Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

EXPLORING THE EFFECTS OF DEVICE VARIABILITY ON FORENSIC SPEAKER COMPARISON USING VOCALISE AND NFI-FRIDA, A FORENSICALLY REALISTIC DATABASE

David van der Vloed[1], Finnian Kelly[2], Anil Alexander[2]

[1]Netherlands Forensic Institute, The Hague, The Netherlands,
[2]Oxford Wave Research, Oxford, United Kingdom

d.van.der.vloed@nfi.nl, {finnian|anil}@oxfordwaveresearch.com

ABSTRACT

In this paper we present NFI-FRIDA (Netherlands Forensic Institute - Forensically Realistic Inter-Device Audio), a database of speech recordings acquired simultaneously by multiple forensically-relevant recording devices, and demonstrate how this database can be used to support forensic speaker comparison (FSC) casework. We use VOCALISE (Voice Comparison and Analysis of the Likelihood of Speech Evidence), an x-vector based automatic speaker recognition system that allows a forensic practitioner to perform speaker comparisons in a flexible way. After establishing how variability of the recording device affects speaker recognition discrimination performance, we explore how variability of the recording device of the relevant population affects the resulting likelihood ratios. These experiments demonstrate a research methodology for how a forensic practitioner can corroborate their subjective judgment of the 'representativeness' of the relevant population in FSC casework.

1. INTRODUCTION

In this paper the Netherlands Forensic Institute's Forensically Realistic Inter-Device Audio Database (NFI-FRIDA) is presented. This 250 speaker database contains 333 hours of speech in 1500 hours of simultaneous recordings from multiple recording devices. The database is designed to serve as relevant population data for NFI forensic speaker comparison (FSC) casework, and to enable research into the relationship between different recording devices and the output of automatic speaker recognition systems.

The automatic speaker recognition experiments in this paper are conducted using VOCALISE (Voice Comparison and Analysis of the Likelihood of Speech Evidence) [1], an automatic speaker recognition system that allows a forensic practitioner to perform speaker comparisons in a flexible way. VOCALISE supports x-vector and i-vector speaker recognition frameworks, with PLDA (Probabilistic Linear Discriminant Analysis) and Cosine Distance scoring, as well as classic GMM (Gaussian Mixture Modelling), with and without MAP (Maximum a Posteriori) adaptation. All of these approaches can be applied to both spectral MFCC (Mel frequency cepstral coefficient) and 'auto-phonetic' (automatically-extracted phonetic) features [2].

VOCALISE is provided with pre-trained models, which can be fully or partially re-trained by the user. Additionally, VOCALISE supports both model- and score-based condition adaptation with small amounts of user-provided data; this provides a practical way for the practitioner to adapt the system towards the conditions of a specific case.

In this paper, a VOCALISE x-vector PLDA system is used to explore the extent to which recording device mismatches influence automatic speaker recognition output. The results were used to assess the choices a practitioner of FSC can make in different real cases. This type of research can guide the selection of relevant population data that is representative of case data, by informing the practitioner if a specific device mismatch must be accounted for, or can be disregarded. This process allows the practitioner to substantiate their judgment of the 'representativeness' of the relevant population, which is one of the most prominently subjective decisions to be made when employing automatic speaker recognition in FSC. In this study we investigate the effect of device variability on speaker recognition performance and show how that can be used as a controlled evaluation of representativeness in a forensic speaker comparison context.

2. DESCRIPTION OF THE DATABASE

2.1. Speaker demographics

All 250 speakers were males who were not university educated, and about 80% of them were between 18 and 35 years of age, and 20% older, up to 55 years of age. 50% of speakers were from a Native Dutch background, 25% were from a Moroccan immigrant background and 25% were from a Turkish immigrant background. The majority of speakers were born and raised in Amsterdam, and all speech collected in this database is in Dutch, often including colloquialisms and street language. The speakers with an immigrant background speak a variety of Dutch that is associated with immigrant groups, and the speakers with a native Dutch background speak Amsterdam Dutch. These speaker demographics were chosen because they are relevant to casework at the NFI. Furthermore, linguistic data sets with these speaker demographics are rare, making NFI-FRIDA a valuable resource from a sociolinguistic perspective.

