Speaker-targeted Synthetic Speech Detection

Diego Castan, Md Hafizur Rahman, Sarah Bakst,
Chris Cobo-Kroenke, Mitchell McLaren, Martin Graciarena, Aaron Lawson

Speech Technology and Research Laboratory,
SRI International, California, USA

firstname.lastname@sri.com

Abstract

Text-to-speech technologies are evolving quickly towards realistic-sounding human-like voices. As this technology improves, so does the opportunity for malpractice in speaker identification (SID) via spoofing, the process of impersonating a voice biometric via synthesis. More data typically equates to a more realistic voice model, which poses an issue for well-known subjects, such as politicians and celebrities, who have vast amounts of multimedia available online. Detection of synthetic speech has relied on signal processing techniques that focus on the generation of new acoustic features and train deep-learning models to detect when an audio file has been manipulated through the characterization of unnatural changes or artifacts. However, these techniques do not use any information from the speaker they are evaluating. This paper proposes to incorporate information from the speaker-of-interest (SoI) into the models to avoid specific spoofing attacks for certain vulnerable people. The wealth of data for well-known people can also be used to train a speaker-specific spoofing detector with a higher level of accuracy than a speaker-independent model. The paper proposes a new xResNet-PLDA system and compares it to three different baseline systems: a state-of-the-art speaker identification system, an xResNet system trained to discriminate between bonafide and fake speech, and a speaker identification system in which the PLDA and calibration models were trained with bonafide and fake speech. We evaluated the systems in two different scenarios — a cross-validation scenario and a hold-out scenario — with three different databases. We show how the proposed system outperforms dramatically the baseline systems in each scenario and for each database. Finally, we show how using a small amount of the SoI's speech to adapt global calibration parameters improves the performance of the system, especially in unseen conditions.

Index Terms: Deep-fake speech, speaker information, xResNet, calibration

1. Introduction

Fake speech has challenged the status of multimedia documents as evidence of past situations. Synthetic speech generated by deep-fake algorithms can be used to falsify events and spread online misinformation. While the generation and manipulation of speech is not new, the quality of the state-of-the-art text-to-speech (TTS) methods and the quantity of available data to train the algorithms allow users with limited knowledge to create convincing speech for targeted individuals. End-to-end models like WaveNet [1], Tacatron1/2 [2, 3], Deep Voice 3 [4], FastSpeech 1/2 [5], ClariNet [6], or EATS [7] have improved the TTS technologies considerably with their abilities to generate natural and intelligible speech. As a result, the amount of deep-fake content has consistently increased over the last few years.

Training a high-quality TTS system that mimics a specific speaker requires a large amount of transcribed speech of the speaker-of-interest (SoI). Therefore, celebrities and politicians are easy targets of malicious TTS attack. However, even when data resources are limited, there are techniques that focus on leveraging data from other speakers to improve the quality of the SoI [8, 9].

These recent developments in TTS have heightened the need for methods and systems to detect deep-fake attacks. Most deep-fake detectors are based on signal processing techniques and deep-learning methods that detect artifacts in the speech signal [10]. Although some of the artifacts have similar uncommon energy distribution, unnatural prosody, or clippings in the high frequencies, the artifacts are easy to mask by adding some background noise or music, applying filters to the speech signal, or using some specific codecs [11]. Moreover, the TTS systems mentioned above can reduce the amount of artifacts to a minimum if enough data are available to train the models properly. Therefore, detection methods based on artifacts are not reliable for high-performance synthetic speech and cannot generalize their decisions in different scenarios other than those used for training [12]. Also, these general methods do not use any information about the SoI since they are intended to be applied to all speakers without previous knowledge.

Very recently, Belli et al. [13] investigated the face enrollment approach associated with each query image for face spoof-attack detection. In contrast to contemporary systems that rely only on query images to detect fake images, this approach significantly improved spoof-attack detection. Along the same line of research, we propose a new deep-fake detection approach that can leverage the information of the SoI (enrollment) to differentiate between generated and bonafide speech. A considerable amount of literature has been published on automatic speaker verification (ASV) from the threat of spoofing thanks to the ASV Spoof evaluations [12]. While these evaluations focused on ASV as the core task, they also reported performance on fake-audio detection. However, we have not found previous work where SoI information was used as enrollment to detect fake audio to the best of our knowledge. The proposed system in this work allows a user to enroll bonafide SoI speech that the back-end of the system uses for three specific goals: a) to compare the bonafide speech (enrollment) with the test speech (query), b) to re-calibrate the system output for that specific SoI, and c) to output a likelihood ratio between the bonafide hypothesis and the fake hypothesis. This proposed approach is based on an xResNet architecture [14] trained to classify bonafide versus fake speech. From that architecture, we

extract embeddings that we use in a Probabilistic Linear Discriminant Analysis (PLDA) backend. We also propose a new challenging database of deep-fake videos from YouTube that has background noises and music. The videos are composed of 22 celebrities and politicians with many audio documents on the Internet. The human ear cannot easily distinguish between the deep fake audios from the bonafide samples.

