

NON-NATIVE SPEAKER VERIFICATION FOR SPOKEN LANGUAGE ASSESSMENT

Linlin Wang[†], Yu Wang[†], Mark J. F. Gales

ALTA Institute / Engineering Department, Cambridge University, UK

ABSTRACT

Automatic spoken language assessment systems are becoming more popular in order to handle increasing interests in second language learning. One challenge for these systems is to detect malpractice. Malpractice can take a range of forms, this paper focuses on detecting when a candidate attempts to impersonate another in a speaking test. This form of malpractice is closely related to speaker verification, but applied in the specific domain of spoken language assessment. Advanced speaker verification systems, which leverage deep-learning approaches to extract speaker representations, have been successfully applied to a range of native speaker verification tasks. These systems are explored for non-native spoken English data in this paper. The data used for speaker enrolment and verification is mainly taken from the BULATS test, which assesses English language skills for business. Performance of systems trained on relatively limited amounts of BULATS data, and standard large speaker verification corpora, is compared. Experimental results on large-scale test sets with millions of trials show that the best performance is achieved by adapting the imported model to non-native data. Breakdown of impostor trials across different first languages (L1s) and grades is analysed, which shows that inter-L1 impostors are more challenging for speaker verification systems.

Index Terms— speaker verification, non-native speech

1. INTRODUCTION

Automatic spoken assessment systems are becoming increasingly popular, especially for English with the high demand around the world for learning of English as a second language [1, 2, 3, 4]. In addition to assessing a candidate's English ability such as fluency and pronunciation and giving feedback to the candidate, these automatic systems also need to ensure the integrity of the candidate's score by detecting malpractice, as shown in Figure 1. Malpractice is the action by a candidate that breaks the assessment regulation and potentially

threatens the reliability of the exam and associated certification. Malpractice can take a range of forms in spoken language assessment scenarios, such as using or trying to use unauthorised materials, impersonation, speaking irrelevant to prompts/questions, speaking in his/her first language (L1) instead of the target language for spoken tests, *etc.* This work aims to investigate the problem of automatically detecting impersonation, in which a candidate attempts to impersonate another in a speaking test. This is closely related to speaker verification.

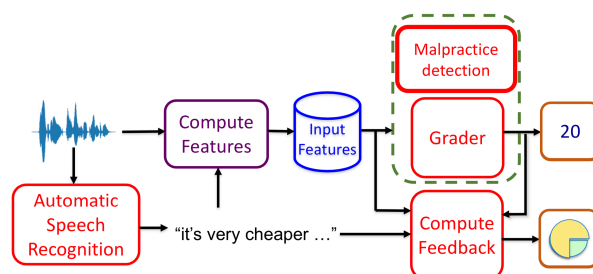


Fig. 1. Diagram of an automatic spoken language assessment system.

Speaker verification is the process to accept or reject an identity claim by comparing the speaker-specific information extracted from the verification speech with that from the enrolment speech of the claimed identity. These approaches can be directly applied to detect impersonation in spoken language tests. The performance of speaker verification systems has advanced considerably in the last decade with the development of i-vector modelling [5], in which a speech segment or a speaker is represented as a low-dimensional feature vector. Extraction of i-vectors is normally based on a Gaussian mixture model (GMM) based universal background model (UBM). This fixed length representation can then be used with a probabilistic linear discriminant analysis (PLDA) model to produce verification scores by comparing speaker representations, which are then used to make valid or impostor speaker decisions [6, 7, 8, 9]. Recently, with developments in deep learning, performance of speaker verification systems has been improved by replacing the GMM with a deep neural network (DNN) to derive statistics for extracting speaker representations. This DNN is usually trained to take a fixed length window of the acoustics and discriminate

This paper reports on research supported by Cambridge Assessment, University of Cambridge. Thanks to Cambridge English Language Assessment for support and access to the BULATS and Linguaskill data. [†] Both authors contributed equally. The authors would also like to thank Dr Kate Knill and Dr Anton Ragni for valuable discussions during the preparation of this manuscript.

between speakers using supplied speaker labels as targets. To handle the variable-length nature of the acoustic signal, a pooling layer is used to yield the final fixed-dimensional speaker representation. In [10], a DNN was trained at the frame level, and pooling was performed by averaging activation vectors of the last hidden layer over all frames of an input utterance. In [11, 12, 13], segment-level embeddings were extracted, which are referred to as x-vectors [13] with data augmentation. By leveraging data augmentation based on background noise and acoustic reverberation, these x-vectors based systems can achieve better performance than i-vector and d-vector based systems on standard speaker verification tasks.

