INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Analysis of Critical Metadata Factors for the Calibration of Speaker Recognition Systems

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

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, California, USA
[2]Instituto de Investigacón en Ciencias de la Computación, UBA-CONICET, Argentina

{mahesh.nandwana, mitchell.mclaren, diego.castan, aaron.lawson}@sri.com,
lferrer@dc.uba.ar

Abstract

In this paper, we analyze and assess the impact of critical metadata factors on the calibration performance of speaker recognition systems. In particular, we study the effect of duration, distance, language, and gender by using a variety of datasets and systematically varying the conditions in the evaluation and calibration sets. For all experiments, the system is based on i-vectors and a probabilistic linear discriminant analysis (PLDA) back-end and linear calibration. We measure system performance in terms of calibration loss. Our experiments reveal (i) a large degradation when the duration used for calibration is significantly different from that in the evaluation set; (ii) no significant degradation when a different gender is used for calibration than for evaluation; (iii) a large degradation when microphone distance is significantly different between the sets; and (iv) a small loss for closely related languages and languages with shared vocabulary. This analysis will be beneficial in the development of speaker recognition systems for use in unseen environments and for forensic speaker recognition analysts when selecting relevant population data.

Index Terms: speaker recognition, calibration, metadata, calibration loss.

1. Introduction

A speaker identification (SID) system produces an output score when comparing the voice similarity of an unknown speaker sample to a speech sample from a claimed identity. A more positive score indicates stronger support for the voices being from the same individual [1]. Unfortunately, these scores contain bias due to condition mismatch between the data that the system or speaker model was trained on and the conditions of the verification speech. Unless accounted for, this bias results in an inaccurate assessment of voice similarity; the similarity appears either stronger or weaker than it should. Bias is commonly corrected with a technique termed calibration, in which a calibration model, typically consisting of shift and scale parameters, is trained using a calibration dataset that reflects the expected end-use conditions. The calibration step converts the system scores into meaningful output, known as log-likelihood ratios (LLRs). The LLRs have a clear probabilistic interpretation and can be either used directly in some applications, like forensic voice comparison, or converted to binary decisions by applying a score threshold for other applications, such as user authentication [2, 3].

In this scientific study, rather than focusing on improving the performance of the speaker recognition system, we focus on gaining a deeper understanding of the impact of different metadata factors and their effect on calibration performance. The motivation comes from the fact that different factors influence calibration (or mis-calibration) to varying degrees and understanding and attending to the most critical factors during calibration of a system will reduce the potential of mis-calibration in a deployed system. To exemplify the importance of these factors, Figure 1 shows the distinct difference between score distributions of two datasets despite both consisting of English, close-talking microphone samples. Using one set to calibrate a system for the evaluation of the other would result in mis-calibrated LLRs. The main objectives of this analysis are two fold. First, this analysis will be beneficial for developing speaker recognition systems intended for use in unseen environments. Second, this analysis will assist forensic analysts when selecting relevant populations for their casework in forensic speaker recognition.

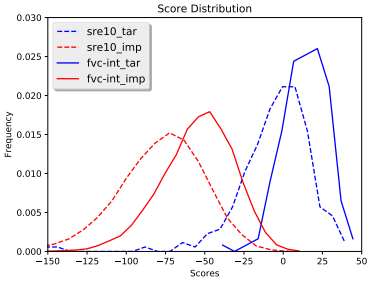

Figure 1: *Score distribution bias is evident across two seemingly similar datasets. An decision threshold tuned on one dataset, would not be applicable to the other.*

A speaker recognition system can be used in various operating conditions. These conditions vary in terms of language, distance, noise levels, vocal efforts, duration, compression etc. However, knowing a priori the precise operating conditions once deployed is very difficult, thereby reducing the ability of a pre-trained calibration model to generalize to such conditions. Similarly, in forensic casework, forensic analysts consider the degree of voice similarity between the samples and their typicality with respect to the relevant population. The relevant population is the population of speakers from which the target speaker could have conceivably come from. The selection of a relevant population in forensics is of critical importance and its choice is conducted on a case-by-case basis. The factors affecting the choice of the relevant population can be extrinsic or intrinsic in nature [4, 5, 6, 7].

