INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

HABLA: A dataset of Latin American Spanish accents for voice anti-spoofing

Pablo Andrés Tamayo Flórez[1], *Rubén Manrique*[1], *Bernardo Pereira Nunes*[2]

[1]Department of Systems and Computing Engineering, Universidad de los Andes, Colombia
[2]College of Engineering, Computing and Cybernetics, Australian National University, Australia
{p.tamayo,rf.manrique}@uniandes.edu.co, bernardo.nunes@anu.edu.au

Abstract

Research on improving automatic speaker verification systems to detect speech spoofing has focused mainly on English, with little attention given to other languages creating a significant gap in language coverage. This paper introduces HABLA, the first voice anti-spoofing dataset in the Spanish language including Argentinian, Colombian, Peruvian, Venezuelan, and Chilean accents. The dataset provided by HABLA comprises over 22,000 authentic speech samples from male and female speakers hailing from five distinct Latin American nations as well as 58,000 spoof samples that were generated through the use of six different speech synthesis strategies, including recent voice conversion and text-to-speech algorithms. Finally, initial findings on the efficacy of pre-existing Antispoofing Systems models are presented along with concerns regarding their performance in languages other than English.

Index Terms: Voice anti-spoofing, anti-spoofing dataset in Spanish, Voice conversion, Spanish spoof samples

1. Introduction

Voice recognition methods and automatic speaker verification (ASV) systems provide unique and cost-effective means of identifying individuals and re-identifying spoken utterances, respectively [1, 2, 3, 4]. ASV systems are commonly integrated into smart devices like Amazon Alexa and Google Home, enabling voice-activated tasks such as turning lights on/off, sending messages, or making calls. However, despite their benefits, both voice recognition and ASV systems are highly susceptible to spoofing attacks.

Previous works [2, 5] suggest four different forms of attacks: (1) voice imitation; (2) playback of pre-recorded audio of the target user; (3) Voice Conversion (VC); and, (4) Text-To-Speech (TTS). While the first two types rely on playing back or mimicking the original voice, the latter two create synthetic speech that is indistinguishable from human perception. Voice conversion algorithms, for example, produce high-quality forgeries by preserving the prosodic characteristics of the voice [4].

Several datasets have been created to facilitate research and training of models for detecting voice spoofing, including the VCTK, Spoofing and Anti-Spoofing (SAS), ASVSpoof, and Fake or Real (FoR) corpora [6, 7, 8]. However, all existing datasets are focused on the English language, which leaves the performance of models trained on these datasets in other languages largely unknown. Given the variability in pronunciation, accent, and prosody across different languages, it is unclear whether models trained on English datasets would generalize well to other languages. Previous research [9] argues that acoustic features and frequencies can vary considerably between languages influencing generalization. As such, there is a need for more diverse and comprehensive datasets that cover a wider range of languages and accents to evaluate the effectiveness of existing voice recognition and anti-spoofing models. Moreover, the recent development of generative models has created new challenges for anti-spoofing systems as they can deceive both current anti-spoofing solutions and human perception [10].

This paper introduces HABLA, the first voice anti-spoofing dataset in the Spanish language, including Argentinian, Colombian, Peruvian, Venezuelan, and Chilean accents. The generation of spoof examples focuses on contemporary speech synthesis algorithms, including *voice conversion* and *text-to-speech*. In total, there are more than 22,000 real samples available in HABLA's dataset, representing both males and females from five different Latin American countries, and 58,000 spoof samples generated by six different speech synthesis strategies. Finally, we discuss our findings and present the efficacy of pre-existing Antispoofing architectures [10] in our new dataset.

2. Related work

2.1. Datasets and spoof examples generation strategies

The development of multiple academic challenges that provide datasets comprising both genuine and spoofed voice samples has spurred research into techniques for voice anti-spoofing [7, 11, 6]. Reimao and Tzerpos [8] introduced a dataset comprising 111,000 and 87,000 genuine and synthetic samples, respectively. Synthetic samples were generated using commercial Text-To-Speech algorithms such as Amazon AWS Polly, Baidu TTS, Microsoft Azure TTS, and Google Wavenet TTS.