Section 2 of this paper describes the different parts of the proposed system. Section 3 describes the databases used to evaluate the proposed solution and several other databases used during the training stage. Section 4 shows the experimental settings and results of comparing the proposed system to a general deep-fake detector and to a speaker ID system. Section 5 presents the conclusions of this work.

2. System Description

The system is composed of a) the front-end part that of the system that does not contain speaker-specific information, and b) the back-end part of the system that contains speaker information.

2.1. Front-end

The front-end of the system is composed of the acoustic features, the speech activity detector (SAD), and the deep-fake embedding extractor.

2.1.1. Acoustic Features: Linear Frequency Cepstral Coefficients

Linear Frequency Cepstral Coefficients (LFCC) is an acoustic feature that uses a series of filterbanks on a linear frequency scale (uniform separation between filters). LFCC provides higher signal resolution at high frequencies than filterbanks based on the Mel-scale [15] because the separation between filters does not increase with frequency. These high frequencies are essential for detecting deep-fakes because the artifacts in the synthetic speech are usually located in the limits of low and high frequencies of the speech spectrum [15].

2.1.2. Speech Activity Detection (SAD)

Our SAD is DNN-based with two hidden layers containing 500 and 100 nodes, respectively. The SAD DNN is trained using 20-dimensional Mel-frequency cepstral coefficients (MFCC) features, stacked with 31 frames. Before training the SAD DNN, the features were mean and variance normalized over a 201-frame window. In our previous work [16], we investigated the impact of SAD on the performance of speaker-embeddings-based speaker recognition systems. It was shown that a low SAD threshold during training tended to benefit the embeddings extractor, while maintaining a strict threshold during evaluation was necessary. The thresholds for selecting the speech versus non-speech frames was 2.0 for evaluation and -1.5 for DNN training.

2.1.3. Deep-fake embedding extractor

Deep residual networks (ResNets) were introduced to address the neural network degradation and generalization problem [17]. The skip connections in residual modules has partially relieved the degradation problem, and the ResNet architecture has demonstrated impressive generalization for image recognition. We used a variation of ResNet called xResNet trained to classify generated versus bonafide speech. The xResNet architecture comes with a small modification in the downsampling block [14] to use more information that is typically discarded in the regular ResNet models. To improve DNN generalization, we used a one-class feature learning approach [18] to train the deep embedding space with only bonafide speech. This prevented the model from over-fitting to known generated speech classes. The following OC-Softmax function is used for DNN training.

$$L_{OC} = \frac{1}{N}\sum_{i=1}^{N} log(1 + e^{\alpha(m_{y_i} - \hat{w}_0 \hat{x}_i)(-1)^{y_i}}) \quad (1)$$

where $\hat{x}_i \in \mathbb{R}^D$ and $\hat{w}_0 \in \mathbb{R}^D$ represent the normalized target-class embeddings and weight-vector, respectively. $y_i \in 0, 1$ denotes sample labels, and m_0, $m_1 \in [-1, 1]$, $m_0 > m_1$ are the angular margins between classes.

Figure 1 shows the parts of the xResNet network. This network consists of an input stem, four residual stages, and an output layer. The input stem consists of three 3x3 convolution layers with stride 2 in the first layer for downsampling, 32 filters in the first two layers, and 64 filters in the last layer. Each residual stage includes multiple residual blocks, where each residual block consists of a residual path and an identity path. The first residual stage does not include any downsampling blocks. Starting from the second residual stage, the first residual block has been replaced with a downsampling residual block. The identity path of this downsampling block first downsamples with a 2x2 average pool for anti-aliasing. The 1x1 convolution after is used to increase the number of feature maps, matching the residual path output. The first convolution block in the residual path downsamples with a stride of 2x2. It also doubles the number of feature maps to keep computation constant. To extract embeddings from the xResNet, we compute the mean of the last layer of the xResNet before the output in windows of 2.5 seconds and 0.5 second steps.

