ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096141

QUANTUM TRANSFER LEARNING USING THE LARGE-SCALE UNSUPERVISED PRE-TRAINED MODEL WAVLM-LARGE FOR SYNTHETIC SPEECH DETECTION

Ruoyu Wang[⋆] *Jun Du*[⋆,*] *Tian Gao*[†]

[⋆]University of Science and Technology of China, Hefei, China
[†]iFlytek Research, Hefei, China

ABSTRACT

The development of quantum machine learning demonstrates its quantum advantages over traditional deep learning, which promises to discover new patterns on supervised classification datasets. This work proposes a classical-to-quantum transfer learning system based on the large-scale unsupervised pre-trained model to demonstrate the competitive performance of quantum transfer learning for synthetic speech detection. We use the pre-trained model WavLM-Large to extract feature maps from speech signals, obtain low-dimensional embedding vectors through classical network components, and then jointly fine-tune the pre-trained model and classical network components with a variational quantum circuit (VQC). We evaluate our system on the ASVspoof 2021 DF task, and the experiments using quantum circuit simulations show that quantum transfer learning can improve the performance of the classical transfer learning baseline on the task.

Index Terms— Quantum transfer learning, quantum machine learning, synthetic speech detection, variational quantum circuit, pre-trained model

1. INTRODUCTION

Synthetic speech detection techniques are designed to secure biometric systems built on automatic speaker verification (ASV) technology from the threat of fake speech attacks generated using text-to-speech (TTS) or voice conversion (VC) systems [1]. Conventional systems perform poorly on mismatched data due to the invisibility of synthetic speech attacks in the wild and the lack of real speech variety during training [2]. Transfer learning [3, 4] is considered as an excellent approach to solve such problems by taking the knowledge gained from generic data to a different data domain.

State-of-the-art synthetic speech detection methods usually use a large-scale unsupervised pre-trained model as a front-end combined with a back-end classification network for fine-tuning and substantially improving the generalizability of the system to out-of-set data [5, 6, 7]. However, there are still data privacy and security issues involved for the application, and further research is difficult to go beyond the traditional deep learning framework to discover new patterns in real and fake speech. Thanks to the introduction of quantum machine learning (QML) [8], it is expected that these problems can be solved by using quantum computing to circumvent classical computing technical limitations [9, 10].

Noisy-intermediate-scale-quantum (NISQ) devices are a major hardware class of quantum computing devices which show unique properties and empirical advantages in many applications using only a few quantum bits (5 to 100) [11]. For example, projecting classical data into a high-dimensional quantum feature space has been shown to have quantum advantages in many classification tasks [12, 13], and privacy algorithms built on cloud quantum devices API provide data isolation [10, 14]. Since current NISQ devices allow few quantum bits, using classical-to-quantum transfer learning [15] is a considered option in the present practical application.

Quantum transfer learning was introduced by Xanadu in [15], where they proposed a quantum transfer learning paradigm using pre-trained models to extract embedding vectors and then perform classification through hybrid quantum neural network (QNN). It can efficiently process high-dimensional input samples with any pre-trained deep neural network, and successively process low-dimensional but informative features with quantum circuits, which combines the benefits of quantum machine learning with deep learning models that have proven effective in practice.

Previous quantum transfer learning approaches usually directly concatenate the embedding vectors generated by pre-trained models with quantum networks. They rarely consider using large-scale unsupervised pre-trained models to deal with practical problems on some complex datasets. In particular, for speech pre-trained models, most of them do not generate embedding vectors directly, such as Wav2vec [16], WavLM [17] generating feature maps, and it is a problem to process these feature maps to low-dimensional representation as the input of quantum circuits.

Based on recent successes in acoustic modeling of quantum circuits and quantum transfer learning [18, 19, 20], we propose a quantum transfer learning system based on the large-scale unsupervised pre-trained model to achieve quantum advantages in synthetic speech detection task. In our

[*] Corresponding author.

approach, we use the pre-trained model WavLM-Large [17] to extract feature maps from the speech signal. The feature maps are downscaled by the classical network component to low-dimensional embedding vectors to fed into a variational quantum circuit with learnable parameters. The quantum component is updated jointly with the pre-trained model and the classical network components for fine-tuning. Compared with classical transfer learning systems for synthetic speech detection, our proposed system can combine quantum advantages and the generalizability of unsupervised pre-trained models to maintain a competitive experimental performance.