Subsequent datasets are improved by adding a greater variety of spoofing methods using voice conversion algorithms. Wu et al. [11] built their dataset by implementing five TTS and eight VC algorithms on the VCTK dataset, which contains voice data of 109 people in British English with multiple accents. Among the different VC methods presented, it is worth highlighting: (i) the modification of the spectral slope by changing the first generalized cepstral coefficient of Mel (MGC); (ii) the voice conversion toolkit provided by the open source Festvox system; and, (iii) the Tensor-Based arbitrary Voice Conversion (TVC) system. In [12, 13], generative adversarial networks were used to generate spoofed samples of a target user's voice. In [14], a diffusion probabilistic model was implemented to generate Mel-frequency cepstral coefficients from an input. In [15], variational auto-encoders are implemented to generate spoof samples, and the experiment is conducted on the VCTK corpus. Currently, no known datasets are available in the Spanish language using voice model synthesis algorithms to train specialized voice anti-spoofing models.

10.21437/Interspeech.2023-2272

2.2. Anti-spoofing models

Voice anti-spoofing models can be classified according to voice representation and feature extraction strategies. The most widely used acoustic features include Linear Frequency Cepstral Coefficients (LFCC) and Constant Q Cepstrum Coefficients (CQCC) [16]. Dua et al. [1] uses the following feature extraction strategies: Mel Frequency Cepstral Coefficients (MFCC), CQCC, and Inverse Mel Frequency Cepstral Coefficients (IMFCC). To obtain more discriminative acoustic features, a Long short-term memory (LSTM) network is implemented using an assembly strategy. On the other hand, Tak et al. [17] used the wav2vec 2.0 pre-trained model to extract audio features from wave-forms representation.

Wang and Yamagishi [16] compared different wav2vec pretrained models with LFCC. According to their results, the wav2vec achieves better results even varying the subsequent architecture (aka back-end) that uses the features to predict if the input trial is spoofed or *bona fide*. In another work [18], the same authors compared LFCC and a Lineal Filter Bank (LFB) as feature extraction strategies, plus a back-end selected from a Light Convolutional Neural Network (LCNN), an LCNN + LSTM layer, and an LCNN + Attention layer. The best results were obtained via LFCC with a back-end composed of LCNN + LSTM implemented with a Probability to Similarity Gradient (P2SGrad) activation function.

Arif et al. [4] used two representations for feature extraction: LFCC which extracted features in the frequency domain, and the extended local ternary pattern (ELTP) which extracted features in the time domain. LFCC and ELTP are combined to later enter a BiLSTM neural network to obtain the feature vector. In [2] the anti-spoof model is composed of a Convolutional neural network (CNN) capable of extracting frequency representations, an LSTM for sequence prediction, and a nonlinear classifier. That model extracts spatial local characteristics in the time domain, as well as dependencies within the frequency domain. Wang et al. [10] defined three different strategies for extracting acoustic features: LFCC, spectrograms, and LFB. In the back-end were defined three different architectures, LCNN-LSTM, LCNN-Attention, and LCNN-trim-pad. In [19] the characteristics of the voice spectrum were extracted using MFCC, CQCC, and Log Power Spectrum (LPS). The extracted characteristics are sent to a transformer encoder to extract the deepest features, then, a residual network (ResNet) calculates a composed score. In [20], a model architecture composed of a Siamese network for learning the representations and a multi-layer perceptron for classification is proposed. They use the Siamese network to extract the representation vectors of wav2vec (pre-trained model for audio signals) features.

3. Dataset construction

The authentic sample data used in this study was sourced from [21]. The study selected Argentinian, Colombian, Peruvian, Venezuelan, and Chilean accents, comprising 162 distinct speakers and 22,816 sample files, based solely on the availability of genuine samples with CC BY 4.0 license and with more than 10 speakers per accent (see Table 1). The audio samples were recorded as 48kHz single-channel and were provided in 16-bit linear PCM RIFF format, accompanied by corresponding textual translations. This text was subsequently employed to apply text-to-speech technology and generate counterfeit samples. We conducted downsampling of the data at two distinct sampling rates: 16kHz to align with the ASVSpoof format file and 22.05kHz to conform with the voice conversion models' data format.