In this work, we analyze how various known conditions impact system calibration performance. Once we assess the impact of each condition, this knowledge can be folded into the training of the calibration model for speaker recognition systems intended for use in operational environments or in the selection of relevant population data for forensic casework. In the past, a limited number of studies have focused on analyz-

http://dx.doi.org/10.21437/Interspeech.2019-1808

ing critical metadata factors, and include the impact of duration [8], aging [9], vocal effort [10], Lombard effect [11], and whisper [12]. Here, we analyze four critical metadata factors including gender, duration, distance, and language in a large cross-dataset study.

The remainder of this paper is organized as follows: First, the speaker recognition system used in this study is detailed in Section 2. This is followed by a description of the datasets used in this study in Section 3. Section 4 describes the experimental results of the impact of critical metadata factors on the calibration performance of the speaker recognition. Finally, conclusions and directions for future work are presented in Section 5.

2. Speaker Recognition System

For this work, we used a speaker recognition system based on the i-vector framework [13] with a probabilistic linear discriminant analysis (PLDA) back-end. We select the i-vector framework over the more recent and powerful deep neural network (DNN) speaker-embedding architecture [14, 15] due to the current widespread use of i-vectors, the fact that Gaussian assumptions of the framework align better with forensic case work that may be presented in court, as well as our need to limit system training data in order to provide sufficient cases for our analysis. More specifically, we used a number of datasets for our analysis (detailed later in Section 3) that are more commonly used in the training of SID systems. However, we had to exclude them from the training of the system in this work to avoid overlap between training and evaluation data. For training of the speaker recognition system components, we used the NIST SRE 2008 corpora, which consists of more than 19,000 samples.

The main components of our system are speech activity detection (SAD), front-end feature extractor, i-vector extractor and a PLDA back-end. SAD is similar to the model used in [16] with a threshold of 0.5 to obtain the speech regions.

2.1. MFCC and i-Vector Extraction

The MFCC acoustic features were based on 20-dimensional MFCCs, including C_0, spanning the frequency range of 200–3300 Hz using 24 filter banks, a window of 25 ms, and a step size of 10 ms. MFCCs were contextualized with deltas and double deltas prior to filtering out non-speech frames, as determined by the SAD system, and the application of utterance-level mean and variance normalization over the resulting speech frames. We trained a gender-independent universal background model (UBM) with 2048 Gaussians followed by a 400-dimensional i-vector extractor [13].

2.2. Probabilistic Linear Discriminant Analysis (PLDA)

We used gender-independent PLDA [17] for all our experiments described herein. Before training the PLDA classifier, the dimensions of the i-vectors were reduced to 200 by using linear discriminant analysis (LDA), followed by mean centering to Gaussianize the distribution of the i-vectors, and finally, length normalization. The PLDA classifier used the normalized i-vectors to compute similarity scores between enrolled speaker models and test samples of each data set. These scores served as the input to the calibration stage.

2.3. Score Calibration

Although PLDA outputs LLRs, they are typically not interpretable as calibrated LLRs because of a mismatch between the conditions of the system training data and trial conditions. It is commonplace, therefore, to apply calibration to transform these raw scores into interpretable LLRs. We use a linear calibration transformation in which raw scores s are transformed into calibrated scores s_c, given scaling and offset parameters α and β:

$$s_c = \alpha s + \beta \tag{1}$$

where α and β are obtained by logistic regression optimization on a calibration dataset.

We measure the calibration performance of the speaker recognition system in terms of the cost of likelihood ratio (C_{llr}). The C_{llr} provides an indication of discrimination performance of the system as well as how well calibrated scores are across all operating points on the detection error tradeoff (DET) curve [2]. If the value of C_{llr} is above 1, this indicates the information obtained from the system is worse than random.

