

The rules of the evaluation were quite standard: (1) any publicly available or previous NIST SRE data could be used for training the system, including the SITW development data; (2) enrollment of speaker models had to be treated independently of all other available data; (3) participants had to submit a score (rather than a decision) for each trial, and those scores were treated as log-likelihood ratios for performance computation; and (4) only the **core-core** condition was compulsory, and sites could choose to submit to the alternate conditions.

The primary metric for the evaluation was a standard C_{det}, as used in all NIST SREs, with costs of 1 for both errors and a probability of target of 0.01. The C_{det} was computed by thresholding the scores provided by the participants at the theoretically optimal threshold for these costs (4.59). Participants were provided a scoring script that also computed the minimum C_{det}, C_{llr} and average R_{prec} or $\bar{R}_{prec}$. For details on these metrics, please refer to [3].

3. Evaluation Results

In this section, we show overall evaluation results for all teams, as well as some more detailed analyses of subconditions. The goal of showing these results is to set a baseline performance for the SITW trial conditions and highlight some challenges present in this data. Given the complexity of this dataset, the analysis is not always straightforward, as we will see in many of the results. Nevertheless, interesting conclusions can still be gathered by dissecting results in certain ways.

Table 1: Results for the best submission from each of the sites for each condition. Darker green indicates a better systems.

Cond	Site	C_{det}	minC_{det}	aveR_{prec}	EER	Cllr
core-core	1	0.51	0.50	0.75	0.059	0.21
	2	0.65	0.60	0.66	0.087	0.29
	3	0.65	0.64	0.64	0.077	0.34
	4	0.76	0.74	0.54	0.114	0.38
	5	0.84	0.84	0.47	0.119	0.59
	6	0.87	0.86	0.42	0.160	0.51
	7	0.88	0.87	0.41	0.145	1.55
	8	0.92	0.92	0.35	0.166	0.51
	9	0.94	0.93	0.52	0.121	0.42
	10	1.00	1.00	0.09	0.244	-
	11	1.60	0.95	0.36	0.173	0.78
core-multi	1	0.58	0.57	0.66	0.073	0.31
assist-core	1	0.40	0.40	0.75	0.045	0.17
	3	0.54	0.53	0.66	0.064	0.32
assist-multi	1	0.47	0.46	0.72	0.057	0.24

For some results, we show a 95% confidence interval, which was calculated using a modified version of the joint bootstrapping technique described in [4]. The modification is performed to account for the fact that many models are created for each speaker of interest. Having multiple models per speaker, which might even be enrolled with different snippets from the same session, introduces a very strong correlation across trials involving those models. To this end, we simply add another layer of sampling: speakers are sampled first, then models from those speakers, then test signals. The models themselves might be repeated if a speaker was sampled more than once in the first layer of sampling. The trials corresponding to the selected subset of models and test signals are then used to compute the performance metric. We performed the sampling 20 times for each layer to produce 8000 measurements of the metric. The confidence interval that is reported corresponds to the 5 and 95 percentiles of the resulting empirical distribution.

3.1. Results for all trial conditions

Table 1 shows the results for the best submission from each of the 11 sites. As indicated in the evaluation plan, all sites submitted scores for at least one system for the primary **core-core** condition. For the other conditions, only one or two sites submitted scores. The numbers in the table indicate the site. Note that even though this number is the same across conditions, that does not imply that the same system (architecture, parameters, etc) was run across conditions. In fact, systems varied across conditions to accommodate the different characteristics in the trials.

The first observation we can draw from the results is that the top systems reach impressive performance for this challenging data, with the best system achieving an EER of less than 6%. Clearly, however, this performance is not easily achievable, since only a handful of systems were able to approach that level of performance. Interestingly, only the top three systems leveraged senone Deep Neural Networks (DNN) in their architecture as in [5] or [6, 7]. The fourth system was based on a standard UBM/i-vector architecture including source normalization [8] to reduce mismatch between system training and

[1] Two more conditions in the eval plan corresponded to a subset of the assist enrollment condition which contained only clean data for enrollment. Here, we consider those conditions as subsets of the main assist-core and assist-multi conditions.