Table 1: *Real samples distribution*

Colombian	Male	17	2534
	Female	14	2070
Chilean	Male	17	2487
	Female	12	1602
Peruvian	Male	20	2917
	Female	18	2529
Venezuelan	Male	12	1754
	Female	10	1463
Argentinian	Male	12	1670
	Female	30	3790
Total		162	22,816

For the generation of spoofed examples, five different published voice conversion architectures were evaluated, of which three were selected (StarGAN [12], CycleGAN [13], and a diffusion model [14]). The selection was performed by two human judges who rated the speech conversion in terms of their self-perceived quality. After the counterfeit sample is generated, it is down-sampled to 16khz to keep the sampling rate homogeneous throughout the dataset. Our approach to generating spoofs using voice conversion algorithms involved selecting source-target pairs of speakers. For each accent present in the dataset, we randomly selected four males and four females as source speakers. For each source speaker, we then randomly selected four males and four females from each existing accent as targets and created the different source-target pairs. This ensured that all possible combinations of accent and gender were covered.

The Microsoft Azure TTS service was chosen for the creation of Spanish TTS spoof samples. This service was selected primarily due to its wide range of TTS voices, which cater to each accent and gender featured in the genuine sample dataset. Another approach used for the generation of counterfeit examples involved a combination of TTS technology and the VC algorithms elucidated earlier. Speech samples were initially produced in the Azure TTS service using text as input and were subsequently enhanced through the utilization of VC. Table 2 depicts the distribution of the counterfeit samples. In total, 58,000 samples were produced by implementing six distinct spoof strategies. The resulting dataset is publicly available at `https://zenodo.org/record/7370805`, with a detailed description of the source-target combinations used by the voice conversion algorithms and the folder structure.

Table 2: *Spoof samples distribution*

SPOOF SAMPLES		
Name	**Type**	**# samples**
StarGAN	VC	16,000
CycleGAN	VC	16,000
Diffusion	VC	16,000
TTS	TTS	5,000
TTS-StarGAN	TTS-VC	2,500
TTS-Diff	TTS-VC	2,500

4. Experiments

The purpose of our experiments is to ascertain the efficacy of pre-existing anti-spoofing systems in the newly created Spanish

dataset. To this end, we have selected three architectures that exhibited the highest performance in [10]:

1. LFCC-LCNN-trim-pad: A light convolutional neuronal network architecture (LCNN) with an LFCC as back-end. The input size was fixed to $N = 750$ frames. In case the input is shorter, it will be padded with zeros. Otherwise, it is trimmed by selecting a random window of size N.
2. LFCC-LCNN-attention: The same LCNN, but some layers were replaced with a single-head-attention-based pooling layer and a fully connected (FC) layer.
3. LFCC-LCNN-LSTM-sum: The same LCNN, but the layers after the CNN were replaced with two Bi-LSTM layers, an average pooling layer, and an FC layer.

The three listed architectures above implement MSE for P2SGrad as loss function [22]. We used these state-of-the-art models in all our experiments. After partitioning the dataset into training (40%), validation (20%), and test (40%), we proceeded to evaluate three distinct sets of models. First, we evaluated the performance of English pre-trained models with our anti-spoof test set, using checkpoints for each model as in [10]. These models were trained using the ASVSpoof2019 dataset (English dataset). We refer to these models as EP (English pre-trained models) from now on.

Subsequently, we trained a set of models using our training set from scratch and evaluated their performance using the test set. These trained-from-scratch models constitute the second set of models. We used the same hyper-parameters as [10] for all cases. We refer to these models as SS (Spanish from scratch models). Finally, for our final set of models, we take the pre-trained English models and perform extra training rounds (further-pretrained) with our Spanish dataset. This final experiment is aimed at taking existing knowledge of the English model and refining it for Spanish. We refer to these models with the acronym FT (further-pretrained models). The metric to evaluate the performance of the different models was the EER. The models were trained for 100 (SS) and 50 epochs (FT), batch size of 64, and a learning rate of 0.0003, all trained on a Tesla A40 GPU card[1].

4.1. Experiments results

The performance of the EP models based on their architecture is illustrated in Figure 1. While using the ASVSpoof2019 dataset in English, we achieved an EER similar to that reported in [10]. However, upon evaluating the model's performance using our Spanish anti-spoofing dataset, the EER increased in all cases, surpassing 45%.

