Canonical Voice Conversion and Dual-Channel Processing For Improved Voice Privacy of Speech Recognition Data

Dushyant Sharma[1], Francesco Nespoli[1,2], Rong Gong[3], and Patrick A. Naylor[2]

[1]Nuance Communications Inc., USA
[2]Imperial College London, UK
[3]Nuance Communications GmbH, Austria
Email: dushyant.sharma@nuance.com

***Abstract*—This paper addresses the need for enhancing the privacy of test data in a deployed automatic speech recognition (ASR) system so that what was said cannot be linked to who said it, a process we describe as acoustic de-identification. Existing techniques can be used to modify voice characteristics to make the speaker identity unrecognizable, but normally at the expense of ASR performance. We present a novel approach for improving ASR performance on acoustically de-identified voice data. Our method exploits a dual-channel input to a self-attention channel combinator front-end to an end-to-end ASR system, and data augmentation, where some amount of original speech data is used in model training. The voice data is de-identified by a zero-shot voice style transfer system to the voice of a registered, canonical speaker. We show that the proposed approach achieves a significant improvement in privacy as demonstrated by a 10x increase in the EER of an automatic speaker verification system, while also improving the ASR accuracy as demonstrated by a 18.3% reduction in WER relative to a single channel model baseline model when tested on acoustically de-identified speech.**

I. Introduction

The ubiquitous deployment of automatic speech recognition (ASR) based smart voice systems demands a greater effort in the protection of Personal Identifiable Information (PII). The need for such privacy protection is fuelled not only by recent privacy legislation, such as the general data protection regulation (GDPR) [1], which is an EU law that regulates data privacy and protection in the EU and EEA, but also by an increasing user-awareness of privacy issues. The GDPR defines personal data as any information that is related to an identified or identifiable natural person (also known as the data subject) [1] and outlines two levels of data de-identification: (i) anonymization, where the data is processed in such as way that the data subject is no longer identifiable or (ii) pseudonymization, where the data subject can be identified only through the use of additional information. An example of pseudonymization is the use of PI masking and encryption where the original data is only retrievable through the use of additional information in the form of the correct decryption key. Two elements in speech that could identify the talker are: (a) by reference to an identifier such as a name, identification number, location data or other personal information including for example financial data, health related data, and culturally or ethnically specific information; (b) the sound of the speaker's voice as determined by factors including pitch and pitch variation, vocal timbre, tempo and other accent-related characteristics [2]. We note that a person's voice is by definition PII and therefore an important component of a speech anonymization or pseudonymization system.

As an example, it is clear that many individuals, particularly if well-known public figures, could be identified by the sound of their voice. Therefore, in order to acoustically de-identify[1] a speech signal, such as might be used for ASR-based systems, it is necessary to remove from all such utterances any identifiable acoustic information so that the speaker's voice cannot be recognized, such as by changing vocal characteristics. Whereas such algorithms serve to improve the speaker's privacy, they have the tendency to introduce characteristics that may degrade ASR accuracy as measured, for example, in terms of the Word Error Rate (WER) [3]. Accordingly, our motivation is to find a way to preserve, or even enhance, WER while at the same time using acoustically de-identified speech data as input to a deployed ASR system. We note that typically, the data that is used to train an ASR system is stored and processed in a way that enhances privacy, using for example, pseudonymization techniques such as tokenisation and encryption. In this paper, we concentrate on the privacy enhancement of test data that is provided to an ASR system at run-time.

Recently, a number of approaches for acoustic or voice-based privacy processing have been proposed, including an IEEE challenge for improving Voice Privacy [4]. In the challenge, a high performance baseline acoustic de-identification system is provided that has four main components: an F0 estimator, an ASR based acoustic model, an x-vector based speaker embedding system [5] and a speech synthesizer. After extracting the F0 contour, the phonetic posteriorgram (PPG) and the x-vector from the original utterance, a new de-identified x-vector is sampled from a pool of held-out

[1] Note that we are not addressing the content of the audio, only the acoustic aspects.