2. RELATED WORK

2.1. Quantum transfer learning algorithm

QML networks encode the input data with quantum bits and learn parameterized quantum gate parameters. The expressiveness is measured by the local effective dimension, and it is expected to have a quantum advantage over traditional learning methods for certain computational problems [21]. When it comes to dimensionality reduction of classical inputs, recent work by Qi *et al.* [22] further reveals that the characterization and generalization capabilities of VQC are enhanced as the number of quantum bits used by VQC increases. This suggests that quantum devices have great potential for future applications as the number of supported quantum bits increases.

Quantum transfer learning is an important technique that enables putting quantum advantages into practice when quantum devices are not yet mature enough. Qi *et al.* [18] trained variational quantum circuit-based quantum neural network by pre-trained CNN networks with fixed parameters to process speech signals as embedding vectors, reducing the training difficulty of hybrid classical and quantum networks and improving the performance of their baseline on the spoken command recognition task. Yang *et al.* [19] used BERT and stochastic quantum time convolution for vertical joint learning to obtain competitive results on text classification while ensuring data isolation. However, it remains an open question how the most practical large-scale unsupervised pre-trained speech models in transfer learning can be combined with quantum circuit backends.

2.2. Large-scale unsupervised speech pre-training model

In recent years, pre-trained models have attracted lots of attention from academia and industry in the fields of natural language processing and computer vision, while general speech pre-trained models such as Wav2vec [16], WavLM [17] are proposed in the speech field. Pre-trained models trained with large-scale unsupervised data have powerful generalizability, and only need to be fine-tuned on small-scale labeled data to be applied on the corresponding downstream tasks [23].

Their powerful generalization ability is also of interest to the field of synthetic speech detection. Wang *et al.* [5] show that the generalization ability of synthetic speech detection systems can be significantly improved by introducing the large-scale unsupervised pre-trained model as front-end combined with fine-tuning of the back-end classification network. How to better utilize the pre-trained front-end model for fine-tuning has become an issue of interest.

3. METHOD

3.1. System overview

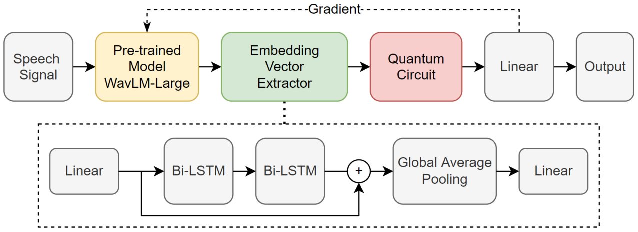


Fig. 1. The proposed quantum transfer learning system using the large-scale unsupervised pre-trained model WavLM-large for synthetic speech detection. The loaded pre-trained model will update the parameters with the other parts.

We illustrate the architecture of the proposed quantum transfer learning system for synthetic speech detection in Figure 1. We use WavLM-Large [17] as the pre-trained speech model. The WavLM-based model consists of convolutional neural networks and transformer encoders, and we use the output of last transformer layer as the output of the pre-trained model to extract the hidden feature map features $\mathbf{z}_{M\times N}$ from the speech signals $\mathbf{s}_{1\times T}$.

The quantum circuits accept only embedding vectors as input, and we mimic the approach in classical transfer learning [5] by adding an embedding vector extractor between the pre-trained model and the quantum circuits to transform the feature maps to the embedding vectors. The feature maps extracted by the large-scale unsupervised pre-trained model usually has a large dimension N (N=1024 for WavLM-Large). For the purpose of reducing the computational cost, the feature map dimension is first downscaled to $\mathbf{z}_{M\times\frac{N}{4}}$ by a linear layer, and then the temporal information is extracted through two Bi-LSTM layers. Here we add skip connection to prevent the loss of information, and finally the embedding vectors $\mathbf{x}_{1\times 4}$ are obtained by global average pooling and a linear layer as the inputs of the 4-qubit variational quantum circuit network. Quantum circuit measurements $\mathbf{y}_{1\times 4}$ are output to the linear layer for processing into real or fake binary classification results.

3.2. 4-qubit variational quantum circuit network

A d-qubit quantum state $|\mathbf{v}\rangle = \otimes_{i=1}^{d} |v_i\rangle = |v_1\rangle \otimes |v_2\rangle \otimes \cdots \otimes |v_d\rangle$ is associated with a 2^d-dimensional vector in a Hilbert space for $\mathbf{v} = [v_1, v_2, \ldots, v_d]^T \in R^d$, where for a scalar v_i, the quantum state $|v_i\rangle$ is:

$$|v_i\rangle = \cos v_i |0\rangle + \sin v_i |1\rangle = \begin{bmatrix} \cos v_i \\ \sin v_i \end{bmatrix}. \tag{1}$$

The network constructed by variational quantum circuits in Figure 2 proposed in [24], consists of three parts: quantum encoding, variational quantum circuits and measurement.