2.2. Back-end

2.2.1. Probabilistic Linear Discriminant Analysis (PLDA)

As an alternative back-end less used for this task, we propose to apply a PLDA back-end [19]. In view of the impressive results achieved with PLDA in speaker verification with embeddings, we apply it to have a reference result of PLDA in embeddings for deep-fake detection. After extracting the embeddings from the ResNet, the embeddings are transformed using linear discriminant analysis (LDA), then mean- and variance-normalized and L2 length-normalized. LDA, mean, and variance statistics are learned from a back-end training dataset. Next, PLDA is used to obtain scores for each pair of examples (verification trial.) Thus, a binary detector is used to determine if a test speech is generated or bonafide using trials. The trial is composed of a bonafide speech of the SoI (the enrollment or model), and a test speech than can be bonafide or generated (the query). The PLDA modeling that we employ can be expressed as

$$y_i = \mu + U_1 \cdot x_1 + \epsilon_i, \quad (2)$$

where μ is the speaker-independent mean vector, U_1 is the eigenspeaker matrix, x_1 is the speaker factor, and ϵ models the residual variability.

2.2.2. Calibration

There are many different calibration strategies in the literature (e.g., [20]). In this work, we applied a common and simple solu-

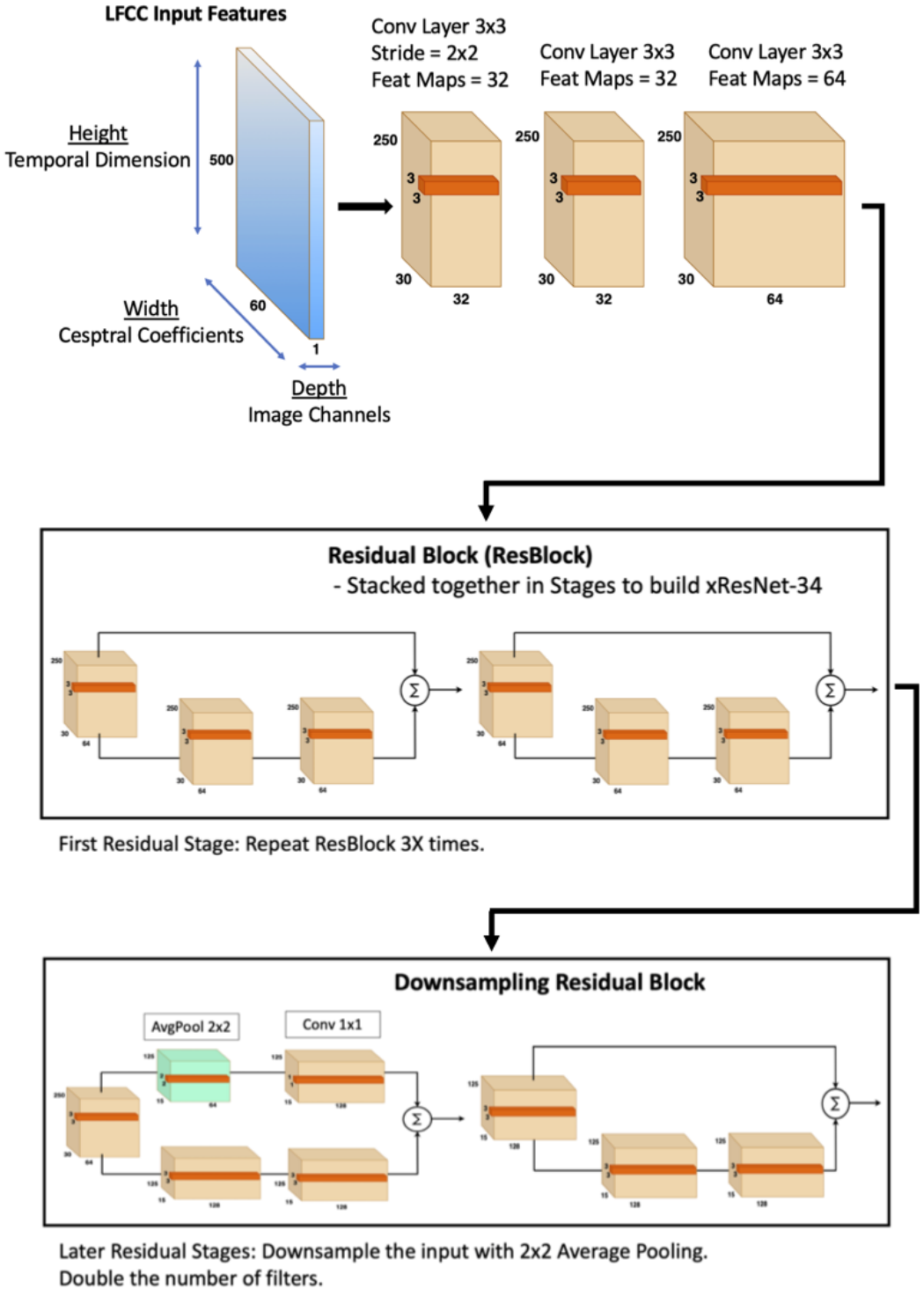


Figure 1: xResNet34 input stem with example residual blocks. The xResNet has a modified input stem compared to its original version. The original 7x7 convolutional layer was replaced with three 3x3 convolutional layers. After the input stem, the ResNet consists of four residual stages. Each stage has a certain number of residual blocks ([3, 6, 4, 3] in this system). The first residual stage always passes through the full-sized input. After the first stage, every other stage downsamples the input and double the number of filters to keep the computation constant. The downsampling is performed with a 2x2 average pooling for its anti-aliasing benefits. The final output of the residual stages is then fed to our statistical pooling and embedding layers for further processing.

tion using a discriminatively trained affine transformation from scores to log-likelihood ratios (LLRs.) The parameters of this transformation (a scale α and an offset β) are trained to minimize a weighted binary cross-entropy objective which measures the ability of the calibrated scores to make cost-effective Bayes decisions when they are interpreted as LLRs [21]. Assuming the calibration training data reflects the evaluation conditions, this procedure has shown great performance in different conditions. Yet, when evaluation conditions differ from those in the calibration training data, the average performance of the hard decisions made with the system can be poor, sometimes worse than that of random decisions [22]. Finally, we propose a regularization approach to adapt the global calibration model using SoI speech for the test trials. Calibration parameter training requires both positive and negative trial scores, and we use the speech of the SoI to increase the score count of SoI bonafide trials to get more matched samples for this process as previously proposed in [23].