ISBN: 978-9-4645-9360-0

speakers and input to the synthesizer which, in combination with the original F0 and the PPGs, outputs the de-identified utterance. The underlying assumption of this system is that speaker information primarily resides in the x-vector. However, it was demonstrated that both the F0 and the PPGs features also contain speaker information [6], [7]. The de-identification capabilities of the systems are tested in terms of equal error rate (EER) of an automatic speaker verification (ASV) system that is based on a x-vextor extractor coupled with a probabilistic linear discriminant analysis (PLDA) [8] classifier. Finally, the WER of an ASR model is used as proxy for the quality of the synthesised speech with both the ASV and ASR systems re-trained only on de-identified data.

This paradigm however is not ideal for two main reasons. First, in a more realistic context researchers or developers in an organization typically have access to purchased or publicly available data which comes with consent to use that data as-is and therefore only a subset of the training data typically needs to be acoustically de-identified. Moreover, this data is typically secured via pseudonymizion techniques as previously mentioned, and such processing is more feasible for training data due to lower constraints on computational complexity or processing delay. Second, it is not guaranteed that training the ASR model only on acoustically de-identified data leads to the lowest WER on the acoustically de-identified test subsets. We therefore propose a novel acoustic de-identification pipeline based on zero-shot Voice Style Transfer (VST) to the voice of a registered, single canonical speaker's voice, in combination with a multi-channel ASR front-end and training data augmentation based on mixing original and acoustically de-identified data. We show that our approach achieves significant voice privacy of test data and also outperforms baseline ASR systems in terms of WER.

The remainder of the paper is organized as follows. In the following section we describe our proposed method for voice privacy and 2 channel ASR. In Section III we present experiments conducted, including the ASV system, the training and test data as well as evaluation metrics, followed by results in Section IV and conclusions in Section V.

II. Methods

In this section we outline the canonical voice conversion based acoustic de-identification algorithm and then describe the two-channel, Self Attention Channel Combinator (SACC) based ASR system followed by the data augmentation technique based on mixing original and de-identification data.

A. Canonical Voice Conversion based Acoustic De-Identification

We propose to use the voice of a registered speaker[2] as the target for a voice conversion (VC) system as a means for achieving de-identification of the original speaker's voice. In previous approaches, such as those published in the Voice Privacy Challenge [4], the target speaker embedding is chosen in a way that improves privacy, by clustering all known speaker embeddings in the training data and finding an embedding that is maximally separated from known centroids. This approach however does not guarantee that the new speaker's voice will be (a) well synthesized nor (b) that they will result in a voice that is indistinguishable from another in the training set (there is a chance that the new voice may sound like someone in the training or test data). In our approach, we always use the speaker embedding of a known and registered speaker as the target for voice conversion. This allows us to be confident that the target voice is one that is registered for us to use and since only one voice is used for all the data, this can lead to a better anonymization. Also, in a system where only one target speaker is needed, it is possible to fine tune or retrain a VC system to synthesize to the target speaker's voice (since this is a many to one transformation compared to a many to many transformation). In our experiments described in the following, we use the Coqui [9] toolkit and the zero-shot voice conversion from the YourTTS [10] model. The YourTTS system builds upon the VITS [11] model with several novel modifications for zero-shot multi-speaker and multilingual training [10].

B. SACC based ASR

The Self Attention Channel Combinator (SACC) front-end was first introduced in [12] as a novel front-end to an End-To-End (E2E) ASR system that leverages self-attention to optimally combine multi-channel audio signals. In [13] we showed how the SACC can be combined with an additional front-end comprising channel shortening with the Weighted Prediction Error (WPE) method followed by a fixed MVDR beamformer. In this work, we exploit the SACC front-end as a means to effectively combine original and acoustically de-identified speech signals.