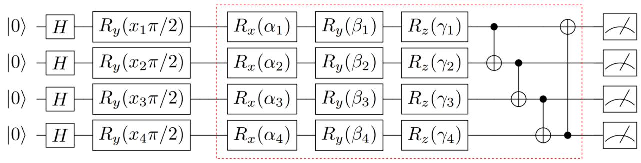


Fig. 2. A framework of 4-qubit variational quantum circuit network. The circuits in the dash square correspond to the learnable variational quantum circuit layer with repeated copies m.

First, we convert the vectors in Euclidean space to Hilbert space by quantum coding. The framework of quantum encoding constructs transformation relations between the classical data input $\mathbf{x} = [x_1, x_2, x_3, x_4]^T$ and its quantum state $|\mathbf{x}\rangle$. The encoding map $\mathcal{E}$ prepares each qubit in a balanced superposition of $|0\rangle$ and $|1\rangle$ and then performs a rotation around the y axis of the Bloch sphere parametrized by the classical vector $\mathbf{x}$, where H is the single-qubit Hadamard gate:

$$\mathcal{E}(\mathbf{x}) = \left(\bigotimes_{k=1}^{4} \left(R_y \left(x_k \pi/2 \right) H \right) \right) |0\rangle^{\otimes 4}. \tag{2}$$

Next, we learn the linear variation between quantum states and the entanglement of each quantum bit through the variational quantum layer. A variational quantum layer $\mathcal{L}$ consists of a CNOT gate and learnable rotation gates R_x, R_x and R_z which separately denote Pauli rotation X, Y and Z gates:

$$\mathcal{L}(\mathbf{w}) : |\mathbf{x}\rangle \rightarrow |\mathbf{y}\rangle = \mathbf{K} \bigotimes_{k=1}^{4} R_z \left(\gamma_k \right) R_y \left(\beta_k \right) R_x \left(\alpha_k \right) |\mathbf{x}\rangle, \tag{3}$$

where $\mathbf{K}$ is an entangled unitary operation consisting of four CNOT gates:

$$\begin{aligned} \mathbf{K} = & \left(\text{CNOT} \otimes I_{1,4} \right) \left(\text{CNOT} \otimes I_{3,4} \right) \\ & \left(\text{CNOT} \otimes I_{2,3} \right) \left(\text{CNOT} \otimes I_{1,2} \right). \end{aligned} \tag{4}$$

The CNOT gate imposes quantum entanglement between any two quantum wires and $\mathbf{K}$ ensures all quantum bits of the quantum wires can be entangled [25]. The rotation angles $\mathbf{w} = \{\alpha_k, \beta_k, \gamma_k | k = 1, 2, 3, 4\}$ are trainable parameters for R_x, R_y and R_z. A variational quantum circuit usually consists of m variational quantum layers.

Finally by $Z = diag(1, -1)$ Pauli matrix, we get the local measurement for each quantum bit:

$$\mathcal{M} : |\mathbf{y}\rangle \rightarrow \mathbf{y} = \begin{bmatrix} \langle \mathbf{y} | Z \otimes I \otimes I \otimes I | \mathbf{y} \rangle \\ \langle \mathbf{y} | I \otimes Z \otimes I \otimes I | \mathbf{y} \rangle \\ \langle \mathbf{y} | I \otimes I \otimes Z \otimes I | \mathbf{y} \rangle \\ \langle \mathbf{y} | I \otimes I \otimes I \otimes Z | \mathbf{y} \rangle \end{bmatrix}. \tag{5}$$

Given 4-dimensional inputs, a linear layer with hidden size 4 will contain 16 learnable parameters, while the variational quantum layer contains only 12, involving fewer model parameters. Moreover, compared with the linear layer, which only learns linear relations in Euclidean space, the variational quantum layer learns linear relations in Hilbert space, implying the possibility of discovering new patterns.