There has been some previous work on tasks related to non-native speech data using speaker verification approaches, such as detection of non-native speech [14], classification of native/non-native English [15] and L1 detection [16]. In [17], meta-data (L1) sensitive bottleneck features were employed within the i-vector framework to improve the performance of speaker verification with non-native speech. In contrast, this paper focuses on making use of the state-of-the-art deep-learning based speaker verification approaches to detect candidate impersonation in an English speaking test. As there is limited amounts of data available for the non-native learner task, it is of interest to investigate adapting a standard speaker verification task to this non-native task. Here a system based on the VoxCeleb dataset [18, 19] is adapted to the BULATS task. Two forms of adaptation are examined: modifying the PLDA distance measure; and adapting the process for extracting the speaker representation by “fine-tuning” the network to the target domain. Furthermore, detailed analysis of performance is also done with respect to speaker attributes. Gender is an important attribute in impostor selection for standard speaker verification tasks, and for non-native speech, there are two additional speaker attributes: the L1 and the language proficiency level¹, which should also be taken into consideration for speaker verification.

This paper is organised as follows. Section 2 gives an overview of speaker verification systems, and Section 3 introduces the non-native spoken English corpora used in this work. Experimental setup is described in Section 4, results and analysis are detailed in Section 5, and finally, conclusions are drawn in Section 6.

2. SPEAKER VERIFICATION SYSTEMS

In this work both i-vector and x-vector representations are used. For the i-vector speaker representation the form described in [5, 20] is used. This section will just discuss the x-vector speaker representation as this is the form that is adapted to the non-native verification task.

¹Language ability level is referred to as “grade” in this work.

2.1. Deep neural network embedding extractor

There are three blocks to form the DNN for extracting the utterance-level speaker representation, or embedding. The first block of the deep embedding extractor is a frame-level feature extractor. The input to this block is a sequence of acoustic feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ of T frames. This part normally consists of a number of hidden layers such as long short-term memory (LSTM) [21] or time delay neural network (TDNN) layers [12, 13]. The activations of the last hidden layer of this block for the input frames, $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T\}$, form the input to the second block which is a statistics pooling layer. This layer converts variable-length frame-level features into a fixed-dimensional vector by calculating the mean vector, $\boldsymbol{\mu}$ and standard deviation vector $\boldsymbol{\sigma}$ of the frame-level feature vectors over the T frames. The third block takes the statistics as the input and produces utterance-level representations using a number of stacked fully-connected hidden layers. The output of the DNN extractor is a softmax layer, and each of the nodes corresponds to one speaker identity. This DNN extractor is trained based on a cross-entropy loss function using the supplied speaker labels to get the targets. Consider there are N training segments and S speakers, the cross-entropy can be written as

$$\mathcal{F}(\boldsymbol{\theta}) = - \sum_{n=1}^N \sum_{k=1}^K \delta(s, s_k^{(n)}) \log P(s | \mathbf{x}_{1:T}^{(n)}, \boldsymbol{\theta}), \quad (1)$$

where $\boldsymbol{\theta}$ represents the parameters of the DNN and $\delta(\cdot)$ represents the Kronecker delta function. $s_k^{(n)}$ represents that the speaker label for segment n is s_k . After the DNN is trained, the utterance-level embeddings, \mathbf{e}_d , are normally extracted from the output of the affine component that is with or without the nonlinear activation function applied of one hidden layer in the third block of the DNN [12, 13].

2.2. PLDA classifier and adaptation

After the speaker embeddings are extracted, they are used to train a PLDA model that yields the score (distance) between speaker embeddings. The training of the PLDA models aims to maximise the between-speaker difference and minimise the within-speaker variation, typically using expectation maximisation (EM). A number of variants of PLDA models have been introduced into the speaker verification task based on this “standard” PLDA [6]: two-covariance PLDA [22] and heavy-tailed PLDA [7]. The variant implemented in the Kaldi toolkit [20], and used in this work, follows [23] and is similar to the two-covariance model. This model can be written as

$$\mathbf{e} = \mathbf{y} + \mathbf{z}, \quad (2)$$

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \boldsymbol{\Gamma}), \quad (3)$$

$$p(\mathbf{e}|\mathbf{y}) = \mathcal{N}(\mathbf{e}; \mathbf{y} + \boldsymbol{\mu}, \boldsymbol{\Lambda}), \quad (4)$$

where \mathbf{e} is the speaker embedding. The vector \mathbf{y} represents the underlying speaker vector and $\boldsymbol{\mu}$ represents its mean. \mathbf{z}

is the Gaussian noise vector. For speaker verification tasks, estimation of this PLDA model can be performed by estimating the between-speaker covariance matrix, Γ , and within-speaker covariance matrix, Λ , using the EM algorithm.