Based on these findings, we have drawn two significant conclusions. Firstly, we have been able to replicate state-of-the-art results in the English dataset ASVSpoof2019. Secondly, we observed a significant decrease in performance when applying the anti-spoofing system to languages other than the language it was trained on. While we anticipated this outcome, the magnitude of the error surpassed our expectations, indicating a lack of multi-language capability in pre-existing speech anti-spoof models.

Figure 2 illustrates a comparison of the EP and SS models in their classification of genuine voice examples from the Spanish dataset. Across the three distinct architectures, the EP models correctly classified an average of only 62.6% of genuine samples, while the SS models achieved an average accuracy

[1]Models can be found at `https://github.com/Ruframapi/HABLA`

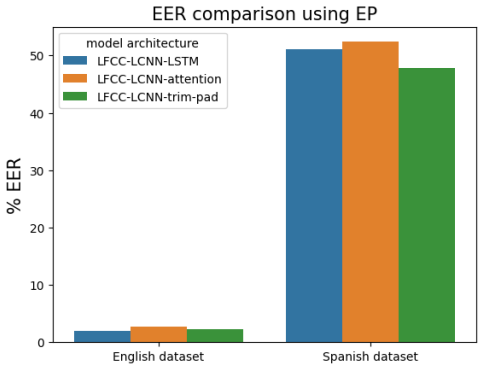

Figure 1: *EP models evaluation results over English and Spanish datasets.*

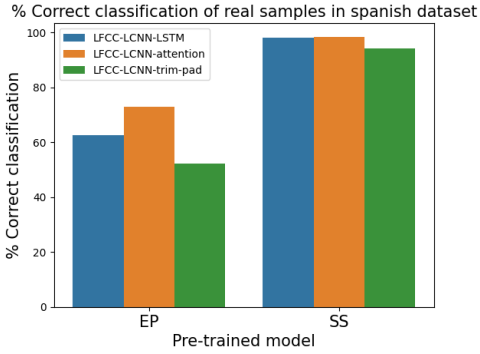

Figure 2: *EP vs SS models comparison on real voice examples from the Spanish dataset.*

of 96.08% in their classification of genuine samples in the test set. This outcome serves to reinforce the notion that language is a crucial factor in speech anti-spoofing models, and a sample originating from an unknown language may be misidentified as a genuine example.

In the subsequent analysis, we evaluate and contrast the performance of EP and SS models across the different accents present in the Spanish dataset. Figure 3 depicts the outcomes obtained exclusively for the LFCC-LCNN-LSTM-sum architecture, yet similar results were obtained for the others. Based on our observations, the EP model did not attain an accuracy greater than 80% for any of the accents. In fact, for the Colombian accent, the EP model's classification accuracy was a mere 40%. In contrast, the SS model performed with an accuracy of over 80% across all accents.

We also evaluate the performance of EP and SS models for each type of spoof generation strategy employed in our Spanish dataset (see Figure 4). Our findings suggest that the TTS and TTS-VC strategies did not pose any significant challenges for both EP and SS models. However, the results obtained for StarGAN, CycleGAN, and diffusion strategies were comparatively poor, with classification accuracy of 22%, 45%, and 62% respectively. Out of all the spoof generation techniques evaluated, the GAN-based VC strategies yielded the worst results, indicating that the generated spoof samples using these techniques were the most challenging to classify. Notably, SS models achieved 100% correct classification results (note that only spoof samples in the test set were considered for

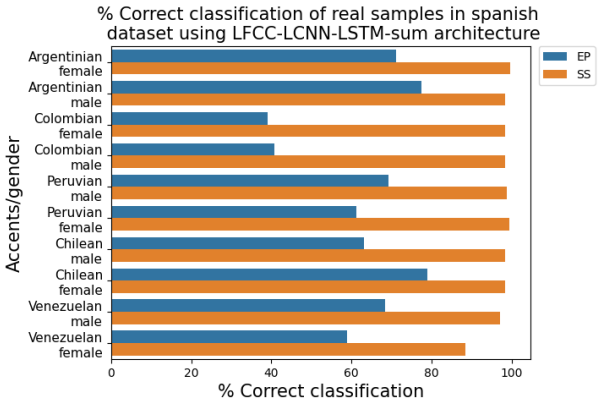

Figure 3: *EP vs SS models comparison on real voice examples grouped by accent and gender from the Spanish dataset. Only the results for LFCC-LCNN-LSTM-sum architecture are shown.*