3. Data

3.1. Front-end training data

For better DNN generalization and to prevent the model from over-fitting to any known speech generators or datasets, our training pool includes bonafide and synthetic speech from a diverse collection of state-of-the art TTS and VC algorithms. We used 17 speech generators samples from the training subset of the ASV Spoof 2019 [24] logical attack task, which includes 2,580 genuine and 22,800 generated speech utterances. To avoid model over-fitting, we used half of the generated data from each speech generator. Another publicly available synthetic speech data collection is the Fake or Real (FoR) dataset [25], which contains more than 111,000 bonafide and 87,000 synthetic speech samples from 33 speech generators. For DNN training, we used 53,000 bonafide and 53,000 synthetic speech samples originating from 32 speech generators from this dataset. We also used in-house generated 30,000 TTS samples from RTVC [26] and Tacotron2 [3] trained with 80 LibriTTS speakers.

We augmented training data with four types of audio degradation: (1) reverb, (2) compression, (3) instrumental music, and (4) noise. Noises included babble, restaurant noises, different in-door and outdoor sounds, traffic, mechanical, and natural sounds at 5 dB signal-to-noise (SNR) ratio. We also applied a frequency-masking [27] technique to randomly dropout frequency bands during training ranging from f_0 to $f_0 + f$, where f is chosen from a uniform distribution from 0 to maximum number of masked channels, F.

3.2. Testing data

3.2.1. ASV Spoof 2019

ASV Spoof is a well-established community led challenge for evaluating spoof attacks in speaker-verification systems. These databases mainly come with two types of attacks: physical attack (PA) and logical attack (LA). The PA task consists of recorded speech spoof attack, while the LA task consists of synthetic speech attack generated with TTS and VC systems. In this paper, we used ASV Spoof 2019 [24] LA task for our system evaluation. This task consists of 7355 bonafide samples from 67 speakers and 63882 generated samples from 48 speakers. Generated samples are produced with 2 known and 11 unknown state-of-the-art TTS synthesis and VC algorithms.

3.2.2. VCC data

The voice conversion challenge (VCC) is a bi-annually held challenge to progress and compare the latest voice conversion technologies in the research community. Three past editions of VCC provided us with a sufficient amount of data for synthetic speech detector testing. From VCC 2016 [28], we used 324 bonafide samples and in-house generated 108 VC samples with CycleGAN [29] and 80 samples with VQ-VAE [30] from 2 unseen speakers. From VCC 2018 [31] dataset, we used 470 bonafide samples and participant submitted 20,627 generated samples from 4 different speakers.