PLDA is a powerful approach to classifying speakers given a large amounts of training data with speaker labels [24, 25, 26]. However, large amounts of labelled training data may not be available in the domain of interest such as the one considered in this paper, the non-native speaker verification. One approach to alleviate this problem is to do adaptation from a pre-trained out-of-domain model to the target domain. There are a number of methods for adapting the PLDA model in both supervised and unsupervised manners [27, 26]. The Kaldi toolkit implements an unsupervised adaptation method which does not require knowledge of speaker labels [20]. This method aims at adapting Γ and Λ of the out-of-domain PLDA model to better match the total covariance of the in-domain adaptation data.

3. NON-NATIVE SPOKEN ENGLISH CORPORA

The Business Language Testing Service (BULATS) test of Cambridge Assessment English [28] is a multi-level computer-based English test. It consists of read speech and free-speaking components, with the candidate responding to prompts. The BULATS spoken test has five sections, all with materials appropriate to business scenarios. The first section (A) contains eight questions about the candidate and their work. The second section (B) is a read-aloud section in which the candidates are asked to read eight sentences. The last three sections (C, D and E) have longer utterances of spontaneous speech elicited by prompts. In section C the candidates are asked to talk for one minute about a prompted business related topic. In section D, the candidate has one minute to describe a business situation illustrated in graphs or charts, such as pie or bar charts. The prompt for section E asks the candidate to imagine they are in a specific conversation and to respond to questions they may be asked in that situation (e.g. advice about planning a conference). This section is made up of 5x 20 seconds responses.

Each section is scored between 0 and 6; the overall score is therefore between 0 and 30. This score is then mapped into Common European Framework of Reference (CEFR) [29] language proficiency levels, which is an international standard for describing language ability on a six-level scale. Each candidate is finally assigned a “grade”, ranging from minimal (A1) and basic (A2) command, through limited but effective (B1) and generally effective (B2) command, to good operational (C1) and fully operational (C2) command of the spoken language.

In this work, non-native speech from the BULATS test is used as both training and test data for the speaker verification systems. To investigate how the systems generalise, data for testing is also taken from the Cambridge Assessment English

Linguaskill² online test. Like BULATS, this is also a multi-level test and has a similar format composed of the same five sections as described before but assesses general English ability.

4. EXPERIMENTAL SETUP

A set of 8,480 candidates from BULATS was used for training. The approximately 280 hours of speech covers a wide range of more than 70 different L1s. There are 15 major L1s with more than 100 candidates for each, including Tamil, Gujarati, Hindi, Telugu, Malayalam, Bengali, Spanish, Russian, Kannada, Portuguese, French, *etc.* Data augmentation was applied to the training set, and each recording was processed with a randomly selected source from “babble”, “music”, “noise” and “reverb” [13], which roughly doubled the size of the original training set. Another set of 8,318 BULATS candidates was used as one test set to evaluate the system performance. There are 7 major L1s in this set, each of which has more than 100 candidates: Spanish, Thai, Tamil, Arabic, Vietnamese, Polish and Dutch. There are no overlapping candidates between the BULATS training and test sets. The other test set of 2,540 candidates came from the Linguaskill test, of which there are 6 major L1s each with more than 100 candidates: Hindi, Portuguese, Japanese, Spanish, Thai and Vietnamese. Each of the training set and two test sets was fairly gender balanced, with approximately one third of candidates graded as B1, one third graded as B2, and the rest graded as A1, A2, C1, or C2, according to CEFR ability levels. For each test set candidate, responses from sections A and B were used for speaker enrolment (approximately 180s), while the more challenging free-speaking sections C, D, and E were used for whole section-level verification (approximately 60s for each section).

