A Compact End-to-End Model with Local and Global Context for Spoken Language Identification

Fei Jia, Nithin Rao Koluguri, Jagadeesh Balam, Boris Ginsburg

NVIDIA Corporation, United States

`fjia, nkoluguri, jbalam, bginsburg@nvidia.com`

Abstract

We introduce TitaNet-LID, a compact end-to-end neural network for Spoken Language Identification (LID) that is based on the ContextNet architecture. TitaNet-LID employs 1D depthwise separable convolutions and Squeeze-and-Excitation layers to effectively capture local and global context within an utterance. Despite its small size, TitaNet-LID achieves performance similar to state-of-the-art models on the VoxLingua107 dataset while being 10 times smaller. Furthermore, it can be easily adapted to new acoustic conditions and unseen languages through simple fine-tuning, achieving a state-of-the-art accuracy of 88.2% on the FLEURS benchmark. Our model is scalable and can achieve a better trade-off between accuracy and speed. TitaNet-LID performs well even on short utterances less than 5s in length, indicating its robustness to input length.

Index Terms: spoken language identification, LID, VoxLingua107, fleurs

1. Introduction

Spoken Language Identification (LID) is an important preprocessing step in various applications such as Automatic Speech Recognition (ASR) and speech translation. It requires robust performance while introducing minimal latency in the pipeline. There are two important characteristics of LID models that have been rarely investigated. Firstly, LID models must be able to adapt easily to unseen languages and be robust to unseen domains. Secondly, while the performance of the LID task is heavily influenced by speech duration and the amount of information present [1], the model should still exhibit reliable performance on short speech segments.

Early approaches to LID mainly relied on acoustic, phonetic, prosodic, morphologic, and semantic level representations [2]. However, recent advancements in deep neural networks have led to better performance [3, 4, 5, 6, 7, 8, 9] compared to traditional systems [10]. For instance, [11] explores using Long Short Term Memory (LSTM) for LID task on NIST LRE 2009, and demonstrates that it can exploit temporal dependencies effectively. [12] conducts experiments using bottleneck features from Convolutional Neural Network (CNN) on noisy RATS Radio traffic data [13] and shows it has consistent improvement regardless of various acoustic conditions and durations. [14] proposes the addition of a time-frequency attention mechanism to Time Delay Neural Network architecture (TDNN) and CNN-LSTM-TDNN models to capture longer temporal dependencies, showing that frequency attention is more effective than time attention. More recently, ECAPA-TDNN [15] further improves TDNN performance on speaker tasks by introducing additional skip connections. This architecture has been adopted to LID task, referred to as ECAPA-

TDNN-LID [16] below, training on VoxLingua107 [17] and achieving 6.7% error rate on the evaluation set.

Although CNN-based models have been effective in speech processing tasks, their limited kernel size often results in only capturing local context. This can be particularly detrimental for the LID task, where the amount of available context is critical. Incorporating global context models may address this limitation and improve the performance of CNN-based LID models. Among several proposed methods [15, 18, 19, 20], ContextNet, proposed by [20], builds upon previous work [21] and incorporates a Squeeze-and-Excitation (SE) layer [22] to compress a sequence of local feature vectors into a single global context vector. This global context vector is then broadcasted back to each local feature vector and combined via multiplications. In the ASR domain, [20] demonstrated that incorporating global context improved accuracy. Another model, TitaNet [23], utilizes the encoder of the ContextNet model as a top-level feature extractor, and applies an attentive pooling layer to capture utterance-level speaker representation, achieving outstanding performance in speaker recognition and diarization tasks. In this study, we aim to explore whether this encoder architecture can also enhance the performance for LID task.

Another approach for LID is to first learn a good language representation vector, which can then be used for language category classification. To learn such a vector, one can can use self-supervised learning (SSL). Large scale speech understanding models like XLS-R (0.3B) [24] and w2v-bert-51 (0.6B) [25] are able to achieve great results by pre-training with more than 400k unlabeled data and then fine-tuning on LID datasets. More recently, [26] feeds the outputs of XLS-R (0.3B) model through an attentive pooling layer [27], achieving state-of-the-art (SOTA) results of 4.7% on VoxLingua107. Though SSL-pretrained models show excellent performance, their enormous model size makes it challenging to deploy them in compute-constrained scenarios.