3.2.3. Speaking-of-AI

A corpus of deep-fake audio was created from the Speaking-of-AI YouTube channel[1] via the *youtube-dl* program[2]. This channel contains synthesized videos of a collection of celebrities such as Betty White, Adam Driver, and characters like Homer Simpson (Dan Castellaneta). AI models were developed by the Speaking-of-AI author and are not publicly available. Each video of the channel was diarized to keep the deep-fake voice and delete the other speakers. Ground-truth versions of these voices were found from various sources across YouTube. Audio was extracted from videos as wav files with ffmpeg at a sampling rate of 16kHz. In many cases, the ground-truth audio contained multiple voices (e.g., in conversation) or periods of background noise (e.g., applause). In-house diarization software was used to separate irrelevant voices and background from the ground-truth audio. These clips were manually corrected by one of the authors.

3.3. Trial preparation

To enroll the speaker-of-interest (SoI), we created trials where the enrollment utterance is always bonafide speech, and test utterance can be either bonafide or fake speech. In contrast to speaker identification trials where the enrollment and the test utterances can or cannot be from the same speaker, in this task, all the enrollment and test utterances are from the same speaker. We defined the target trial as the trail where both utterances are bonafide speech; the impostor trial is composed of a bonafide speech in the enrollment and fake speech in the test.

Table 1: Number of speakers, target trials, and impostor trials for ASV Spoof 2019, VCC, and Speaking-of-AI datasets

	Num. Spk.	**Tgt. Trials**	**Imp. Trials**
ASV Spoof 2019	67	380533	7277893
VCC	17	633552	2390919
Speaking-of-AI	22	10748	2147

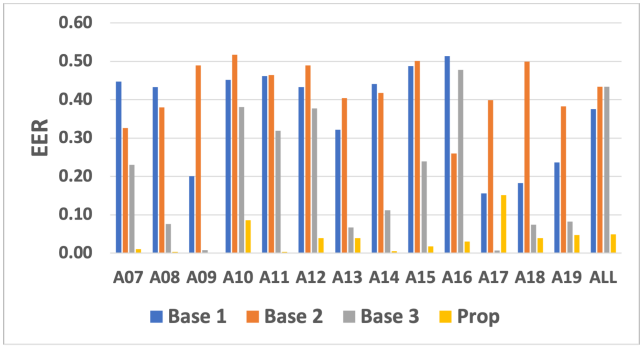
Table 1 shows the number of speakers and the number of target and impostor trials for each database used for the evaluation.

4. Experimental Settings

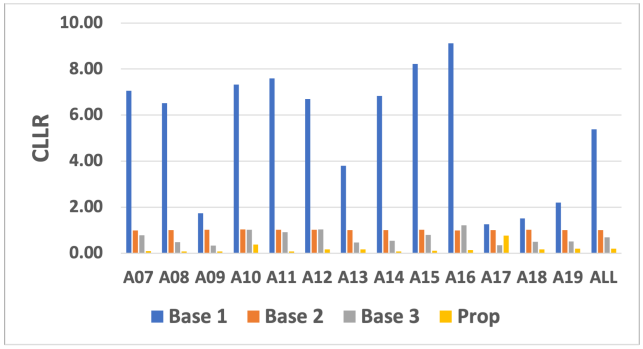
We conducted two sets of experiments that show different scenarios and conditions.

[1] Accessed at https://www.youtube.com/channel/UCID5qusrF32kSjoSGq3rJg

[2] Available here: https://github.com/ytdl-org/youtube-dl



(a) Equal Error Rate (EER)



(b) Actual Cost of the Likelihood Ratio (Cllr)

Figure 2: Equal Error Rate (EER) and the Actual Cost of the Likelihood Ratio (Cllr) of each system for each condition of the ASV Spoof 2019 testing set and the average over all conditions

4.1. Cross-validation scenario

In this section, we evaluate four different systems: three baseline systems and the proposed system.