5. EXPERIMENTAL RESULTS

5.1. Baseline system performance

Gender is generally considered an important speaker attribute, and impostor trials were first selected from the same gender group as the reference speaker, as commonly done in standard speaker verification tasks. This resulted in a total of 104.8 million verification trials for the BULATS test set and 9.7 million trials for the Linguaskill test set.

An i-vector/PLDA system and an x-vector/PLDA system were first trained on the “in-domain” BULATS training set. For the i-vector system, 13-dimensional perceptual linear predictive (PLP) features were extracted using the HTK toolkit [30] with a frame-length of 25ms. A UBM of 2,048 mixture components was first trained with full-covariance matrices, and then 600-dimensional i-vectors were extracted

²<https://www.cambridgeenglish.org/exams-and-tests/linguaskill/>

for both training and test sets. For the x-vector system, 40-dimensional filterbank features were also extracted using HTK with a frame-length of 25ms. DNN configurations were the same as used in [13], and 512-dimensional x-vectors were extracted from the affine component of the segment-level layer immediately following the statistics pooling layer.

Performance of the two baseline systems is shown in Table 1 in terms of equal error rate (EER). The x-vector system yielded lower EERs on both BULATS and Linguaskill test sets.

Table 1. % EER performance of BULATS-trained baseline systems on BULATS and Linguaskill test sets.

System	BULATS	Linguaskill
BULATS i-vector/PLDA	0.69	0.72
BULATS x-vector/PLDA	0.66	0.70

In addition to the models trained on the BULATS data, it is also interesting to investigate the application of “out-of-the-box” models for standard speaker verification tasks to this non-native speaker verification task as there is limited amounts of non-native learner English data that is publicly available. In this paper, the Kaldi-released [20] VoxCeleb x-vector/PLDA system was used as imported models, which was trained on augmented VoxCeleb 1 [18] and VoxCeleb 2 [19]. There are more than 7,000 speakers in the VoxCeleb dataset with more than 2,000 hours of audio data, making it the largest publicly available speaker recognition dataset. 30 dimensional mel-frequency cepstral coefficients (MFCCs) were used as input features and system configurations were the same as the BULATS x-vector/PLDA one. It can be seen from Table 2 that these out-of-domain models gave worse performance than baseline systems trained on a far smaller amount of BULATS data due to domain mismatch. Thus, two kinds of in-domain adaptation strategies were explored to make use of the BULATS training set: PLDA adaptation and x-vector extractor fine-tuning. For PLDA adaptation, x-vectors of the BULATS training set were first extracted using the VoxCeleb-trained x-vector extractor, and then employed to adapt the VoxCeleb-trained PLDA model with their mean and variance. For x-vector extractor fine-tuning, with all other layers of the VoxCeleb-trained model kept still, the output layer was re-initialised using the BULATS training set with the number of targets adjusted accordingly, and then all layers were fine-tuned on the BULATS training set. Here the PLDA adaptation system is referred to as **X1** and the extractor fine-tuning system is referred to as **X2**. Both adaptation approaches can yield good performance gains as can be seen from Table 2. PLDA adaptation is a straightforward yet effective way, while the system with x-vector extractor fine-tuning gave slightly lower EERs on both BULATS and Linguaskill test sets by virtue of a relatively “in-domain” extractor prior to the PLDA back-end.

Table 2. % EER performance of VoxCeleb-based systems on BULATS and Linguaskill test sets.

System	BULATS	Linguaskill
VoxCeleb x-vector/PLDA	0.85	1.44
+ PLDA adaptation (X1)	0.55	0.62
+ Extractor fine-tuning (X2)	0.49	0.55

Detection Error Tradeoff (DET) curves of the four x-vector/PLDA systems on the BULATS test set were illustrated in Figure 2. It can be seen that, both adaptation systems outperformed the original VoxCeleb-trained system in any threshold of the false alarm (FA) probability and the miss (MS) probability. The extractor fine-tuning system only gave higher MS probability than the PLDA adapted one with FA probability below 0.4%, while for a large range of FA probabilities above 0.4%, the extractor fine-tuning system outperformed the PLDA adapted one.