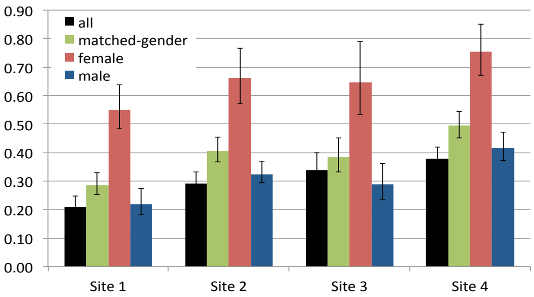

Figure 2: Results (C_{llr}) for the **core-core** trials (all), and the subset of matched-gender trials and gender-dependent trials.

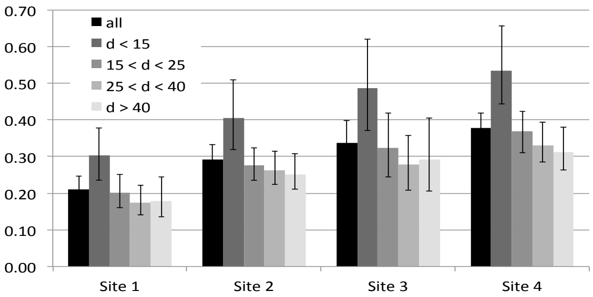

Figure 3: Results (C_{llr}) for the **core-core** trials (all), and the test duration-dependent subsets.

evaluation data sources. Site 2 also utilized source normalization with the SITW development data forming one of the "sources" in this approach. In the next section we will show more detailed results for these top four systems. Additional system characteristics of interest include Site 1's use of a phoneme recognizer for SAD, in contrast to other sites' use of energy-based SAD, spectral matching or self-adaptive algorithms. Sites 1, 3, 6, and 7 used the fusion of 2 or 3 subsystems, while others used a single system. Calibration parameters for Sites 1–9 were trained directly on the SITW development trial set, while Sites 10 and 11 did not apply calibration.

Note that all conditions include cross-gender trials. This has the effect of improving performance with respect to a trial set based only on matched gender trials. For example, for the top site (first line in Table 1), the C_{det}, EER and C_{llr} for matched-gender trials are 0.578, 0.0768 and 0.285, respectively. C_{llr} results for matched-gender trials are also shown, along with confidence intervals in Figure 2. Comparing these results with those for all trials, we see that the presence of cross-gender trials (which represent 44% of all trials) makes the task significantly easier for this system. Similar improvements can be observed for other top systems.

Most systems had excellent calibration performance, with values of minimum C_{det} very close to actual C_{det}. This performance is likely due to the fact that the SITW development data was a good match to the evaluation data and most sites used this data to calibrate their scores. It is interesting to observe that the C_{llr}, a metric that measures the quality of the scores over all possible operating points by assuming them to be well-calibrated log-likelihood ratios, correlates well with the C_{det} for the top systems. These are the systems that are, indeed, well-calibrated across all operating points, and not just on the point defined by C_{det}.

We can see from Table 1 that the **core-multi** condition was more difficult than **core-core** for the single site that ran both. Note that the **multi** test signals include all **core** (single-speaker) test segments to allow for analysis of whether the segmentation of test samples adversely affects single-speaker audio. Due to a lack of enough submissions involving **multi** tests, we refrain from this analysis here. All other test samples include multi-speaker segments, many of which have short speaker turns, overlapping speech and other conversational aspects such as backchannel and laughter. An analysis of these nuisances on the effect of speaker recognition is yet to be conducted. It is worth noting that most speaker diarization algorithms have been designed to allocate all detected speech across the automatically-defined speaker clusters. It would be interesting to tailor these algorithms toward the speaker recognition task by only retrieving speech that the system can confidently determine as that of the speaker involved in the trial.

The two **assist** enrollment conditions appear easier than the corresponding **core** enrollment conditions. It should be noted, however, that the comparison in Table 1 is not direct, since the **assist-core** speaker models are based on different annotation lengths, and not all audio used to enroll the **core** models had a corresponding assist version. To allow for a direct comparison, we created two subset conditions: one where the **core-core** trials are subsetted to include only models for which an **assist** version exists, and another one where the **assist-core** trials are subsetted to include only one model for each session using the longest possible seed which coincides with the **core** signal in the **core** model set. Both subsets then include an identical number of comparable trials: the test samples are the same and for each **core** model, there is a corresponding **assist** model that uses the **core** snippet as the annotation and includes additional speech. For Site 1, results for the **assist-core** subset are better than for the **core-core** one. The trend is reversed for Site 3 (results not shown for lack of space). This means that the gain from the additional data is not guaranteed, and is most likely dependent on the quality of the diarization process that is performed to discard any irrelevant speech in the signal. Broadly speaking, both sites applied unsupervised speaker diarization on the audio before considering which speaker cluster from diarization shared the most overlap with the annotation. Further analysis found limited difference between short and long enrollment annotations (5s, 10s, 15s or >15s) when comparing within-site results. Consequently, we can presume that detection of speech from the speaker of interest (recall) was adequate to result in similar enrollment speech. Precision may need to be improved to ensure the additional speech is only from the speaker of interest in the assist audio.