4. EXPERIMENT

4.1. Dataset and setup

We train the system using the training set of the ASVspoof 2019 LA database [26]. It contains real speech from the VCTK database and fake speech data generated using 6 TTS and VC systems. Data used for the training of TTS and VC systems also comes from the VCTK database without overlap. We use the ASVspoof 2021 DF evaluation data as our test set, which simulates scenarios that are very unfavorable for synthetic speech detection systems, where most of the faked and real samples use codec compression. Moreover, because the ASVspoof 2021 DF evaluation data consists of ASVspoof 2019, VCC 2018, and VCC 2020 [2, 27, 28], there are many test samples coming from mismatched data domains or generated by more diverse means of spoofing. Testing the ASVspoof 2021 DF evaluation set provides a better measure of the system's generalization performance. We use the officially recommended Equal Error Rate (EER) as our evaluation metric [29], with a lower EER implying better performance of the synthetic speech detection system. Evaluation of the EER for our test set reflects the system's generalizability to invisible attacks and unknown domains.

We directly use the original speech training data for training. The input samples are cut into segments of 4 seconds duration, without applying any other data augmentation and speech signal processing techniques. Our model is built based on Pytorch and Pennylane [30] is used to build the quantum circuit part. We initialize the pre-trained front-end network using the open-source WavLM-Large [17] model with about 316 million pre-trained parameters. Its outputs are feature maps with 1024 hidden dimensions. We use the fine-tune training strategy and the cross-entropy loss function. The minimum batch size is set to 32, and the pre-trained front-end uses a learning rate of 2×10^{-5}, and the back-end and quantum circuit parts use a learning rate of 2×10^{-3}, with the learning rate decaying by a factor of 0.7 per epoch. We train 10 epochs and select the model with the lowest cross-validation loss for evaluation.

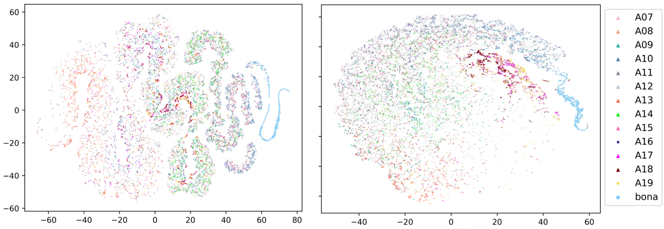
4.2. Experimental results

The quantum transfer learning system is based on Section 3 designed with one variational quantum circuit layers. We design two classical transfer learning systems as for comparison. The first system which is also our baseline system, directly removes the quantum circuit module. The 4-dimensional vector outputs of the embedding vector extractor are passed through a linear layer to produce classification results, i.e., the quantum circuit part is replaced with an identity layer. The second system in which the quantum circuit is replaced by a linear layer with hidden size 4 ensures that the parameters of the quantum method are relatively unincreased with respect to the classical method.

Table 1. EER (%) results for quantum transfer learning and classical transfer learning comparisons on ASVspoof 2021 DF evaluation data.

Transfer learning	Replacement	EER(%)
Classical	Identity baseline	6.80
	Linear: 4 → 4	6.94
Quantum	VQC	**5.51**

Table 1 shows the experimental results of our above methods. There is no significant difference between the EER results of the two classical methods, while the EER of the quantum method achieves absolute decline of 1.29% to our baseline system. Further we visualize the 256-dimensional embedding vectors from the penultimate layer of the embedding vector extractor by the t-SNE [31] method. In Figure 3, We can see that the embedding representations of the quantum and classical methods have different distribution patterns. Based on these considerations, it is reasonable to assume that quantum methods can bring performance gains and mine new patterns on the synthetic speech detection task compared to classical methods.



(a) Classical

(b) Quantum

Fig. 3. The t-SNE visualization results of the penultimate layer outputs of the embedding vector extractor in the classical transfer method and the quantum transfer method. The A07-A19 notation indicates the fake speech class generated by different synthetic speech systems and bona notation indicates the real speech in ASVspoof part of test set.

Table 2. Classification accuracy (%) of real and fake samples in test data of ASVspoof and VCC parts under thresholds corresponding to EERs of the baseline and quantum methods in Table 1.

Transfer learning	ASVspoof		VCC	
	Real	Fake	Real	Fake
Classical	92.28	87.65	**94.19**	95.00
Quantum	**96.09**	**89.62**	90.67	**96.07**

As we mentioned in Section 4.1, there are two data domains in our test data. The ASVspoof part is usually considered as in-set domain while VCC part is considered as out-set, and fake speech detection in VCC will be more difficult than in ASVspoof. In Table 2, the quantum method improves the classification accuracy of fake samples on different data domains relative to the classical method. It indicates that the quantum transfer learning method constructed based on a large-scale unsupervised front-end can improve the generalization ability of the classical transfer learning method while realizing the quantum advantage. On the other hand, our method improves the classification accuracy of in-set real samples but leads to performance degradation for out-of-set real samples, implying that quantum circuits still have the risk of overfitting when there are too few learning samples of real class and few quantum bits used in quantum circuits.