For each acoustic condition, we would like the C_{llr} to be close to the C_{llr} that would be achieved on that condition if perfectly matched data were available for calibration. Hence, we show results in terms of C_{loss}, defined as the relative difference between C_{llr_T}, which is the actual C_{llr} on a certain set of trials, and C_{llr_M} obtained when calibrating with a perfectly matched calibration set for that test set.

$$C_{loss} = (C_{llr_T} - C_{llr_M})/C_{llr_M} \tag{2}$$

In this study, C_{llr_T} is calculated using a calibration model trained on scores from a held-out calibration set (development set), whereas C_{llr_M} is obtained by calibrating using a model trained on the evaluation scores directly. We report our results in terms of calibration loss percentage. The values above 20% may be considered a large calibration loss.

3. Datasets

A large effort was made to design datasets consisting of various homogeneous sets of trials, each of them divided into evaluation and calibration sets with disjoint speakers. The homogeneous sets enabled us to obtain a target calibration performance C_{llr_M} for each test set by training the calibration model on the matched calibration set. The scores obtained with this calibration model were used to compute the C_{loss} for each test set. Further, using homogeneous sets enabled our analysis into which specific kinds of variation significantly affect calibration. The homogeneous sets for analysis of metadata factors used were benchmarked on our speaker recognition system with results reported in Table 1 to serve as a reference. With the exception of the SWPH2 and SWCELLP1 datasets, samples consisted of taking from each original file, a single cut of audio that contained 20 seconds of speech content.

3.1. Switchboard Datasets

Switchboard datasets consist of English telephony conversations. We used two Switchboard sets for our impact assessment of duration and gender.

Switchboard-2 (SWPH2) consists of landline telephone conversations. For this set, we included six different subsets produced by cutting each signal to 5, 10, 20, 40, 80, and 160 seconds of speech content for only male speakers. For the female speakers, we created only the 20-second cuts.

Switchboard Cellular (SWCELLP1) consists of cellphone conversations prepared in the same way as SWPH2.

Table 1: *Speaker recognition performance of datasets used in this study. The held-out calibration subset of each dataset was used to calibrate the system prior to benchmarking the eval set.*

Dataset	Target/Imposter	EER (%)	*Cllr*
SWPH2	2.2k/359.1k	9.80	0.363
SWCELLP1	549/34k	8.53	0.388
SRI-Dist	2.3k/133.3k	9.83	0.372
FVC-int	326/54.2k	2.45	0.115
FVC-cod	326/54.2k	5.53	0.211
FVC-rev	326/54.2k	8.28	0.290
FVC-noi	978/162.6k	11.04	0.374
RATS-G	1k/38.1k	18.36	0.615
SRE10	440/57.4k	5.45	0.340
RATS-clean (Alv)	276/3588	12.76	0.469
RATS-clean (Urd)	540/13.5k	12.82	0.475
RATS-clean (Fas)	198/1584	22.22	0.642
RATS-clean (Prs)	294/4116	25.89	0.752
RATS-clean (Pus)	858/34.5k	24.84	0.724

3.2. SRI Distant Speech Collect Dataset (SRI-Dist)

The SRI distant speech corpora, which was generated in-house at SRI by transmitting original Forensic Voice Comparison (FVC) [18] audio files from a loudspeaker and capturing it by using two different microphones (a studio mic and a lapel mic) at a number of distances. The microphones were placed at 76, 94, 140, 150, 155, 178, and 259 centimeters from the speaker. For this dataset, only the conversation speaking style from FVC was used [16]. We only used Room 1 for our experiments.

3.3. Forensic Voice Comparison Dataset (FVC)

The FVC dataset consists of 544 Australian English speakers [18]. The speakers were recorded in three different speaking styles – interview, conversation, and fax (read) – with multiple sessions available for each speaker. For this work, we only used interview speaking style perturbed with noise, reverberation and compression, and only the test segments were perturbed and the enrollment samples left unaltered.

FVC-int was formed from the interview signals of 417 male speakers without any degradation applied.

FVC-cod was prepared by transcoding the test segments of FVC-int using the GSM codec.

FVC-rev was prepared by adding reverberation to the test segments of FVC-int.