2.2. Recording devices

The speech was simultaneously recorded with multiple devices. Depending on the session type (see Section 2.3) the recordings were made with three or six devices, see Table 1. The devices were chosen to reflect conditions encountered in NFI casework.

10.21437/Odyssey.2020-57

Table 1. *Recording devices*

	Recording device	Used in session
d1	Shure WH20 HQ Headset	1,2,3,4,5,6,7,8
d2	Shure SM58 close	1,2,3,4
d3	AKG C400BL close	1,2,3,4
d4	Shure SM58 far	1,2,3,4
d5	Intercepted telephone	1,2,3,4,5,6,7,8
d6	Video by iPhone 4	1,2,3,4,5,6,7,8

Recording device 1 (d1) was worn by the speaker, and provided a high quality recording. Recording devices 2 and 3 (d2 and d3) were positioned on the table in front of the speaker and represent the higher quality police interview recordings. Recording device 4 (d4, the same hardware as recording device 2) was set up on the other side of the room, at about a 3 metre distance from the speaker. Recordings from this device contain considerable reverberation and have a higher noise level than the counterpart device 2. Recordings from device 4 represent lower quality police interview recordings. The intercepted telephone recordings (d5), made possible through the kind cooperation of Dutch police, are recordings that went through the telephone interception system that is used in actual criminal investigations. Either an iPhone 4 or a Nokia 1280 telephone was used, according to the session (see Section 2.3). The iPhone 4 was chosen as it was the most widely used smartphone at the time of recording, and the Nokia 1280 was chosen to represent a cheap 'burner' phone, the type often encountered in casework. Recording device 6 (d6) represents casework material originating from smartphone video recordings. See Figure 1 for the setup used in the indoor recording sessions.

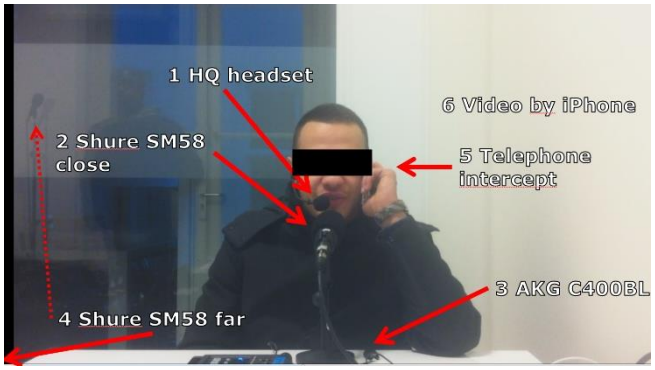

Figure 1. *Screen capture of the smartphone video playback (device 6) showing all other devices for indoor recording sessions*

Table 2. *Sessions per day*

Session	Location	Environment	Telephone
1	Indoor	Quiet	Nokia 1280
2	Indoor	Quiet	iPhone 4
3	Indoor	Noisy	Nokia 1280
4	Indoor	Noisy	iPhone 4
5	Outdoor	Calm location	Nokia 1280
6	Outdoor	Calm location	iPhone 4
7	Outdoor	Busy street	Nokia 1280
8	Outdoor	Busy street	iPhone 4

2.3. Speaker sessions

Speakers were recorded in 16 sessions divided over two days, with an interval of at least one week between the two days. The eight sessions recorded per day varied in location, the telephone used, and the noisiness of the environment, as shown in Table 2.

Each session lasted about 5 minutes. All speech recorded consisted of telephone conversations, and the participants were in all cases talking to another participant.

For indoor sessions, a noisy environment meant the presence of static radio noise, and for outside sessions, the environment alternated between a calm and a noisy street location.