5. CONCLUSION

In this paper, we propose a quantum transfer learning system based on the large-scale unsupervised speech model WavLM-Large for synthetic speech detection. We fine-tune the loaded pre-trained model, classical network embedding vector extractor and quantum circuit jointly to improve classical transfer learning baseline system and achieve quantum advantages. Our experiments on the ASVspoof 2021 DF task demonstrate the importance of quantum transfer learning for improving generalization over unseen domains and discovering new patterns different from classical methods.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

7. REFERENCES

[1] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al., "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[2] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al., "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *arXiv preprint arXiv:2210.02437*, 2022.

[3] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[4] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*. Springer, 2018, pp. 270–279.

[5] Xin Wang and Junichi Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *arXiv preprint arXiv:2111.07725*, 2021.

[6] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[7] Youngsik Eom, Yeonghyeon Lee, Ji Sub Um, and Hoirin Kim, "Anti-spoofing using transfer learning with variational information bottleneck," *arXiv preprint arXiv:2204.01387*, 2022.

[8] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.

[9] Jun Qi, Chao-Han Huck Yang, and Pin-Yu Chen, "Qtn-vqc: An end-to-end learning framework for quantum neural networks," *arXiv preprint arXiv:2110.03861*, 2021.

[10] Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Pin-Yu Chen, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6523–6527.

[11] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.

[12] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme, "A rigorous and robust quantum speed-up in supervised machine learning," *Nature Physics*, vol. 17, no. 9, pp. 1013–1017, 2021.

[13] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean, "Power of data in quantum machine learning," *Nature communications*, vol. 12, no. 1, pp. 1–9, 2021.

[14] Samuel Yen-Chi Chen and Shinjae Yoo, "Federated quantum machine learning," *Entropy*, vol. 23, no. 4, pp. 460, 2021.

[15] Andrea Mari, Thomas R Bromley, Josh Izaac, Maria Schuld, and Nathan Killoran, "Transfer learning in hybrid classical-quantum neural networks," *Quantum*, vol. 4, pp. 340, 2020.

[16] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[17] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[18] Jun Qi and Javier Tejedor, "Classical-to-quantum transfer learning for spoken command recognition based on quantum neural networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8627–8631.

[19] Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Yu Tsao, and Pin-Yu Chen, "When bert meets quantum temporal convolution learning for text classification in heterogeneous computing," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8602–8606.

[20] Soronzonbold Otgonbaatar, Gottfried Schwarz, Mihai Datcu, and Dieter Kranzlmueller, "Quantum transfer learning for real-world, small, and large-scale datasets," *arXiv preprint arXiv:2209.07799*, 2022.

[21] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner, "The power of quantum neural networks," *Nature Computational Science*, vol. 1, no. 6, pp. 403–409, 2021.

[22] Jun Qi, Chao-Han Huck Yang, Pin-Yu Chen, and Min-Hsiu Hsieh, "Theoretical error performance analysis for variational quantum circuit based functional regression," *npj Quantum Information*, vol. 9, no. 1, pp. 4, 2023.

[23] Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv preprint arXiv:2010.12829*, 2020.

[24] Samuel Yen-Chi Chen, Chao-Han Huck Yang, Jun Qi, Pin-Yu Chen, Xiaoli Ma, and Hsi-Sheng Goan, "Variational quantum circuits for deep reinforcement learning," *IEEE Access*, vol. 8, pp. 141007–141024, 2020.

[25] Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms," *Advanced Quantum Technologies*, vol. 2, no. 12, pp. 1900070, 2019.

[26] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[27] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv preprint arXiv:1804.04262*, 2018.

[28] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.

[29] Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, et al., "Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv preprint arXiv:2109.00535*, 2021.

[30] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, M Sohaib Alam, Shahnawaz Ahmed, Juan Miguel Arrazola, Carsten Blank, Alain Delgado, Soran Jahangiri, et al., "Pennylane: Automatic differentiation of hybrid quantum-classical computations," *arXiv preprint arXiv:1811.04968*, 2018.

[31] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.