FVC-noi was generated by adding three different signal-to-noise ratio (SNR) levels (8 dB, 15 dB, and 20 dB) from a cafeteria type of noise to the segments of FVC-int.

3.4. NIST SRE 2010 (SRE10)

The SRE10 set in this study was formed using the close-talking microphone samples of the NIST SRE'10 dataset (channel 02) from 178 male speakers.

3.5. DARPA RATS

From the DARPA Robust Automatic Transcription of Speech (RATS) dataset [19], conversational telephone speech (CTS) data was re-transmitted through eight different radio communication channels. The RATS data represents five different languages: English, Pashto, Farsi, Urdu, and Arabic. We created two subsets from the RATS for our experiments.

RATS-G for our study included only the Pashto language and re-transmitted channel G. We cut the audio into 20-second speech snippets and included only male speakers.

RATS-telephone was used to study the effect of language. To produce this set, we used the original telephone signals with speech from males, and exclude any cross-language trials.

4. Analysis of Critical Metadata Factors

In this section, we analyze the effect that each type of variation has on calibration performance. The following tables show the C_{loss} when calibrating on a specific portion (for example, one language of five) of a pre-defined calibration set and testing on a specific subsets of evaluation set. When the calibration and the evaluation sets are matched, the C_{loss} is 0 by definition (since we compute the C_{loss} with respect to matched calibration performance). Note that we could also, in fact, get negative C_{loss} values. This would mean that the corresponding calibration set produced a better calibration model than the matched calibration model for that test set. As we will see, this occurs in several conditions to only a small degree.

4.1. Duration Analysis

Figure 2 shows the results when varying the duration of the evaluation dataset compared to the calibration set for the two Switchboard datasets.

Cal\Test	5s	10s	20s	40s	80s	160s
5s	0.0	5.1	39.7	56.0	50.9	71.1
10s	2.9	0.0	22.3	27.2	19.0	32.0
20s	47.0	34.5	0.0	5.4	7.0	15.3
40s	60.1	33.0	1.4	0.0	0.4	4.4
80s	74.2	35.0	8.5	1.4	0.0	1.0
160s	72.4	33.4	0.0	0.7	-1.0	0.0

(a) *Switchboard Cellular (SWCELLP1)*

Cal\Test	5s	10s	20s	40s	80s	160s
5s	0.0	2.0	34.1	80.5	117.3	139.1
10s	27.5	0.0	1.5	28.6	56.2	69.9
20s	95.8	35.6	0.0	1.1	15.5	22.2
40s	175.6	92.8	25.4	0.0	0.9	3.0
80s	221.0	133.7	52.3	9.1	0.0	-0.1
160s	221.2	133.8	0.0	9.1	0.1	0.0

(b) *Switchboard-2 (SWPH2)*

Figure 2: *C_{loss}(%) for duration variation when calibrating on one duration and applying it to the evaluation of another duration of the Switchboard datasets.*

We see a large degradation (more than 20 points) when the durations used for calibration are significantly different from those in the evaluation set, supporting the conclusions of [8] and highlighting the requirement for a speaker recognition system to take into account duration information to be robust to mis-calibration. Note that while the enrollment and test duration were the same in Figure 2, the mismatch of enrollment and test sample durations would be expected to contribute a similar impact on calibration performance, and therefore a good system design would ensure calibration parameters account for such mismatch. Fortunately, this can be taken into account in an automatic speaker recognition system by having a pre-trained set of calibration models on hand that represent a range of enrollment-test duration pairs and selecting the closest calibration model dynamically at test time based on the duration of speech detected via SAD [3].

4.2. Gender Analysis

Figure 3 shows the results for calibrating on the same or different gender trials for the two Switchboard datasets. In this case, no significant degradation (fewer than five points) was observed when a different gender was used for calibration with respect to the gender being evaluated. This implies that the gender-independent training of our system provided gender-dependent score distributions that were closely aligned and therefore resulted in similar calibration model parameters.