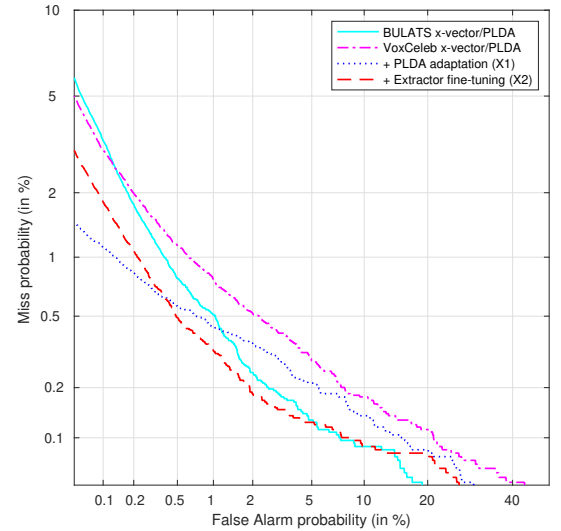


Fig. 2. DET curves of the four x-vector/PLDA systems on the BULATS test set with impostors from the same gender group as the reference speaker.

Furthermore, by leveraging the large-scale VoxCeleb dataset, both adaptation systems produced lower EERs than baseline systems solely trained on BULATS data, especially the extractor fine-tuning one, which gave a reduction rate of 26% in EER over the baseline x-vector/PLDA system on the BULATS test set. It can also be seen from Figure 2 that, the extractor fine-tuning system gave consistently better performance than the baseline systems for almost any threshold of FA and MS.

5.2. Impostor attributes analysis

As mentioned in Section 5.1, gender is an important attribute when selecting impostors. For the non-native English speech data considered in this work, there are two additional attributes that may significantly impact performance, the candidate speaking ability (grade) and L1. In this section, the impact of both attributes on verification performance is analysed on the BULATS test set using the extractor fine-tuning system (X2) detailed in Section 5.1 with impostors selected from the same gender group as the reference speaker. Taking EER as the operating threshold, both grade and L1 breakdown are investigated with respect to the number of impostor trials resulting in false alarm (FA) errors.

As there were only a small number of speakers graded as C1 or C2 in the BULATS test set, the two grade groups were merged into one group as C in the following analysis. Also for a fair comparison, 200 speakers were randomly selected (roughly gender balanced) for each grade group from the BULATS test set, and the grade breakdown is shown in Table 3. For lower grades, impostor trials from the grade group of A1 dominated FA errors as A1 speakers tend to speak short utterances, which is more challenging for the systems. For higher grades (B2 and C), impostor trials from the grade group of C constituted a larger portion of FA errors probably due to the fact that C speakers tend to speak long utterances in a more “native” way and they are also similar to B2 speakers.

Table 3. Grade breakdown of the percentage of impostor trials with FA errors at the operating threshold of EER for the extractor fine-tuning system on a subset of the BULATS test set.

Grade Ref. Spkr.	Grade of Impostor Spkr.				
	A1	A2	B1	B2	C
A1	65.8	27.5	5.8	0.3	0.6
A2	60.9	29.9	7.1	0.9	1.3
B1	46.5	26.8	13.1	7.6	5.9
B2	11.4	11.9	19.2	25.9	31.7
C	17.7	12.0	10.3	24.3	35.6

The numbers of speakers from different L1 groups also varied in the BULATS test set. For a fair comparison, 200 speakers were randomly selected (roughly gender balanced) for each of 6 major L1s. The L1 breakdown is shown in Table 4, where impostor trials from the same L1 group as the reference speaker generally dominated FA errors. English learners from the same L1 group tend to have similar accents when speaking English, which makes them more confusable to speaker verification systems compared to learners from a different L1 group. Particularly, impostors of Thai L1 constitute a considerable portion of FA errors for each L1, as A1 and A2 speakers dominate Thai L1 in the BULATS test set, which is different from other L1s where B1 and B2 speakers dominate.

Table 4. L1 breakdown of the percentage of impostor trials with FA errors at the operating threshold of EER for the extractor fine-tuning system on a subset of the BULATS test set.

L1 Ref. Spkr.	L1 of Impostor Spkr.					
	Ara.	Pol.	Spa.	Tam.	Tha.	Vie.
Ara.	74.9	0.0	0.3	0.6	14.7	9.5
Pol.	0.0	76.9	1.3	0.3	21.6	0.0
Spa.	2.1	16.5	44.7	0.0	28.2	8.5
Tam.	0.0	1.7	0.3	62.4	33.9	1.7
Tha.	0.5	2.4	0.4	1.0	92.9	2.8
Vie.	1.2	0.1	1.3	0.6	12.7	84.0

5.3. Overall system performance

Based on the analysis in the previous section, the impact of speaker attributes beyond gender, the grade and L1, were used as additional restrictions on the impostor set selection. The following forms of impostor selection were examined:

- **gender**, impostors from the same gender group as the reference speaker, as in Section 5.1;
- **grade**, impostors from the same grade group as the reference speaker;
- **>grade**, impostors from higher grade groups than the reference speaker if the grade of the reference speaker is lower than C, otherwise from C; this case is of practical interest for impersonation in spoken language tests;
- **L1**, impostors from the same L1 group as the reference speaker;

The number of total verification trials decreases with further restriction on impostors, which is shown in Table 5. Table 6 shows the impact on EER of restricting the possible set of impostors according to gender, L1 or grade on both BULATS and Linguaskill test sets. Due to the lack of data for each L1 or grade, X1 and X2 systems that are adapted or fine-tuned on all of the BULATS training set are used for verification. As expected, restricting possible impostors according to speaker attributes yielded higher EERs as the percentage of impostors “close” to the reference speaker increased. Take **gender** as the starting point, which is the configuration used in previous experiments in Section 5.1. Further restricting the set of impostors to **L1** again increased EERs agreeing with the results shown in Table 4, similarly to **grade**. An interesting result in terms of handling impersonation is that, if the set of impostors is further restricted to **>grade**, EERs decrease compared to simply restricted to **gender**. The highest EER for both systems was achieved by restricted to **gender+L1+grade**, which indicates that all these are important speaker attributes of non-native data. The **gender+L1+>grade** case is more related to practical scenarios of impersonation, since it is more likely that a candidate

chooses a substitute from the same gender and L1 group but speak the target language better to impersonate him/herself in order to obtain a higher grade in a spoken language test.

Table 5. Number of verification trials (in millions) with different restrictions on impostors for both BULATS and Linguaskill test sets.

Restrictions	BULATS	Linguaskill
gender	104.8	9.7
+ grade	31.6	2.7
+ >grade	36.9	3.6
+ L1	44.3	2.2
+ grade	14.1	0.7
+ >grade	16.7	0.8

Table 6. % EER performance of two adapted systems with different restrictions on impostors on both BULATS and Linguaskill test sets.

Restrictions	BULATS		Linguaskill	
	X1	X2	X1	X2
gender	0.55	0.49	0.62	0.55
+ grade	0.60	0.64	0.66	0.64
+ >grade	0.45	0.49	0.55	0.49
+ L1	0.65	0.71	0.84	0.98
+ grade	0.73	0.79	0.92	1.17
+ >grade	0.62	0.68	0.79	0.87

For the impersonation scenario where the impostor trials are restricted to **gender+L1+>grade**, the DET curves for all systems including the unadapted VoxCeleb and BULATS trained systems are shown in Figure 3 for the BULATS test set. This allows the overall distribution of FA and MS errors for the aforementioned systems to be evaluated. It can be seen that, compared with the fine-tuned X2 system, the PLDA-adapted X1 system had a lower MS probability when the FA probability was low and had a higher MS probability when the FA probability was high. This implies that the X1 system tends to accept imposters as reference speakers while the X2 system tends to reject reference speakers as impostors. For malpractice candidate impersonation in spoken language tests, the X2 system may have a high cost as it may incorrectly identify malpractice in valid candidates. This would require manual checks to confirm this classification. In contrast, the X1 system may result in a lower level of security because it has a higher chance of misidentifying the candidate who is impersonating another. Based on these complementary trends, a score-level linear combination of the two systems was performed with weights of 0.7 and 0.3 for X1 and X2 systems, respectively. The combination system gave consistently better performance for a wide range of FA and MS probabilities than the aforementioned systems with an EER of 0.58% on the BULATS test set, as demonstrated in Figure 3.

The same trend was also observed at these weightings on the Linguaskill test set with an EER of 0.72% for the combination system, approximately 8% relative reduction in EER from the X1 system. Thus, the combination of the two adapted systems making use of both large-scale VoxCeleb data and in-domain BULATS data, can serve as a sensible configuration for impersonation detection in spoken language tests.