The SACC front-end [12] operates in the power domain via Short-Time Fourier Transform (STFT) of the input signals, in this case from the two channels. The output is a single-channel magnitude spectogram, $\mathbf{Y} \in \mathbb{R}^{T\times F}$ obtained by taking an element-wise product and sum over the channel dimension of the weights matrix $\mathbf{W} \in \mathbb{R}^{T\times 2\times 1}$ and the normalized logarithmic power of multi-channel input $\mathbf{X} \in \mathbb{R}^{T\times 2\times F}$, $\mathbf{Y} = \mathbf{W}\mathbf{X}^{\mathsf{T}}$, where T and F are the numbers of time frames and frequency bins, respectively. A scaled dot-product self-attention mechanism is utilized to compute the weights $\mathbf{W}$ [12]. The mixed single channel output signal is then processed through a 64 dimension Mel-Filterbank.

Following the SACC front-end, we use an attention encoder-decoder based E2E ASR system with a ContextNet [14] based encoder and a single layer LSTM decoder [12]. For all experiments reported here, the ASR system is trained for 110 epochs and checkpoint averaging is performed over three checkpoints. An early stopping patience of 20 epochs is used and Spectral-Augmentation [15] is also enabled on the input feature stream.

In order to train a two-channel SACC based ASR system, we consider different options for processing the original and acoustically de-identified signals. In all cases, we assign chan-

[2] We assume that we have the registered speaker(s) consent for the use of their voice in this manner.

nel 1 to the original speech and channel 2 to the de-identified speech. The following table summarizes the different options for processing the original signals when the original signal is to be de-identified (in the following, the original speech is always replaced by either a noise or a zeros signal).

	Channel 1	Channel 2
C-Noise	-3 dBFS white noise	VST(orig.)
M-Noise	white noise + RMS matched to Ch.2	VST(orig.)
Zeros	zeros	VST(orig.)
Zeros+M-Sig	zeros	VST(orig.) + RMS matched to orig.

When testing the SACC based ASR systems, the original test condition is based on a two-channel signal (where channel 1 is the original speech and channel 2 is the acoustically de-identified speech). For the VST condition, the channel 1 signal is zeroed out.

C. Data Augmentation

We propose to mix different proportions of permitted original training data with acoustically de-identified data as a type of data augmentation strategy. This approach is motivated by lower constraints on computational complexity and/or processing delay when securing training data and also due to the use of data that may not be subject to stringent privacy concerns, such as that which is obtained by a deliberate collection with appropriate consent or public data. We show that such processing results in better performing ASR models (single and two channel).

III. Experiments

In this section we describe the data and metrics used for training and testing the ASR systems. Also, for the purpose of evaluating the impact of the acoustic de-identification scheme on voice privacy, we use a speaker verification system, similar to the Voice Privacy Challenge [4].

A. Data

The training and test data are based on the Libri-Speech [16] corpus. For training, we use a combination of the clean-100 and clean-360 data, amounting to 460 hours. of data. The testing uses the Libri test-clean and test-other data-sets.

B. Voice Conversion

In this work, we primarily work with a canonical speaker as the target of the VC system. However, we also evaluated a setup where a set of 4 target speakers are available. For these experiments (labeled 4S in the results section), 2 male and 2 female speakers were selected from the VCTK data-set [17] and their speaker embeddings extracted using the Coqui toolkit [9]. For each utterance in training and test data, the target speaker was chosen as the one whose speaker embedding was most distant, as measured by the cosine distance between pairs of speaker embeddings.