- **Baseline 1:** A standard Speaker-ID system based on x-vectors [32]. The input features for the embedding extraction network are power-normalized cepstral coefficients (PNCC) that, in our experiments, produced better results than the more standard mel frequency cepstral coefficients (MFCCs). We extracted 30 PNCCs, and the features were mean and variance normalized over a rolling window of 3 seconds. Silence frames are discarded using the DNN-based speech activity detection system described in section 2. System training data included 234K signals from 14,630 speakers. This data was compiled from NIST SRE 2004–2008, NIST SRE 2012, Mixer6, Voxceleb1, and Voxceleb2 (training set) data (see [22] for a detailed description). The training data for the PLDA backend and the calibration was a subset of the training data used for the speaker embeddings DNN. In this case, the PLDA was trained to identify real speakers and, therefore, no fake data and no data partitions has been used in this system except for the test partition. For more details about the system and references about the databases, we refer to [23, 22].
- **Baseline 2:** For this baseline system, we used the output of the xResNet described in section 2 since the net was trained to classify generated versus bonafide speech. In this case, we have not used the embeddings, and the final score is computed as the average over the speech segments of the utterance. This system cannot be used to enroll speakers directly as the Baseline 1 system since the detection is done in our previously defined front-end.
- **Baseline 3:** We used the same speaker-ID system described in Baseline 1 but with a different PLDA backend and calibration models. In this case, the PLDA was trained to identify real speech versus fake speech using the partitions of the cross-validation as described below.
- **Proposed system:** This system is the same xResNet described in Section 2, but in this case, we extracted embeddings from the last layer of the net before the softmax layer using a mean over the activations. Then we used the embeddings to train and evaluate the PLDA and calibration system in the same way as for system Baseline 3.

We used a 4-fold cross-validation method where two portions of the data were used to train the PLDA models, one portion of the data was used to train the calibration models, and the final portion of the data was used for testing. Because the data was split in terms of speakers, there are no overlapping speakers across the partitions. The experiments are done with ASV Spoof 2019, the VCC data, and the Speaker-of-AI data previously described.

4.1.1. Metrics

We used the equal error rate (EER) and the actual cost of the likelihood ratio (Cllr) to evaluate the baselines and proposed systems for both scenarios. While EER provides a measure of the discrimination of a system, the Cllr provides both the discrimination and the calibration performance. The calibration is motivated by the requirement to be able to make good decisions in the face of uncertainty [33]. Cllr indicates score calibration across all operating points along a detection error tradeoff (DET) curve. Van Leeuwen et al. [34] provide deeper description about these metrics.

4.1.2. Results

Figure 2 compares the average of the cross-validation partitions for each condition of the ASV Spoof 2019 data, and the total average over all the conditions. The left part of the figure shows the discrimination of the systems in terms of the EER, while the right part of the figure shows the Cllr of the same conditions and systems. Proposed system outperforms the rest of the systems in discrimination (EER) and calibration (Cllr) for all the conditions except condition A17 where Baseline 3 system shows better performance. Comparing the systems for all conditions shows that the proposed system provides an improvement of 77% for EER and 72% for Cllr relative to the second-best system (Baseline 3). A Cllr above 1 indicates a poor-calibrated system. Figure 2 also shows that the Baseline 3 system presents better calibration than the other baseline systems, sespecially when compared to Baseline 1. This is due to the exposure of the backend systems (PLDA and calibration) to the conditions using the rest of the partitions of the cross-validation. However, the Baseline 3 system shows similar discrimination performance as the Baseline 2 system. Baseline 2 shows slightly worse performance than Baseline 1 in terms of EER, but Baseline 2 is better calibrated than the Baseline 1 because the xRes-

Net was trained using 17 speech-generator samples from the training subset of the ASV Spoof 2019 dataset.

Table 2: EER and Cllr for ASV Spoof 2019, VCC, and Speaking-of-AI datasets

	Base 1	**Base 2**	**Base 3**	**Prop**
DATA		**EER**		
ASV Spoof 2019	0.38	0.43	0.22	**0.05**
VCC	0.29	0.29	0.23	**0.07**
Speaking-of-AI	0.09	0.41	0.08	**0.07**
		Cllr		
ASV Spoof 2019	5.38	1.01	0.68	**0.19**
VCC	3.34	1.00	0.74	**0.54**
Speaking-of-AI	12.08	1.02	0.76	**0.67**

Table 2 presents the results obtained in all the conditions of the ASV Spoof 2019 and the results with VCC and Speaking-of-AI databases for all the systems. The results using VCC and Speaking-of-AI data show similar trends. Again, the proposed system shows improvement in calibration and discrimination for both databases. Similar to the results for ASV Spoof2019, the Baseline 3 system is better than the other two baseline systems, especially in terms of calibration. Strong evidence indicates that training the backend systems with similar conditions to those the system sees during the test stage significantly improves calibration performance.

4.2. Hold-out scenario

Based on the results, we designed a scenario where the backend systems were not trained using in-domain data, as happened with the cross-validation experiments. Because the backend systems need enough data to train good models that can perform properly with out-of-domain data, we used ASV Spoof 2019 and the VCC data to train the PLDA and the calibration model, and we tested the systems with Speaking-of-AI data, which contains realistic utterances with background noises and music.