2.4. Orthographic transcription

All transcriptions were made using the recordings of device 1 (the highest quality recordings). Because of limited resources, only the even-numbered sessions (those in which an iPhone was used) were transcribed. The transcriptions are orthographic, using a transcription protocol based on [3], adapted for this specific database.

Three native speakers of Dutch were involved in producing each transcription: one made the first transcription and each of the two others checked the transcription. They swapped roles for each speaker pair. The transcriptions were made using Praat [4]. Praat TextGrids were initialized using Praat speech detection (via the command 'To Textgrid (silences)'), producing an interval per speech utterance per speaker. If necessary, the transcribers could adjust these intervals. The final transcriptions therefore provide start and end time information per utterance.

The orthographic transcriptions can be used to train and test models for speech-to-text systems. They also provide speech-non-speech-information, making it possible to create edits of the sound files that only contain speech. The transcriptions were made using recordings of device 1, but due to the simultaneous

nature of the recordings, are equally applicable to the recordings from the other devices. Consequently, the resulting transcriptions for the lower quality recordings are likely to be of better quality than could have been achieved by transcribing those lower quality recordings directly.

2.5. Aim of the database

The database is primarily collected to be used in FSC casework using automatic speaker recognition. In the case method that is described in more detail in Section 4, it is necessary to use relevant population data representative of the case conditions, both for obtaining performance measures for an automatic speaker recognition system in those specific conditions, and to generate likelihood ratios (LRs).

Furthermore, the simultaneous multi-device recording design allows for testing the performance and robustness of automatic speaker recognition systems to different recording devices, since the recordings of the same session only differ in recording device. This allows for research into the boundaries of the 'representativeness' of the relevant population (an example of which will be given in Section 4.2). The presence of higher quality recording devices along with lower quality recording devices also makes it possible to explore how well artificially degrading higher quality recordings to other quality recordings can represent reality (e.g. microphone to telephone). Finally, the data along with the orthographic transcriptions can be used to perform research into within-speaker and between-speaker variation of various phonetic variables, strengthening auditory-acoustic-phonetic methods of forensic speaker comparison (as in [5], for example).

3. VOCALISE

For the experiments in this paper, we use the pre-trained VOCALISE 2019A x-vector PLDA session[1] based on MFCC features. In this session, 22-dimensional MFCCs (including energy) are extracted over 25 ms Hamming windows with a 10 ms overlap, using 23 Mel filterbanks in the range 20 to 3,700 Hz. CMS (cepstral mean subtraction) is applied over a sliding window of 3 seconds, and silence frames are dropped according to VAD.

The session is trained with a diverse set of speech recordings from several thousand speakers; the training set contains various channels (e.g. telephone and microphone) and multiple languages. To expand the quantity and diversity of the training data, 'data augmentation' [6, 7] is applied. In this procedure, copies of the training recordings are augmented with noise and reverberation, before being combined with the original training set. The DNN architecture was the same as that in [6], with the x-vector speaker embedding taken at the output of the 512-dimensional seventh layer. The full training set was used for the DNN, LDA and PLDA models of 150 dimensions.

We also evaluate the effect of score-based condition adaptation, in the form of top-N symmetric score normalisation (S-norm) [8]. Generally speaking, score (or reference) normalisation is an approach to adjust comparison scores using a set of relevant reference speakers. In VOCALISE, top-N S-norm is applied to a test score by comparing each of the test files with a set of reference speakers. The highest N scores (i.e. the top N) in each set of reference scores are extracted (the value of N is set by the user). The mean and standard deviation of the top N scores are calculated, resulting in a set of normalisation statistics for each test file. Two normalised scores are then generated by separately applying (subtracting the mean and dividing by the standard deviation) each set of normalisation statistics to the test score. The final S-normed score is then given by the mean of the two normalised scores.

Calibration was applied to the VOCALISE comparison scores using the accompanying Bio-Metrics performance metrics software[2]. Bio-Metrics uses a linear logistic regression [9] procedure for calibration.