In this work we make the following contributions:

- We propose *TitaNet-LID*, a compact end-to-end neural network (NN) for LID, which is based on ContextNet encoder, combining local features from 1D depth-wise separable convolutions and global context from Squeeze and Excitation (SE) layers [22]. The decoder contains a statistic pooling layer followed by two linear layers. We open-source the model through NVIDIA NeMo [28].[1]

- On VoxLingua107 dataset, we show that the model matches SOTA performance among other models that are trained from scratch. Its accuracy is similar to XLS-R model which uses SSL pretraining. Note that TitaNet-LID is 10X smaller than XLS-R.
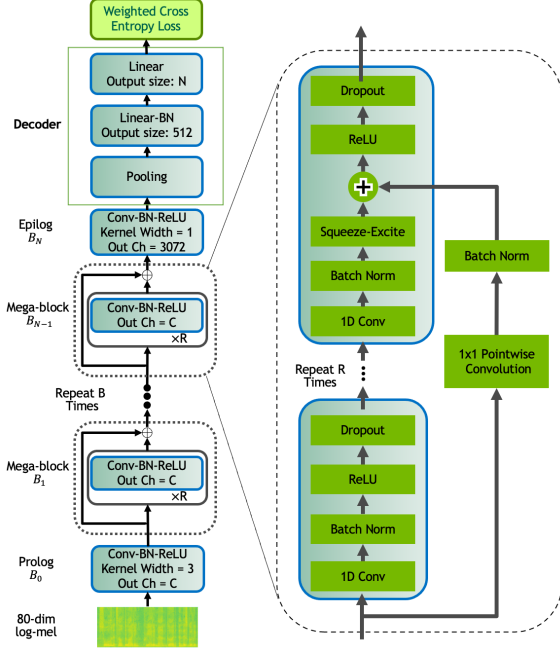
---

[1] `https://github.com/NVIDIA/NeMo`

Figure 1: *TitaNet-LID-$BxRxC$ Encoder is based on ContextNet architecture [20], where B is the number of blocks, R is the number of repeated "basic" blocks and C is the number of filters in the convolution layers of each block.*

- We examine TitaNet-LID's adaptation capability to unseen languages and new acoustic conditions. On the FLEURS dataset, TitaNet-LID outperforms the previous SOTA model by 16.8% while using 20 times lesser number of parameters.
- We study the effect of segment length on accuracy and demonstrate that model performs well even on short $\leq 5s$ segments.
- We show that the model can be easily scaled up or down for better accuracy or lower latency respectively.

2. MODEL ARCHITECTURE

As illustrated in Figure 1, TitaNet-LID-$BxRxC$ consists of an encoder which is a 1D depth-wise channel separable convolutional model with a ContextNet-like [20, 23] architecture and a decoder. The model's encoder includes B blocks, each block comprising R repeated "basic" blocks and C filters in the convolution layers of each block.

The TitaNet-LID encoder starts with a prologue block B_0, and followed by residual mega blocks $B_1...B_{N-1}$. Each mega block consists of R "basic" blocks and a squeeze-excite module in the end. A basic block is combined from 1D time-channel separable convolutional [21] (noted as 1D Conv in Figure 1) module with Kernel K, followed by BatchNorm, ReLU, and Dropout. Each 1D Conv module consists of two parts: a depth-wise convolutional layer and a 1x1 pointwise convolutional layer. These basic blocks are repeated R times and connected residually with SE [22] layers with global average pooling as in [23, 29]. The kernel K in repeated "basic" blocks are 7,11,15. The encoder ends with an epilogue block B_N and outputs intermediary audio features.

The encoded audio features are passed to the decoder to perform classification. The decoder contains two parts: a statis-tics pooling layer (mean and std) [30] to map variable length of input audio features to fixed-length feature representation, and two linear layers, one of output size 512 and another for a linear transformation from 512 to the final number of classes N.

3. Experiment Setup and Results

3.1. Train on VoxLingua107

3.1.1. Dataset

VoxLingua107 [17] is a large dataset designed for LID task which contains diverse YouTube data for 107 languages. The dataset is quite noisy with samples containing synthetic speech, singing, yelling, etc. The size of the official training set is 6628 hours, 62 hours per language on average, but it's highly imbalanced.

We split 10% of data of each language in training set into validation set. The input utterances have been split into non-overlap 3-second segments resulting in 5.9M and 656K segments in train and validation set respectively. The dataset comes with an official evaluation set of 1609 verified utterances covering 33 languages. The duration of evaluation samples varies from 1.7 seconds to 19.98 seconds. From 1609 samples, 253 are found to be less than 5 seconds.

3.1.2. Training setup

The audio segments are pre-processed into 80-dimensional log-melspectrograms calculated using 25ms window with stride of 10ms. Due to the extreme imbalance of the number of samples for each language in training data, the TitaNet-LID model is trained with weighted cross entropy loss. The weight of each class is calculated from data using

$$w_i = \frac{\sum_{n=0}^{N} c_n}{c_i}, \tag{1}$$

where N is number of classes and c_i is number of samples of the given class i. The weights are then normalized to sum up to one. To value the minority class, macro accuracy which computes the accuracy for each class and then take the average are used during validation to select the best checkpoint for each run.