3.2. Results for subsets of the core-core condition

In this section, we analyze performance on the **core-core** condition by splitting the trials into different subsets. For this analysis, we focus on the four top systems from Table 1. These results are shown in terms of C_{llr}, since this is a more general metric than the actual DCF, which focuses on a single operating point. While the evaluation keys were designed to discard any symmetric trials (that is, trials that interchanged enrollment file with test file), we decided to include these trials in the results for this section before the trials were subsetted, since subsets are mostly done in terms of test samples. To this end, we simply assumed that systems would generate identical scores for the symmetric trials and, consequently, automatically created the scores for the missing trials of submitted systems.

3.2.1. Results by gender

Figure 2 shows the results for all **core-core** trials, for the subset of matched-gender trials, and for the two gender-dependent sub-

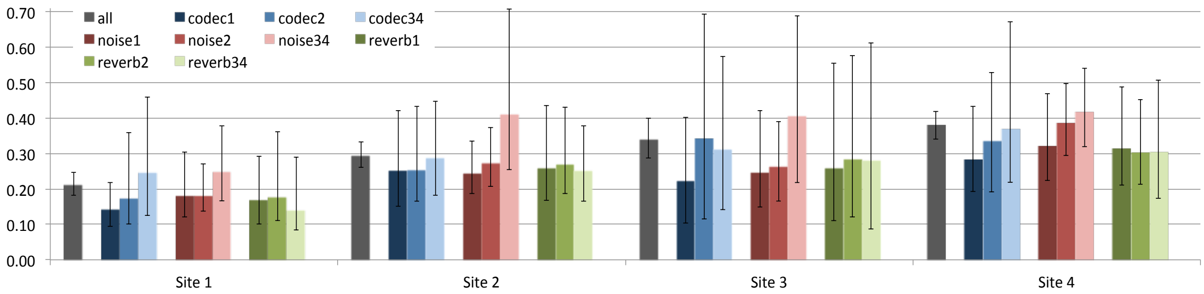

Figure 4: Results (C_{llr}) for the **core-core trials** (all) and some degradation-dependent subsets denoted by *type* and *level*.

sets (the matched-gender subset is the union of the two gender-dependent subsets). Results show that females pose a much harder challenge to the top systems than males. While the fact that females usually have worse speaker recognition performance than males is a well-known fact (e.g., [9, 10]), the difference in this case is somewhat larger than expected. This seems to be due mostly to poor discrimination power rather than poor calibration, since the EER (a metric that is independent of calibration) for the first system is 13.7% for females and 5.8% for males, with similar relative differences for the other top systems.

3.2.2. Results by duration

Figure 3 shows the results for all **core-core** trials and for subsets of these trials where the test files have been binned by their detected speech duration as detected by our SAD system (described in [11]). We can see that the trends by duration are as expected, with shorter files being significantly harder than longer files. Interestingly, the degradation seems to saturate after 25 seconds (the last two bins have similar performance) for Sites 1 and 3. It is possible that duration mismatch between enrollment and test samples is responsible for this trend. Specifically, the enrolled speaker models have the same speech duration distribution as the test segments with a bias toward 20-25 seconds. Due to the limited number of trials that result from a model and test subset required for this analysis, this hypothesis is difficult to support using the SITW database.

3.2.3. Results by degradation level and type

Figure 4 shows the results for different degradation levels for three common types of degradation. These degradation types (noise, reverb, codec) and levels (0-4) were those perceived by a single human annotator. Test files in this analysis exhibit only a single degradation type: a small subset of all audio files in the SITW database, which contains mostly files with multiple types of degradation. We can see that, for both codec and noise types, the degradation level is a good predictor of performance: higher degradation levels imply worse performance. This is not the case for the reverberation type, for which the degradation level seems to have no correlation with performance. Interestingly, the degradation due to the highest level of noise affects the top system relatively less than the other systems. This system seems to be especially good at mitigating the effect of noise.