C. Automatic Speaker Verification

In this work, we use the automatic speaker verification task as a means of assessing the improvement in voice security achieved on the test data. The evaluation is performed using two systems, the first of which is the taken from the Voice Privacy Challenge [3] and is based on x-vector speaker embeddings followed by a PLDA classifier [18], that outputs a log-likelihood ratio (LLR) score for a pair of enrollment and trial x-vectors, which is compared to a threshold to make a speaker verification decision. The x-vector embeddings are based on a DNN using 24 dimension Mel-filter-bank features extracted with a frame-length of 25 ms and mean-normalized over a sliding window of up to 3 s [19]. The DNN is trained to classify between speakers in the training data and an embedding of dimension 512 is extracted for each utterance. We also use a second speaker verification system from speechbrain [20] that is built on an ECAPA-TDNN [21] based speaker embedding system. We re-trained this model using the single channel version of the 50% mixed data-set (where half of the training data was acoustically de-identified using a single speaker's voice) for 100 epochs with a batch size of 128, 80 dimension Mel-Filter-bank coefficients and a speaker embedding of size 192 (as in the pre-trained model from speechbrain). Additional data augmentation was also performed as per the original scripts, including speed augmentation, spectral augmentation, additive noise and reverberation [20]. The Additive Angular Margin (AAM) [22] loss was used with an Adam optimizer (learning rate initialized with a value of 0.001) and learning rate decay using cyclical learning rate [23]. This retraining represents the use case where an attacker has knowledge of the canonical speaker and voice conversion system used in our privacy enhancement system and thus allows for a more stringent evaluation of the voice privacy aspect of this work.

D. Metrics

We use the Word Error Rate (WER) metric for assessing ASR performance. In addition, the Equal Error Rate (EER) and Minimum Detection Cost Function (minDCF) metrics are used for assessing the degradation that the VST process causes in speaker recognition systems as a means for voice security improvement. An ASV system has to typically trade-off between two types of errors: a false acceptance or false alarm (FA), when the system incorrectly accepts an impostor and a false rejection (FR), where the system incorrectly rejects a speaker as an imposter. These can be defined as follows.

$$\text{FAR} = \text{number of FA} \div \text{number of imposter trials} \tag{1}$$

$$\text{FRR} = \text{number of FR} \div \text{number of user trials} \tag{2}$$

The EER metric represents the operating point (θ_{EER}) at which the false reject rate ($P_{FRR}(\theta)$) and false alarm rate ($P_{FAR}(\theta)$) are equal, i.e. $EER = P_{FAR}(\theta_{EER}) = P_{FRR}(\theta_{EER})$ [3]. Following this definition, an EER of 50% corresponds to a perfect de-identification.

Another metric that is often used in evaluating ASV systems is proposed in the NIST Speaker Recognition Evaluations

(NIST-SRE) [24] as a weighted sum of false reject and false-alarm error probabilities for some decision threshold θ, denoted as the Detection Cost Function (DCF) and defined as [25]

$$DCF = \begin{cases} \text{C(FR)} \times \text{P(user)} \times \text{FRR+} \\ \text{C(FA)} \times \text{1-P(user)} \times \text{FAR}, \end{cases} \tag{3}$$

where P(user) is the prior probability that a user will use the system (set to 10^{-1}), $1 - \text{P(user)}$ is the prior probability that an impostor will use the system, C(FR) is the cost of a false rejection, and C(FA) is the cost of a false acceptance [24], [25]. For the ECAPA-TDNN [21] based ASV system, we evaluate the performance using the accuracy metric directly.

Table I: Different ways of generating the two-channel signals for SACC based ASR trained with 5% de-id mixed with original speech. The Original test set comprises two-channel signals (ch.1 original speech and ch2. VST speech). The VST test set has a zeros ch.1 signal and the evaluation metric is WER (%).

		Test-Clean		
Model	Proc.	Original	VST	Avg.
SD-5	C-Noise	5.7	8.4	7.1
SD-5	M-Noise	5.7	7.8	6.8
SD-5	Zeros	5.6	8.0	6.8
SD-5	Zeros+M-Sig.	5.6	7.7	6.7

Table II: WER (%) results for baseline systems without data mixing augmentation (i.e. 0 or 100% de-id). 1S and 4S refer to 1 or 4 target speakers for vst processing, respectively.