We focused our experiments for this scenario on effects that occur when the backend was trained with out-of-domain data. The following information reports the results for Baseline 3 and the proposed system. (Baseline 1 is very similar to Baseline 3, Baseline 2 is not designed to enroll new speakers, and previous results (Section 4.1) suggested that Baselines 1 and 2 perform poorly in comparison to Baseline 3 and the proposed system.)

4.2.1. Metrics

To analyze what part of the Cllr is due to miscalibration, we used the minimum Cllr that is computed when the test scores are calibrated optimally. This is usually done by using the PAV algorithm [35] on the test scores. Also, we show results using the detection cost function (DCF) used in NIST SREs, which, in this case, is the average cost on the test data with a prior probability of bonafide speech, $P(Bs) = 0.5$, and unity costs. That is, we computed $DCF = 0.5 * Pmiss + 0.5 * Pfa$, where $Pmiss$ is the probability of labeling a bonafide trial as a fake trial and Pfa is the probability of labeling a fake trial as a bonafide trial. The errors are computed on hard decisions made by thresholding the scores with the threshold that would result in the best expected DCF if the scores were well calibrated. For the Cllr, a minimum value of DCF can also be obtained to determine what part of the DCF is due to misscalibration. In this case, the minimum is obtained by simply sweeping the threshold and choosing the one that minimizes the DCF.

4.2.2. Results

The upper part of Table 3 shows the results on Baseline 3 and the proposed system when the PLDA model was trained with 75% of ASV Spoof 2019 and VCC databases and 25% was used to train the calibration model (as known as Global Calibration in the table). The proposed system shows a relative improvement of approximately 50% for all the metrics, outperforming Baseline 3. This is the same trend shown in the experiments in the cross-validation scenario. However, the differences between the Cllr and the minCllr and between the DCF and the minDCF suggest that the calibration of the proposed system could be improved.

While it is likely that the calibration model (and the PLDA model) was trained partially with the same type of TTS used in the Speaking-of-AI database due to the limited number of TTS systems with good quality, the model was not trained with the speakers and background noises that we are evaluating. The problem that we are tackling is that the celebrities and politicians (SoI's) have enough bonafide speech data in the Internet to be able to train a good TTS to impersonate them. Therefore, it seems reasonable to have access to some extra bonafide speech from the SoI to use it in our calibration system. In order to do that, we collected 3 more utterances of 30 secs for each of the SoI to be able to create an extra 6 target trials per speaker that we use to adapt the global calibration system. We used the new target trials and 396 impostor trials (three times the amount of target trials) randomly selected from the global calibration pool of data (ASV Spoof 2019 and VCC) to adapt the global calibration model using regularized linear logistic regression. In our previous work [23], we investigated the impact of different regularization factors and determined that values between 0.02 and 0.05 lead to the best performance. For this experiment, we used a 0.05 regularization factor to adapt the default parameters, $\alpha 0$ and $\beta 0$, given by the global calibration model. Figure 3 compares the score distributions of the bonafide and fake trials with global calibration (dotted lines) and after adapting the global calibration parameters (solid lines). The figure shows the shift of the score distributions calibrated with the global modal toward the center (zero value) with the adapted parameters.

Table 3: EER, Cllr, minCllr, DCF and MinDCF for Speaking-of-AI database in a hold-out scenario

	EER	**Cllr**	**MinCllr**	**DCF**	**MinDCF**
			Global Calibration		
Base 3	0.065	1.12	0.30	0.83	0.12
Prop	**0.038**	**0.48**	**0.18**	**0.34**	**0.07**
			Speaker-specific Calibration		
Base 3	0.065	0.58	0.30	0.29	0.12
Prop	**0.038**	**0.40**	**0.18**	**0.25**	**0.07**

The bottom part of Table 3 shows the results obtained with the adaptation. While the calibration does not affect the discrimination of the system (same EER, minCLLR, and minDCF), the Cllr and the DCF show values closer to minCllr and minDCF, respectively. Therefore, using some bonafide speech from the SoI's provides a better calibration especially in unseen scenarios.