4. Conclusions and future directions

The SITW database, which is freely available for research purposes, provides a new context for the evaluation of speaker recognition evaluation: real world conditions associated with audio from open-source multimedia. Based on the submissions of 11 international research teams, the analysis presented in this article has shed light on some of the fundamental issues that remain yet unaddressed in the technology, as well as aspects of the database that require further investigation. We summarize these here as future research directions.

Perhaps the most obvious factor for further study is the significant performance difference observed between male and female trials in Section 3.2.1. Our preliminary attempts to dissect these results to determine whether female trials consist of generally greater degradation, degradation type (i.e., babble instead of outside noise) or duration have not provided a clear indication as to why female trials are twice as difficult as male trials.

Assisted enrollment is a new paradigm for many speaker recognition research groups. Submitted systems utilized unsupervised speaker diarization prior to leveraging the information of the provided annotation to determine the enrollment speech for a speaker model. Development of methods that use the annotation directly in the segmentation process to target speech of a known speaker (the annotation of the assist conditions or the speaker model of a trial), rather than first allocating all speech to unsupervised speaker clusters, may improve performance for this speaker recognition task. This approach may be particularly useful in the context of spontaneous, conversational speech as exhibited in the SITW audio.

Regarding system design trends, almost all submissions consisted of energy-based SAD as opposed to the more advanced, noise-aware SAD that was used in the top performing submission. Although much of the SITW data was sourced from interview scenarios which naturally involve a high speech vs non-speech ratio, simple energy-based SAD may not be the most appropriate selection to cope with factors such as babble, background music, and spontaneous noises. Given the uncontrolled nature of the SITW audio, we expect robust SAD algorithms such as those developed under the DARPA RATS program [12, 13, 14, 15, 16] to be a key component to good performance on the SITW data.

Calibration is a key component of any deployed speaker recognition system. Many teams calibrated using the SITW development data which provided a suitable “overall” representation of the dataset. However, as the trial conditions are far from homogeneous, calibration methods that dynamically take into account trial conditions [17, 18] can be expected to improve on a simple calibration model (shift and scale) when a single threshold is applied (C_{det}), or calibration across all operating points is considered (C_{llr}).

In this article we have tried to indicate trends across submissions and draw conclusions where statistical significance between systems exists. As future research is pursued on the SITW database, we recommend that care be taken when attempting to dissect results and draw conclusions from trial subsets, since conditions are often biased toward or dependent on one another due to the nature of real world data.

5. References

[1] "NIST Speaker Recognition Evaluations," http://www.nist.gov/itl/iad/mig/sre.cfm.

[2] J. Gonzalez-Rodriguez, "Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014)," *Loquens*, vol. 1, no. 1, 2014.

[3] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," in *submitted to Interspeech 2016*, 2016.

[4] N. Poh and S. Bengio, "Estimating the confidence interval of expected performance curve in biometric authentication using joint bootstrap," in *Proc. ICASSP*, Honolulu, Apr. 2007.

[5] 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.

[6] P. Matejka, L. Zhang, T. Ng, S. H. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in *Proc. Odyssey-14*, Joensuu, Finland, Jun. 2014.

[7] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Proc. ICASSP*, Brisbane, Australia, May 2015.

[8] M. Mclaren and D. Van Leeuwen, "Source-normalized lda for robust speaker recognition using i-vectors from multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 755–766, 2012.

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

[10] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, Florence, Italy, Aug. 2011.

[11] 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*, Lyon, France, Aug. 2013.

[12] "DARPA RATS program," http://www.darpa.mil/program/robust-atuomatic-transcription-of-speech.

[13] S. Thomas, G. Saon, M. Van Segbroeck, and S. S. Narayanan, "Improvements to the IBM speech activity detection system for the DARPA RATS program," in *Proc. ICASSP*, Brisbane, Australia, May 2015.

[14] J. Ma, "Improving the speech activity detection for the DARPA RATS phase-3 evaluation," in *Proc. Interspeech*, Singapore, Sep. 2014.

[15] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. Hansen, A. Janin, B. S. Lee, Y. Lei, V. Mitra *et al.*, "All for one: Feature combination for highly channel-degraded speech activity detection," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[16] L. Ferrer, M. Graciarena, and V. Mitra, "A phonetically aware system for speech activity detection," in *Proc. ICASSP*, Shanghai, China, March 2016.

[17] 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.

[18] 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.