		Test-Clean			Test-Other		
Model	Ch	VST	Orig.	Avg.	VST	Orig.	Avg.
Orig.	1	8.4	**4.5**	**6.5**	28.3	**14.5**	**21.4**
VST-1S	1	**7.1**	8.8	8.0	**24.1**	30.9	27.5
SACC-1S	2	7.9	5.2	6.6	26.3	17.5	21.9
VST-4S	1	7.5	10.3	8.9	25.0	26.7	25.9
SACC-4S	2	9.7	5.3	7.5	32.3	18.2	25.3

IV. Results

As described in Section II, we consider different options for assembling the two-channel signals to use with the SACC based ASR front-end. The results for these are shown in Table I, where it can be seen that the best ASR results (lowest WER) are obtained with the Zeros+M-Sig. In this experiment, the two-channel SACC front-end is used and 5% of the training data is acoustically de-identified using the canonical voice conversion.

In Table II we present the results for the single channel model trained on original speech (i.e. no de-id) as well as with 100% de-identification processing with 1 or 4 target speakers (denoted VST-1S and VST-4S respectively). Also, presented are WER results for the stereo SACC systems (with channel 1 set to zeros). It can be seen that for the de-identified test set (VST column), the single channel VST-1S model performs the best, but performs poorly on the original speech. The SACC based models outperform the single channel models on the mean WER scores (offering a balance in performance on the original and de-id test sets). In Table III we show the results for the 1S condition and with a mix of de-id and original speech (we explore 5%, 25%, 50% , 75% and 95% mix of de-id data during training). We can now observe a large improvement in performance for both single and two-channel systems on all test sets and processing conditions. The best result for de-identified test data (VST) is obtained with a 50% mix of de-id and original data with the two-channel SACC ASR system, outperforming the single channel VST-1S system by 3.8% WERR. The improvement relative to the model trained with original speech is an 18.3% relative reduction in WER. This shows that using such data augmentation and a two-channel SACC based system can achieve good WER scores and outperform the matched condition system.

Table III: WER results with varying amounts of data mixing augmentation and acoustic de-identification with a canonical target speaker (i.e. 1S). The test data here is fully acoustically de-identified.

Ch.	Mix(%)	Test-Clean	Test-Other	Avg.	WERR-Orig.
1	5	7.5	26.0	16.8	8.7
2	5	7.7	25.8	16.8	8.7
1	25	7.2	24.2	15.7	14.4
2	25	8.6	26.9	17.8	3.3
1	50	7.4	25.2	16.3	11.2
2	50	**6.9**	**23.1**	**15.0**	**18.3**
1	75	7.3	24.2	15.8	14.2
2	75	7.0	23.5	15.3	16.9
1	95	7.1	23.8	15.5	15.8
2	95	7.1	23.3	15.2	17.2

The next consideration is the evaluation of the voice privacy aspect of the proposed canonical voice transformation process. For this purpose, we evaluate the performance of two ASV systems and observe the increase in the EER when switching from original speech to the de-id speech signals. In Table IV we present the accuracy and the threshold score when the re-trained ASV system (ECAPA-TDNN labeled columns) is evaluated with the original utterances, for which the system has a high accuracy of 95.0% and this drops significantly when the acoustic de-id processing is performed (with accuracy dropping to 26.4%). A similar pattern is observed for the threshold score. In Table IV we present the EER and min. DCF scores for the x-vector PLDA based ASV system from the Voice Privacy Challenge [3] (ASV-CPC labeled columns), where we can see that the EER increases nearly 10-fold, from 3.8 for original speech to 39.4 for the acoustically de-id speech. A similar pattern is observed for the min. DCF metric. This confirms that the proposed canonical voice conversion processing significantly improves the voice privacy of the data.

V. Conclusions

In this paper, we presented a novel scheme for improving the privacy of a voice signal and its use with a two-channel SACC front-end based E2E ASR system. We show that the canonical voice conversion process achieves significant improvement in voice privacy as demonstrated by a 10-fold increase in the

Table IV: Speaker verification results using the fine-tuned ECAPA-TDNN speaker embedding model, measured using mean accuracy and threshold score, followed by results using the baseline system from the voice security challenge (asv-vpc), using the EER and min. DCF metrics.