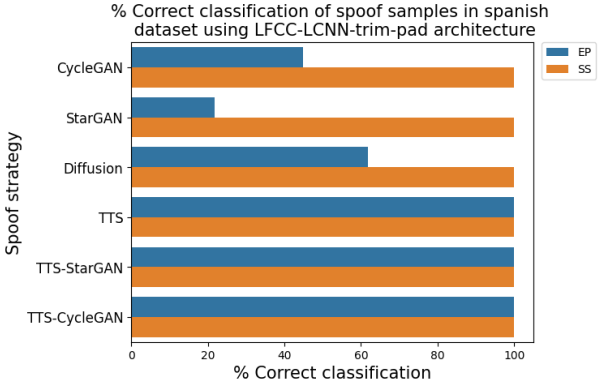

Figure 4: *EP/SS models comparison on the different kinds of spoof examples in the Spanish dataset. Only the results for LFCC-LCNN-trim-pad architecture are shown.*

this analysis), representing a significant improvement in performance compared to EP models.

Our final experiments focus on evaluating the performance of the FT models. The FT models were trained from the latest checkpoint of EP, and further trained for 50 additional epochs on our Spanish dataset. The EER results obtained for the various types of models (i.e., EP, SS, and FT) on both the Spanish and English datasets, using two distinct architectures, are illustrated in Figure 5. The cross-language combinations (i.e. when we evaluate on a language not seen in training) yield the worst results. These combinations correspond to EP models evaluated on the Spanish dataset and SS models evaluated on the English dataset. Conversely, when the model was evaluated in the same language as the one in which it was trained on, the results were better.

The FT models improve the cross-language combination effect. For instance, the SS model trained on LFCC-LCNN-LSTM-sum architecture and evaluated on the English dataset yielded 22,489% of EER, but the FT models using the same architecture achieved a 14,29% EER, which is an improvement of 8.19%. For LFCC-LCNN-Attention the improvement was 2.10%. For Spanish samples, the improvement was more than 50% in both architectures in comparison with EP models. These results show that taking a checkpoint from a pre-trained

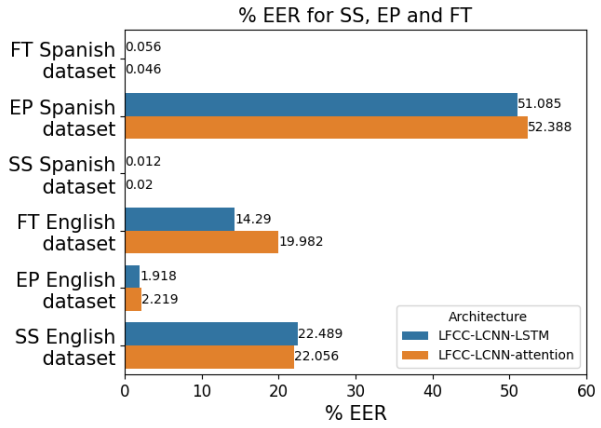

Figure 5: *EER for different models (SS, EP, and FT) and datasets (Spanish and English datasets)*

model in English and doing additional training rounds in Spanish achieves the best balance of EER in the two languages. One added benefit of this transfer learning approach is that it reduces the number of training epochs required to attain results on par with those of the SS models in our Spanish dataset.

As expected for the SS models, the EER for samples in Spanish decreased considerably but increased for samples in English when compared to the EP. A direct and global comparison suggests that the SS models are better than EP with an average error over the two languages of 11.69% compared to 26.36% of EP models. One possible hypothesis is that using the Spanish language as a training base language is better than using an English base. Nonetheless, further experimentation is required to validate this hypothesis, as the datasets are not directly comparable.