	SWCELLP1		SWPH2	
Cal\Test	Female	Male	Female	Male
Female	0.0	0.7	0.0	-0.6
Male	-0.6	0.0	1.5	0.0

Figure 3: *C_{loss}(%) when calibrating with matched or mismatched gender scores with respect to the evaluation set. Results are for both of the Switchboard datasets.*

4.3. Distance Analysis

Figure 4 shows the results when varying the distance to the microphone, for the SRI Distant Speech Collect Dataset for two microphone types: lapel (top) and studio (bottom). In this case, we see a large degradation when the distance to the microphone is significantly different between calibration and evaluation conditions. For this analysis, we kept the room acoustics constant by varying distance in the same room.

Cal\Test	76 cm	94 cm	140 cm	150 cm	155 cm	178 cm	259 cm
76 cm	0.0	17.4	-10.4	12.3	29.7	44.3	228.8
94 cm	2.6	0.0	7.0	-1.5	2.1	6.4	97.1
140 cm	20.3	37.7	0.0	32.4	58.1	72.3	299.9
150 cm	-4.9	3.4	-5.6	0.0	8.5	17.9	144.0
155 cm	-3.9	0.7	4.0	-2.2	0.0	7.9	103.0
178 cm	19.0	4.4	34.9	4.6	-0.4	0.0	49.0
259 cm	122.9	74.5	181.9	77.2	46.6	37.6	0.0

(a) *Lapel microphone*

Cal\Test	76 cm	140 cm	178 cm	259 cm
76 cm	0.0	11.5	154.2	225.5
140 cm	0.3	0.0	90.4	141.2
178 cm	73.3	13.4	0.0	0.7
259 cm	72.2	11.7	1.3	0.0

(b) *Studio microphone*

Figure 4: *C_{loss}(%) for distance variation between calibration and evaluation sets on the SRI Distant Speech Collect Dataset.*

The asymmetry in Figure 4 may be attributed to the reverberation variation (early and late reflections) in signals captured by microphones placed at different distances. The calibration models learned with cleaner data (i.e. mic closer to the source) are prone to bad generalization since its very hard to estimate the bias as there is a very little overlap between the tails of the score distribution. Hence, it is always better to calibrate with same or farther placed microphone than the evaluation set.

4.4. Language Analysis

Figure 5 shows the C_{loss} for different languages used in calibration and evaluation. In this case, we see that using very closely related languages—such as Farsi, Dari, and Pashto—result in only minimal calibration loss (fewer than five points) when used to calibrate the other. Languages with significant shared vocabulary—like Arabic and Urdu—also show a relatively small loss in calibration. The greatest loss in calibration comes from using distantly related or unrelated languages in the calibration and evaluation sets, such as Farsi, Dari, or Pashto to calibrate Arabic or Urdu.

Cal\Test	Alv	Urd	Fas	Prs	Pus
Alv	0.0	-1.8	33.7	51.5	66.2
Urd	7.5	0.0	5.6	17.0	21.0
Fas	49.7	37.6	0.0	2.3	1.8
Prs	65.7	48.5	-1.6	0.0	0.5
Pus	35.2	21.2	-6.8	-2.8	0.0

Figure 5: *C_{loss}(%) for language variation on DARPA RATS clean data. The languages are Alv=Levantine Arabic; Urd=Urdu; Fas=Farsi; Prs=Dari; and Pus=Pashto.*

4.5. Cross-Set Analysis

Finally, Figure 6 shows the cross-set results for a subset of the test sets. The sets are sorted by their performance when calibrated with the SRE10 close-talking microphone data. Surprisingly, FVC has the highest C_{loss} when the calibration model is trained with data that is similar by definition. Therefore, the calibration loss must be a consequence a nuisance factor not typically considered as part of dataset characteristics. The accent is not the major contributing factor as the FVC-noi, FVC-cod, and FVC-rev datasets were derived from the same clean data, and these sets offer a considerably lower C_{loss} *after* being perturbed from the source audio. While we do not yet understand why this corpus in its original form is an outlier, our hypothesis is the mismatch between our system training data and FVC-int data. In addition, with both enroll and test samples being collected in the same manner, our system is unable to suppress the condition variation of this dataset and instead treats it as speaker information. The consequence of this enroll-to-test match and their mismatch to our training set is that all scores, both the impostor and target scores, have a large positive bias, shifting both distributions to the right (as depicted in Figure 1) resulting in a large calibration loss.