All models were trained for 40 epochs on 4 nodes with 8 GPUs in each node with a batch size of 128 per GPU. We utilized the Adam optimizer and Cosine Annealing learning rate scheduler with a warm-up ratio of 10%. A maximum learning rate of 0.001 and a minimum learning rate of 0.0001 were used. We apply noise augmentation, speed perturbation with 0.95x and 1.05x, and RIR impulse corpora [31] perturbation as well as SpecAugment [32].

3.1.3. Results

We compare TitaNet-LID with two other models trained from scratch: Resnet34 [17] and ECAPA-TDNN-LID [16]. We also add a comparison with three pre-trained self-supervised learning (SSL) models finetuned on LID task: wav2vec2.0 [24], XLS-R [24] and XLS-R-attentive [26]. XLS-R [24] is a cross-lingually wav2vec 2.0 model which has been pretrained on 436K hours of 128 languages data including VoxLingua107, and it has 300M parameters. It achieves 5.7% error rate on the official evaluation set. XLS-R-attentive [26] extends XLS-R with an additional attentive pooling layer and reaches 4.7% error rate. We have also experimented with attentive pooling layer in [23] instead of statistic pooling layer but we don't observe performance gain.

Model	params	error rate (%)		
	(M)	0...5s	5...20s	avg
trained from scratch				
Resnet34 [17]	21	12.3	6.1	7.1
ECAPA-TDNN-LID [16][2]	21	11.9	6.7	7.5
TitaNet-LID-3x5x512	12	10.7	6.3	7.0
TitaNet-LID-3x5x1024	29	**7.5**	5.2	5.6
finetuned from SSL				
wav2vec2.0 [24]	300	11.5	6.3	7.2
XLS-R [24]	300	9.1	**5.0**	5.7
XLS-R-attentive [26]	300	-	-	**4.7**

Table 1: *Error rate on VoxLingua107 evaluation set. We compare TitaNet-LID with (1) two other models trained from scratch: Resnet34 and ECAPA-TDNN-LID; (2) three pre-trained self-supervised learning (SSL) models finetuned on VoxLingua107: wav2vec2.0, XLS-R and XLS-R-attentive. Accuracy numbers are computed for different utterance length categories and on all samples.*

In Table 1, we show that TitaNet-LID-3x5x1024 achieves 5.6% error rate on all samples, and it achieves 7.5% and 5.2% error rate for the short and long recordings respectively. TitaNet-LID gets SOTA performance for models trained from scratch and it has accuracy close to pretrained SSL models while being ten times smaller. The lowest error rate on 0-5 seconds samples also suggests that TitaNet-LID is robust even on relatively short samples. Smaller model TitaNet-LID-3x5x512 with 12M parameters also outperform Resnet34 and ECAPA-TDNN-LID with almost half size comparing to those models. TitaNet-LID which employs 1D depth-wise separable convolutions and SE layers could effectively capture local and global context within an utterance which is important for LID task, therefore, it can achieve better result even with less number of parameters.

In Figure 2, we demonstrate the number of the wrongly identified samples of TitaNet-LID-3x5x1024 in each language in the evaluation set with bright blue ink. And we can notice that the most wrongly predicted languages are Urdu, English, and Norwegian. Among the errors, most of them occur between closely related languages such as "Urdu → Hindi", "Norwegian → Norwegian Nynorsk" and "Spanish → Galician". The English samples that have been wrongly identified as Welsh could possibly be due to the bad quality of the retrieved Welsh data as stated in [17].

3.2. Finetune on FLEURS

Trained on VoxLingua107, TitaNet-LID has seen 82 out of 102 languages from FLEURS. It has 77.9% macro accuracy on these seen languages. To investigate the ability to adapt TitaNet-LID to unseen languages and acoustic conditions, we conduct two experiments: fine-tune TitaNet-LID-3x5x1024 1) on FLEUR training set, and 2) on the union of training sets of FLEURS and VoxLingua107.

3.2.1. Dataset

FLEURS [33] was created by recording 2009 sentences from FLoRes-101 benchmark by three different native speakers for
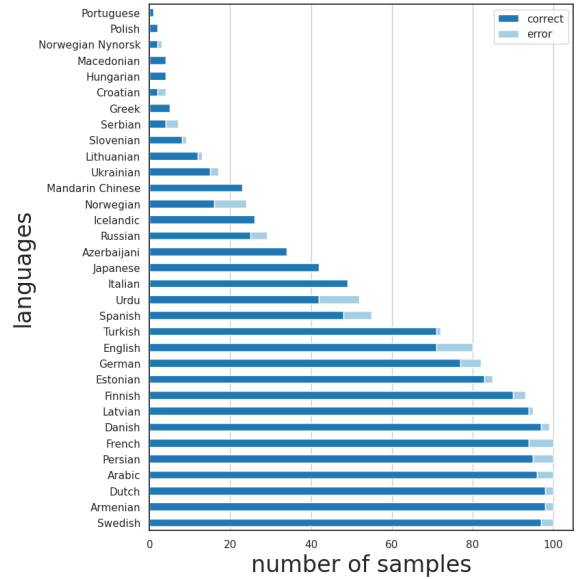
Figure 2: *Number of samples for each language in VoxLingua107 evaluation set. Dark blue represents accurately identified samples of TitaNet-LID-3x5x1024.*