	ECAPA-TDNN		ASV-VPC	
	Acc.	Score	EER	min. DCF
Original	95.0	0.61	3.8	0.11
VST	26.4	0.09	39.4	0.99
Δ (%)	72.2	85.2	90.4	88.9

EER of a ASV system when operated with the acoustically de-identified speech signals. By processing the acoustically de-identified signals through level matching prior to assembling with either a zeros signal or the original signal, input to a SACC front-end allows balancing the trade-off in performance on a test set that is optionally de-identified. Furthermore, if data augmentation via mixing of acoustically de-identified and original speech is performed during training, a large improvement in ASR performing is achieved. We showed that a mixing ratio of 50% acoustically de-identified and original speech allows a two-channel SACC based ASR model to outperform the matched condition model on acoustically de-identified speech (i.e. a model that is trained and tested with acoustically de-identified speech) by 3.5% WER relative (averaged over the test-clean and test-other, acoustically de-identified data-sets). This represents an 18.3% WER reduction relative to the single channel model trained with original speech.

In future work, we would explore the adaptation of the zero-shot VC system using the canonical speaker's voice, with the expectation that will would result in higher quality speech signals and thus lead to an additional improvement in ASR performance.

VI. Acknowledgements

We would like to thank Amy Uhrbach for her guidance and insights regarding the data privacy aspects of this work. We also acknowledge the European Union for funding Fracesco's work within the Horizon 2020 program (Marie Skłodowska-Curie grant No. 956369).

References

[1] "General data protection regulation," https://gdpr-info.eu, Accessed: 2023-02-22.

[2] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, et al., "Preserving privacy in speaker and speech characterisation," *Computer Speech Language*, vol. 58, pp. 441–480, 2019.

[3] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. No©, A. Nautsch, N. Evans, J. Yamagishi, B. OÄôBrien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech Language*, vol. 74, pp. 101362, 2022.

[4] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The voiceprivacy 2022 challenge evaluation plan," 2022.

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

[6] E. Gaznepoglu and N. Peters, "Exploring the importance of f0 trajectories for speaker anonymization using x-vectors and neural waveform models," 2021.

[7] M. Tran and M. Soleymani, "A speech representation anonymization framework via selective noise perturbation," 2022.

[8] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*. Springer Berlin Heidelberg, 2006, pp. 531–542.

[9] "Coqui tts," https://github.com/coqui-ai/TTS, Accessed: 2023-01-25.

[10] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[11] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[12] R. Gong, C. Quillen, D. Sharma, A. Goderre, J. Lainez, and L. Milanovic, "Self-attention channel combinator frontend for end-to-end multichannel far-field speech recognition," in *Proc. of Interspeech*, 2021.

[13] D. Sharma, R. Gong, J. Fosburgh, S. Y. Kruchinin, P. A. Naylor, and L. Milanović, "Spatial processing front-end for distant asr exploiting self-attention channel combinator," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7997–8001.

[14] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context," in *In Proc. of Interspeech*, Oct. 2020.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. of INTERSPEECH*, Graz, Austria, 2019, pp. 2613–2617, ISCA.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. of ICASSP*, Brisbane, Australia, 2015, IEEE.

[17] J. Yamagishi and X. Wang, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[19] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, 2017, vol. 2017, pp. 999–1003.

[20] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[21] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. 2020, pp. 3830–3834, ISCA.

[22] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.

[23] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.

[24] "Nist 2021 speaker recognition evaluation plan," https://www.nist.gov/system/files/documents/2021/07/12/2021_SRE_Evaluation_Plan_V5.pdf, Accessed: 2023-02-15.

[25] S. Bengio and J. Mariéthoz, "The expected performance curve: a new assessment measure for person authentication," Tech. Rep., IDIAP, 2003.