Cal\Test	SRE10	SWCELLP1	SWPH2	FVC-noi	FVC-rev	FVC-cod	RATS-G	FVC-int
SRE10	0.0	8.5	10.5	19.7	26.9	103.7	107.7	359.4
SWCELLP1	-7.5	0.0	10.6	22.5	36.6	114.5	81.6	353.6
SWPH2	4.2	22.3	0.0	-0.8	0.9	13.9	49.3	138.0
FVC-noi	-0.3	14.1	-1.6	0.0	1.7	33.6	64.7	195.3
FVC-rev	8.6	26.5	2.4	1.1	0.0	18.7	65.3	162.8
FVC-cod	24.8	49.9	14.0	9.4	7.8	0.0	56.3	101.6
RATS-G	49.5	70.4	57.1	57.3	101.8	95.2	0.0	113.0
FVC-int	110.9	165.3	103.9	96.4	120.5	25.9	49.0	0.0

Figure 6: *C_{loss}(%) when varying calibration and evaluations sets to invoke condition mismatch.*

5. Conclusions

In this work, we assessed the impact of critical metadata factors that should be considered when calibrating a speaker recognition system. We observed a large degradation when the duration used for calibration were significantly different from those in the evaluation set and, surprisingly, no significant degradation when a different gender was used for calibration than for evaluation. A large degradation was observed when microphone distance was significantly different between the sets, and a small loss was seen for closely-related languages and languages with a shared vocabulary. We believe this study will provide a basis for practical speaker recognition system design and will be beneficial to forensic analysts for relevant population selection.

6. Acknowledgements

The research by authors at SRI International was funded through a development contract with Sandia National Laboratories (SNL) (Subcontract#1758993/ DO 1872160). The views herein are those of the authors and do not necessarily represent the views of the funding agencies.

7. References

[1] J. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[3] L. Ferrer, M. K. Nandwana, M. McLaren, D. Castan, and A. Lawson, "Toward fail-safe speaker recognition: Trial-Based Calibration with a reject option," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 140–153, 2019.

[4] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 331–355, 2006.

[5] G. S. Morrison, F. Ochoa, and T. Thiruvaran, "Database selection for forensic voice comparison," *Speaker Odyssey*, pp. 62–77, 2012.

[6] V. Hughes and P. Foulkes, "The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age," *Speech Communication*, vol. 66, pp. 218–230, 2015.

[7] ——, "What is the relevant population? considerations for the computation of likelihood ratios in forensic voice comparison," *Proc. Interspeech*, pp. 3772–3776, 2017.

[8] 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.

[9] F. Kelly and J. H. Hansen, "Score-aging calibration for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2414–2424, 2016.

[10] M. K. Nandwana, M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "Analysis and mitigation of vocal effort variations in speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6001–6005, 2019.

[11] F. Kelly and J. Hansen, "Evaluation and calibration of Lombard effects in speaker verification," *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 205–209, 2016.

[12] F. Kelly and J. H. Hansen, "Detection and calibration of whisper for speaker recognition," *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 1060–1065, 2018.

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

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[15] M. McLaren, D. Castan, M. K. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," *Speaker Odyssey*, pp. 327–334, 2018.

[16] M. K. Nandwana, J. van Hout, M. McLaren, A. Stauffer, C. Richey, A. Lawson, and M. Graciarena, "Robust speaker recognition from distant speech under real reverberant environments using speaker embeddings," *Proc. Interspeech*, pp. 1106–1110, 2018.

[17] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *Proc. IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

[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," *URL: http://databases. forensic-voice-comparison. net*, 2015.

[19] K. Walker and S. Strassel, "The RATS radio traffic collection system," *Speaker Odyssey*, pp. 291–297, 2012.