3.1. Performance of VOCALISE using NFI-FRIDA

3.1.1. Data used

Recordings of 135 NFI-FRIDA speakers from session 2 (inside, quiet, iPhone 4) from both days were used for experiments in this paper. Of those sessions, recordings from all devices except the smartphone video were used, resulting in two different speech sessions per speaker and in ten recordings per speaker. The edited versions (i.e. after extraction of speech using the transcriptions) of the recordings were used, and were further reduced to 40 seconds of net speech (this was the net speech duration of the shortest recording).

The recordings of 90 speakers were used as test recordings, with the recordings of the remaining 45 speakers used as a cohort for reference normalisation (S-norm). The top-N value was set to 45, and therefore the full normalisation cohort was always used to calculate normalisation statistics.

3.1.2. Comparisons using VOCALISE

All comparisons were performed using VOCALISE 2019A x-vector PLDA session. All comparisons were done twice: once without any reference normalisation and once with reference normalisation. Since all the recordings were edited, the voice activity detection (VAD) option was disabled.

All test recordings from each of the five devices were compared with all test recordings from each of the five devices, while making sure that comparisons were always between recordings from different days, to avoid comparing different recordings of the exact same event. This resulted in 90 same-speaker trials and 8010 different-speaker trials for the experiments with matching recording devices. It is twice that number for the experiments with mismatching recording devices, since in that case recording device A from day 1 can be compared with recording device B from day 2 and vice versa.

[1] A VOCALISE 'session' contains all of the trained models and settings required to carry out a speaker comparison.

[2] Bio-Metrics 1.8 performance metrics software, Oxford Wave Research Ltd., https://www.oxfordwaveresearch.com/products/bio-metrics, accessed April 8th 2020

3.1.3. Results

Table 3 shows the resulting system performance in terms of convex hull EER (as calculated by Bio-Metrics), with and without reference normalisation.

Table 3. *EER% without and with reference normalisation ('no RN' / 'with RN').*

		d1	d2	d3	d4	d5
d1	no RN	1.53	1.68	1.49	3.41	2.75
	with RN	1.20	1.22	1.15	2.28	2.70
d2	no RN		1.40	1.55	3.12	2.92
	with RN		1.20	1.19	2.61	2.93
d3	no RN			1.08	2.62	2.88
	with RN			0.92	2.28	2.97
d4	no RN				2.80	4.72
	with RN				2.42	4.29
d5	no RN					2.74
	with RN					2.13

3.2. Discussion of results

We observe good baseline performance from the system; for all but one comparison (d4 vs d5), EERs are less than 4%. After applying reference normalisation we see further improvement in EERs for the majority of comparisons. As expected, the high quality devices lead to the lowest EERs; the within and cross-device comparisons of d1, d2 and d3 are all close to 1% after reference normalisation. The lower quality of d4 and d5 are reflected in the higher EERs, particularly in the cross-device comparison of these two devices.

This discrimination performance may satisfy the forensic practitioner that this system is appropriate to use for a matched or mismatched condition case involving any of these devices. Our subsequent experiments explore the importance of the device in the selection of a relevant population in the context of FSC casework.

4. USING AUTOMATIC SPEAKER RECOGNITION IN FORENSIC SPEAKER COMPARISON CASEWORK

In FSC the practitioner is asked to compare two recordings in the context of a criminal case. First, there is a questioned recording, in which a speaker with unknown identity is speaking (e.g. the offender), and a known recording, in which a known speaker is speaking (e.g. the suspect). The hypotheses to be considered are typically (see [10]):

H0: "the suspected speaker recording and the questioned recording have the same source" and
H1: "the suspected speaker recording and the questioned recording have different sources".

An example case would be the comparison of a telephone intercept as questioned material, spoken in Dutch with 30 seconds of net speech and a police interview recording as known material, spoken in Dutch with 180 seconds of net speech.