5. Conclusions

This paper introduced HABLA, the first dataset of Latin American Spanish accents for voice anti-spoofing. Our contributions can be summarized as follows:

1. We consolidate a corpus with five different accents (Colombian, Peruvian, Chilean, Argentinian and Venezuelan) from Latin-America Spanish, with six different voice spoof strategies: three VC algorithms, one TTS system, and two TTS-VC combinations, generating 58,000 spoof samples with high quality.
2. The results of the EP models revealed that the anti-spoof models trained with datasets in English have a noticeable degradation in performance when used on datasets in Spanish. Multi-language capabilities cannot be assumed even using state-of-the-art architectures.
3. The results of the SS models reveal the importance of using training data in the target language. In these models the EER for Spanish samples dropped to 3% or less and correctly classified more than 90% of the samples labeled as real.
4. None of the evaluated models had problems detecting spoof samples generated with TTS services.
5. Interestingly, all models, irrespective of their architecture, experienced greater difficulty in classifying the Colombian accent. As a potential avenue for future research, we suggest exploring the underlying reasons for this observation.

6. References

[1] M. Dua, C. Jain, and S. Kumar, "Lstm and cnn based ensemble approach for spoof detection task in automatic speaker verification systems," *Journal of Ambient Intelligence and Humanized Computing*, 2021. [Online]. Available: https://doi.org/10.1007/s12652-021-02960-0

[2] H. Dinkel, Y. Qian, and K. Yu, "Investigating raw wave deep neural networks for end-to-end speaker spoofing detection," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, pp. 2002–2014, 2018.

[3] Q. Fu, Z. Teng, J. White, M. Powell, and D. C. Schmidt, "Fastaudio: A learnable audio front-end for spoof speech detection," 2021. [Online]. Available: http://arxiv.org/abs/2109.02774

[4] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice spoofing countermeasure for logical access attacks detection," *IEEE Access*, vol. 9, pp. 162 857–162 868, 2021.

[5] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, pp. 1985–1999, 2019.

[6] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi, "Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2021.

[7] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-Janua, pp. 2037–2041, 2015.

[8] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," *2019 10th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019*, 2019.

[9] J. Schepens, T. Dijkstra, F. Grootjen, and W. van Heuven, "Cross-language distributions of high frequency and phonetically similar cognates," *PloS one*, vol. 8, p. e63006, 05 2013.

[10] X. Wang and J. Yamagishi, *A Practical Guide to Logical Access Voice Presentation Attack Detection*. Singapore: Springer Nature Singapore, 2022, pp. 169–214. [Online]. Available: https://doi.org/10.1007/978-981-19-1524-6_8

[11] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagish, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, pp. 768–783, 2016.

[12] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*, pp. 266–273, 2019.

[13] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," *European Signal Processing Conference*, vol. 2018-Septe, pp. 2100–2104, 2018.

[14] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," 9 2021. [Online]. Available: http://arxiv.org/abs/2109.13821

[15] K. Akuzawa, K. Onishi, K. Takiguchi, K. Mametani, and K. Mori, "Conditional deep hierarchical variational autoencoder for voice conversion," 12 2021. [Online]. Available: http://arxiv.org/abs/2112.02796

[16] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," 2021. [Online]. Available: http://arxiv.org/abs/2111.07725

[17] H. Tak, M. Todisco, X. Wang, J.-W. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation." [Online]. Available: https://github.com/pytorch/fairseq/tree/

[18] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 6, pp. 4685–4689, 2021.

[19] Z. Zhang, X. Yi, and X. Zhao, "Fake speech detection using residual network with transformer encoder," *IH and MMSec 2021 - Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, pp. 13–22, 2021.

[20] Y. Xie, Z. Zhang, and Y. Yang, "Siamese network with wav2vec feature for spoofing speech detection," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 6, pp. 4700–4704, 2021.

[21] A. Guevara-Rukoz, I. Demirsahin, F. He, S. H. C. Chu, S. Sarin, K. Pipatsrisawat, A. Gutkin, A. Butryna, and O. Kjartansson, "Crowdsourcing latin american spanish for low-resource text-to-speech," *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pp. 6504–6513, 2020.

[22] X. Zhang, R. Zhao, J. Yan, M. Gao, Y. Qiao, X. Wang, and H. Li, "P2sgrad: Refined gradients for optimizing deep face models," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 2019.