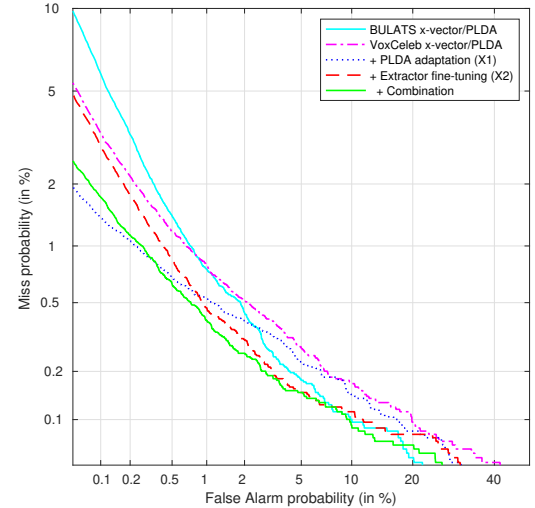


Fig. 3. DET curves of various systems on the BULATS test set with impostor trials selected from the group of the same gender, same L1 and higher grade as/than the reference speaker.

6. CONCLUSIONS

This paper has investigated malpractice in the form of candidate impersonation for spoken language assessment. This task has close relationships to standard speaker verification, but applied to the domain of non-native speech. Advanced neural network based speaker verification systems were built on both limited non-native spoken English data from the BULATS test, and a large standard corpus VoxCeleb. For the configuration used all systems yielded relatively low EERs of less than 1%. Though built with only limited data the systems trained on just BULATS systems outperformed the “out-of-the-box” VoxCeleb based system. However by adapting both the PLDA model and the deep speaker representation, the VoxCeleb-based systems could yield lower EERs. The attributes of the “impostors” was then analysed in terms of both the impostor’s grade and L1. As expected, L1 was the most important attribute of the impostor selected, though the grade did also influence performance. With the most likely scenario of impersonation by restricting impostors to be from the same gender, same L1, and higher grade group, the combination of the two adapted systems gave consistently better performance for a wide range of FA and MS probabilities, making it a sensible configuration for impersonation detection.

7. REFERENCES

- [1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.
- [2] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.
- [3] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners," in *Proc. Interspeech*, 2014, pp. 1468–1472.
- [4] Y. Wang, M. J. F. Gales, K. M. Knill, K. Kyriakopoulos, A. Malinin, R. C. van Dalen, and M. Rashid, "Towards automatic assessment of spontaneous spoken English," *Speech Communication*, vol. 104, pp. 47 – 56, 2018.
- [5] 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.
- [6] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.
- [7] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, 2010, vol. 14.
- [8] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.
- [9] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. ICASSP*, 2012, pp. 4257–4260.
- [10] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, 2014, pp. 4052–4056.
- [11] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proc. SLT Workshop*, 2016, pp. 165–170.
- [12] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.
- [13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.
- [14] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, "Detecting nonnative speech using speaker recognition approaches," in *Odyssey: The Speaker and Language Recognition Workshop*, 2008.
- [15] B. Tan, Q. Li, and R. Foresta, "An automatic non-native speaker recognition system," in *IEEE International Conference on Technologies for Homeland Security (HST)*, 2010, pp. 77–83.
- [16] M. K. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," in *Proc. ICASSP*, 2010, pp. 4398–4401.
- [17] Y. Qian et al., "Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment," in *Proc. Interspeech*, 2016.
- [18] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.
- [19] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
- [20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, G. Nagendra, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU Workshop*, 2011, pp. 4141–4144.
- [21] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*, 2016, pp. 5115–5119.
- [22] N. Brümmer and E. De Villiers, "The speaker partitioning problem," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, 2010.
- [23] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.
- [24] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," in *Proc. SLT Workshop*, 2014, pp. 378–383.

- [25] D. Garcia-Romero, A. McCree, S. Shum, N. Brümmer, and C. Vaquero, “Unsupervised domain adaptation for i-vector speaker recognition,” in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2014.
- [26] J. Villalba and E. Lleida, “Unsupervised adaptation of PLDA by using variational Bayes methods,” in *Proc. ICASSP*, 2014, pp. 744–748.
- [27] D. Garcia-Romero and A. McCree, “Supervised domain adaptation for i-vector based speaker recognition,” in *Proc. ICASSP*, 2014, pp. 4047–4051.
- [28] L. Chambers and K. Ingham, “The BULATS online speaking test,” *Research Notes*, vol. 43, pp. 21–25, 2011.
- [29] Council of Europe, *Common European Framework of Reference for Languages: learning, teaching, assessment*, Cambridge University Press, 2001.
- [30] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. C. Woodland, and C. Zhang, *The HTK book (for HTK version 3.5)*, University of Cambridge, 2015.