The practitioner uses an automatic speaker recognition system to compare the case material to produce a case score. Although this case score may be produced as a ratio of likelihoods by the system, it is not yet a forensic likelihood ratio, in the sense that it cannot directly be interpreted as evidential value in the light of the two hypotheses.

Next, the practitioner selects relevant population material that is representative of the case material. The first part of this relevant population material should be representative of the questioned recording, and therefore it would consist of intercepted telephone recordings with 30 seconds of spoken Dutch. The second part should be representative of the known recording, and therefore it would consist of police interviews with 180 seconds of spoken Dutch. If there are overlapping speakers in those two sets, the relevant population material will contain both same-speaker pairs and different-speaker pairs. The practitioner then proceeds to compare all those pairs from the relevant population material using the exact same automatic speaker recognition system and settings as used with the case comparison. The resulting same-speaker and different-speaker scores can then be used to evaluate the case score under the competing same-speaker and different-speaker hypotheses. An LR can be calculated by modelling the same-source score distribution (H0) and the different-source score distribution (H1) as two probability density functions, and then dividing the value of the case score given the H0 hypothesis by the value of the case score given the H1 hypothesis. Alternatively, the two distributions can be used to establish a score to LR function by training coefficients with which to scale and shift the scores, optimizing for Cllr (log likelihood ratio cost). If the Cllr is low, the warped scores can be interpreted as LRs with good calibration. In either case, the performance of the method as a whole can be assessed at this stage, and the practitioner can decide whether the method is good enough to proceed. Finally, if these criteria are satisfied, the calculated LR can be used as the result of the automatic speaker recognition analysis in the FSC case.

Note that the above description is the core of a method involving automatic speaker recognition in FSC; steps to establish whether the case should be done at all, preprocessing steps and additional steps that test the validity of the system are left out for brevity.

4.1. Judging representativeness

One of the steps in the case method in the previous section relies almost entirely on the informed personal judgment of the practitioner: deciding whether relevant population material is

representative of the case material. In principle, representativeness means that the relevant population comparison pairs are in the same conditions as the case comparison pair. This applies to speaker conditions (spoken language, gender of the speaker, level of vocal effort, etc.) and to recording conditions (microphone, distance to microphone, room acoustics, telephone codecs, recording durations, etc.). As the variation with which real case material can present itself is near endless, the list of conditions to be considered when selecting relevant population material is also near endless. Furthermore, as the provenance of case material is not always completely clear, some of the conditions in the case material may be unknown, further complicating the selection of relevant population material.

The practitioner may have to make a practical decision to disregard some of the conditions (which may or may not be relevant) and proceed with the metadata that is known. If the practitioner tries to match on all the specific conditions that exist within the case, it may be difficult or indeed impossible to source appropriate relevant population data. For instance, in the example case above, the practitioner may have a dataset that has both intercepted telephone speech and police interviews from the same speakers. However, if the practitioner wants to further select only those recordings that match the exact age of the known speaker, the exact language variety and the exact type of microphone and distance to the microphone, it is not hard to see that the practitioner will have insufficient data to do any meaningful comparisons very quickly.

One potential solution is by training a score to LR function that encompasses a mismatch between case material and relevant population material. See [11] for an example involving a large time interval between questioned and known material. Another approach would be to systematically estimate the robustness of automatic speaker recognition and the calculation of the likelihood ratio for differing database conditions. If it can be shown that the exact setting of some condition has little or no impact on the results of an automatic speaker recognition system, that condition can be disregarded when selecting relevant population material, relieving some of the data needs.

4.2. Experiment into representativeness using NFI-FRIDA.

As NFI-FRIDA contains simultaneous recordings only differing in recording device, it can be used to investigate how robust automatic speaker recognition performance is for those different devices. An example experiment is shown below, in which the aim is to chart which conditions can be disregarded and which need to be considered when selecting relevant population data.