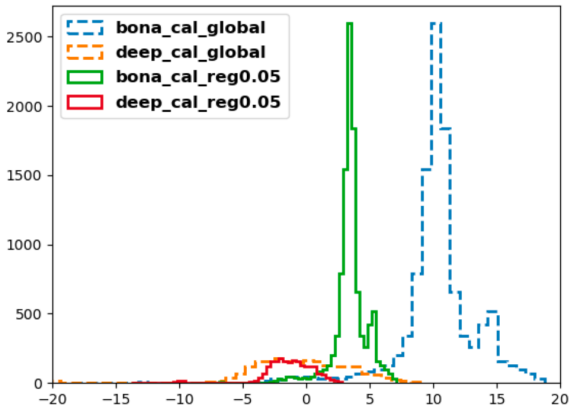


Figure 3: Scores of the bonafide and deep-fake trials with global calibration (dotted lines) and adaptation (solid lines)

5. Conclusions

This paper describes the process and results of our investigation into the detection of deep-fake speech based on modern techniques of TTS. The purpose of the study was to determine how including bonafide and fake speaker-of-Interest (SoI) data in the backend models would affect system performance. We compared the proposed system, based on xResNet and PLDA backend, to three different baseline systems: a state-of-the-art speaker identification system, an xResNet system trained to discriminate between bonafide and fake speech, and a speaker identification system in which the PLDA and calibration models were trained with bonafide and fake speech. Results in two different scenarios and with three different databases suggest that the xResNet-PLDA proposed system outperforms the baseline systems. Finally, we proposed an adaptation of the global calibration model with bonafide speech from the SoI. Adapting the calibration model with a small amount of speech reduces the miscalibration of the system compared to calibration with global parameters.

6. Acknowledgement

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0124. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). This research was funded in part by Sandia National Laboratories, a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

7. References

[1] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WAVENET: A GENERATIVE MODEL FOR RAW AUDIO," Tech. Rep., 2016.

[2] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and Q. Le, "Tacotron: Towards end-To-end speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, pp. 4006–4010. [Online]. Available: https://google.github.io/tacotron

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, pp. 4779–4783. [Online]. Available: https://google.github.io/tacotron/publications/tacotron2.

[4] W. Ping, K. Peng, A. Gibiansky, S. Arık, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[6] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[7] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-End Adversarial Text-to-Speech," in *7th International Conference on Learning Representations, ICLR 2021*, 2021.

[8] S. Arık, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 019–10 029. [Online]. Available: https://audiodemos.github.io

[9] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, C. Gulcehre, A. Van Den Oord, O. Vinyals, and N. De Freitas, "Sample efficient adaptive text-to-speech," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[10] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection," no. 1, pp. 22–28, 2021.

[11] J. Stehouwer, A. Jourabloo, Y. Liu, and X. Liu, "Noise Modeling, Synthesis and Classification for Generic Object Anti-Spoofing," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7292–7301, 2020.

[12] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVSpoof 2019: Future horizons in spoofed and fake audio detection," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1008–1012.

[13] D. Belli and B. Major, "A Personalized Benchmark for Face Anti-spoofing," *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 338–348, 2022.

[14] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional

neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.

[15] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 2087–2091.

[16] M. McLaren, D. Castan, M. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," in *Proc. of Speaker Odyssey*, Les Sables d'Olonne, France, June 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[18] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[19] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey-10*, Brno, Czech Republic, June 2010, keynote presentation.

[20] N. Brümmer, A. Swart, and D. van Leeuwen, "A comparison of linear and non-linear calibrations for speaker recognition," in *Proc. Odyssey-14*, Joensuu, Finland, June 2014.

[21] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, Apr. 2006.

[22] L. Ferrer, M. McLaren, and N. Brümmer, "A speaker verification backend with robust performance across conditions," *Computer Speech and Language*, vol. 71, p. 101258, 2021.

[23] 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 Trans. Audio Speech and Language Processing*, vol. 27, Jan. 2019.

[24] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[25] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, pp. 1–10.

[26] C. Jemine and G. Louppe, "Real-time-voice-cloning."

[27] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 132–137.

[28] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016." in *Interspeech*, 2016, pp. 1632–1636.

[29] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[30] A. Tjandra, S. Sakti, and S. Nakamura, "Transformer vq-vae for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge," *arXiv preprint arXiv:2005.11676*, 2020.

[31] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv preprint arXiv:1804.04262*, 2018.

[32] 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. of Spoken Language Technology Workshop (SLT)*, 2016.

[33] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017.

[34] D. A. Van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker classification I: Fundamentals, Features, and Methods*. Springer-Verlag, 2007.

[35] N. Brümmer and J. du Preez, "The PAV algorithm optimizes binary proper scoring rules," 2013.