4.2.1. Data used

Recordings of 135 NFI-FRIDA speakers from session 2 (inside, quiet, iPhone 4) from both days were used. Of those sessions, all devices except the smartphone video were used, resulting in two different speech sessions per speaker and in ten recordings per speaker. All recordings were edited speech and were further reduced to 40 seconds of net speech. 45 of the 135 speakers were used to create mock cases.

All mock cases consisted of a matched-device comparison; there were therefore five types, one for each device. This resulted in five sets of 45 same-speaker trials and 1980 different-speaker trials.

The remaining 90 speakers were used as relevant population data as described in Section 4.1, again in five matched device types: resulting in five sets of 90 same-speaker trials and 8010 different-speaker trials.

4.2.2. Comparisons using VOCALISE

The mock cases were compared using VOCALISE (see Section 3). No reference normalisation was applied, and since all the recordings were edited, the voice activity detection option was disabled. Next, the relevant population data was compared using VOCALISE with the same settings as the mock cases. This procedure was repeated three times: each time using a different set of 45 speakers as mock case speakers and the remaining 90 as relevant population speakers, such that all 135 speakers served as a mock case speaker once. Using each of the five sets of relevant population data, linear logistic regression was applied to each of the five sets of mock cases using Bio-Metrics. For every mock case type, this yielded a 'correct' set of LRs, meaning that matched-device relevant population data was used to calibrate the case comparisons, and four sets of 'incorrect' LRs, in which there was a device mismatch between relevant population data and mock case data.

4.2.3. Results

The results for the three repetitions were pooled and what resulted were 25 sets of log LRs, one set for each mock case type calibrated by a relevant population data type. An example of 3 sets of log LRs are given in Figure 2, which show Tippett plots for case type d1-d1, when calibrated with relevant population data types d1-d1 (green, continuous line), d2-d2 (dark green, dotted line) and d5-d5 (red, dashed line). We can observe that the log LR distributions using slightly mismatched relevant population data (d2-d2, dark green) are significantly closer to the log LRs obtained using correctly chosen relevant population data (d1-d1, green) than when using the relevant population data with the greater mismatch (d5-d5, red).

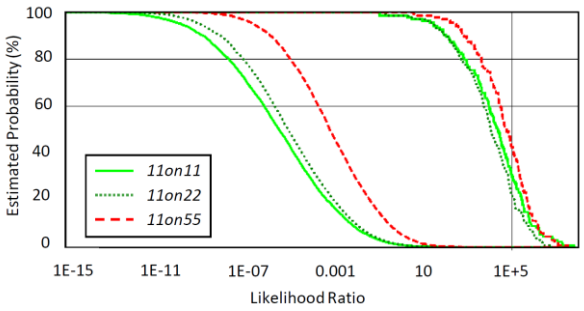

Figure 2. *Tippett plots*
(11on11: mock case type d1-d1, rel. pop. data d1-d1;
11on22: mock case type d1-d1, rel. pop. data d2-d2;
11on55: mock case type d1-d1, rel. pop. data d5-d5)

For each of the sets the Cllr was calculated and is shown in Table 4. The Cllr values can function as a measure of which relevant population data is best used for each mock case type.

Table 4. *Cllr values*

Mock case type:	Calibrated with relevant population data:				
	d1-d1	d2-d2	d3-d3	d4-d4	d5-d5
d1-d1	**0.049**	**0.045**	**0.047**	**0.061**	**0.107**
d2-d2	**0.049**	**0.049**	**0.049**	**0.068**	**0.120**
d3-d3	**0.062**	**0.060**	**0.054**	**0.061**	**0.096**
d4-d4	**0.168**	**0.165**	**0.137**	**0.114**	**0.126**
d5-d5	**0.282**	**0.262**	**0.220**	**0.154**	**0.114**

For each horizontal line in Table 4, the Cllr values can be compared for a relevant population data type. The diagonal from the top left to the bottom right shows the Cllr values if the 'correct' relevant population data is chosen, i.e. data recorded with the same device as the mock case. All other off-diagonal values show the Cllr when there is an 'incorrect' choice of mismatched-device relevant population data. The difference between the Cllrs for incorrect and incorrect selections are a measure of loss due to device mismatch. As can be seen in the first three lines of Table 4 (the three case types using close, direct microphones) choosing any of the three close direct microphones as relevant population data works well. Interestingly, the far microphone (d4-d4) is not much worse as relevant population data in those cases. It is also clear that choosing telephone intercepts (d5-d5) as relevant population data for the close direct microphone cases is the worst option. For cases involving the far microphone (d4-d4) and the telephone intercepts (d5-d5), on the 4th and 5th line of the table, the matching relevant population data performs the best.

4.2.4. Discussion

When the practitioner has to choose relevant population data based on the case data, these results show that it is best to choose matching data with regard to device type. For instance, if the case data consists of telephone intercepts, the practitioner should choose telephone intercepts as relevant population data. It is not surprising that recordings from the same recording device represent each other better than recordings from other devices. However, these results also suggest that, when the case data consists of high quality direct microphones, it is not necessary to exactly replicate the type of microphone for the relevant population data. A surprising result is that for these cases using relevant population data from a far microphone is not that detrimental – suggesting distance to microphone is a variable that is of less importance when choosing relevant population data.

5. CONCLUSION

The NFI-FRIDA database has been collected for validation research of automatic speaker recognition for use in forensic speaker comparison. With its simultaneous recordings of different recording devices it is particularly helpful for researching the performance of automatic speaker recognition under different recording circumstances. The low EERs achieved with VOCALISE on NFI-FRIDA support the use of automatic speaker recognition in FSC even for mismatched-device comparisons. An important consideration for the forensic practitioner is the choice of a relevant population for a case. Our experiments show that while a matched-device relevant population is the best choice, the practitioner may have grounds to disregard the exact type of microphone when dealing with high quality direct microphone recordings. These experiments have demonstrated a process for evaluating the representativeness of relevant population data, which is of central importance to the practitioner in real forensic casework.

6. REFERENCES

1. F. Kelly, O. Forth, S. Kent, L. Gerlach, and A. Alexander, "Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors", Audio Engineering Society (AES) Forensics Conference 2019, Porto, Portugal, 2019.
2. A. Alexander, O. Forth, A. A. Atreya, and F. Kelly, "VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features", Odyssey 2016, Bilbao, Spain, 2016.
3. W. Goedertier and S. Goddijn, "Orthographic Transcription of the Spoken Dutch Corpus", in LREC-2000 (Second International Conference on Language Resources and Evaluation), vol II, pp 909-914, 2000.
4. P. Boersma and D. Weenink, "Praat: doing phonetics by computer", version 6.0.37, retrieved from http://www.praat.org/, 2020.
5. M. Bon, W. Heeren, and D. van der Vloed, "The speaker-dependency of features of hesitation markers in Dutch spontaneous phone conversations", in International Association for Forensic Phonetics and Acoustics conference, Huddersfield, UK, pp. 71-72, 2018.
6. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition", In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5329-5333, Calgary, Canada, 2018.
7. M. McLaren, D. Castán, M. K. Nandwana, L. Ferrer, and E. Yılmaz, "How to Train Your Speaker Embeddings Extractor", In Odyssey 2018, pp 327-334, 2018.
8. S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification", In Odyssey 2010, 2010.
9. S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST '99 1-speaker submissions", Digital Signal Processing, vol. 10, pp. 237–248, 2000.
10. A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, "Methodological Guidelines for best practice in forensic semiautomatic and automatic speaker recognition, including guidance on the conduct of proficiency testing and collaborative exercises", Wiesbaden, Germany. European Network of Forensic Science Institutes (ENFSI), 2015.
11. G. S. Morrison, and F. Kelly, "A statistical procedure to adjust for time-interval mismatch in forensic voice comparison" In Speech Communication, vol. 112, pp. 15-21, 2019.