each language. It contains 102 languages. It has official split for LID task where speakers of the train sets are different than speakers from the dev/test sets. The sizes of training, dev, and test are 987h, 120h, 283h respectively. After splitting the utterances into non-overlapping 3-second segments, there are 678K and 83K segments in the FLEURS training and validation sets, respectively. The training sets of FLEURS and VoxLingua107 together encompass 127 languages, while the test sets cover 103 languages. It is worth noting that VoxLingua107 is more diverse and noisy than FLEURS.

3.2.2. Fine-tuning setup

We freeze TitaNet-LID encoder and change number of output classes in decoder to 102 (FLEURS) and 127 (VoxLingua107 union FLEURS). During fine-tuning, the weight of loss is calculated by occurrence of each class the dataset(s). Model is fine-tuned for 10 epochs on a node with 2 GPUs. We use the same hyper-parameters as for training step with VoxLingua107, except: a peak learning rate of 5e-5, a dropout of 0.1, and speed perturbation for augmentation.

3.2.3. Results

We compare the finetuned models with wav2vec-BERT model, w2v-bert-51 which has 600M parameters and present result in Table 2. w2v-bert-51 [25] was first SSL-pretrained on 429k unlabeled data from 51 languages, and then fine-tuned on FLEURS. It achieves 71.4% macro accuracy on FLEURS test set. Our TitaNet-LID-3x5x1024 (29M) achieves a 16.8% higher accuracy than the 20x larger SSL pre-trained model w2v-bert-51 (0.6B) on the FLEURS test set. Furthermore, we demonstrate that by finetuning our model on the union of FLEURS and VoxLingua107, it can achieve an even higher macro accuracy of 93.8% on the 103-languages test set, and is capable of predicting 127 languages.

Table 3 shows the most common classification errors observed in the evaluation set in finetuning on FLEURS only ex-

model	params (M)	# languages			macro acc (%)
		pre-train	fine-tune	test set	
w2v-bert-51 [25]	600	51	102	102	71.4
TitaNet-LID	29	107	102	102	88.2
TitaNet-LID	29	107	127	103	93.8

Table 2: *TitaNet-LID-3x5x1024 finetuned on FLEURS vs pretrained w2v-bert-51 fine-tuned on FLEURS. We also finetune the model on the union of FLEURS and VoxLingua107 (127 languages), achieving higher macro accuracy 93.8% on 103 languages test set; and the model has the capability to predict 127 languages.*

# samples	# errors	error	
		True → Pred	#
700	570	Serbian → Bosnian	478
728	345	Javanese → Indonesian	344
299	256	Urdu → Hindi	253
749	261	Malay → Indonesian	250
914	337	Croatian → Serbian	229
998	540	Occitan → Lingala	214
946	212	Asturian → Galician	190
925	209	Bosnian → Croatian	189
927	247	Galician → Catalan	170
980	167	Sindhi → Hindi	149

Table 3: *Most common errors when finetuning on FLEURS only.*

periment. Similar to the observation in VoxLingua107, most of the errors come from related languages such as "Serbian, Bosnian, and Croatian". The errors such as misclassifying Occitan to be Lingala could possibly be due to the limitation of the model or the inaccuracy of the training and evaluation data.

4. Ablation study

Since LID model is frequently used as a pre-processing step for ASR, it is expected to introduce less latency and extra processing time as possible, especially for online mode. Hence both the input segment length for inference and the model size are crucial parameters.

4.1. Input length for inference

Typically, the longer the segments, the more language information is presented thus leading to higher accuracy of identification. In this section, we investigate how the input length would impact the performance of the model. TitaNet-LID is able to classify the input segments with different lengths by mapping variable-length utterances to a fixed-length representation. In Figure 3, we split the evaluation samples into segments of different lengths with stride 2, resulting less number of samples with large segment length. It can be seen that even though the model was trained with 3s segments, the model can still perform well and even better with segments of larger lengths.

4.2. Model scalability

The depth and width of TitaNet-LID could be increased or decreased by number of repeated layers R and changing filter size
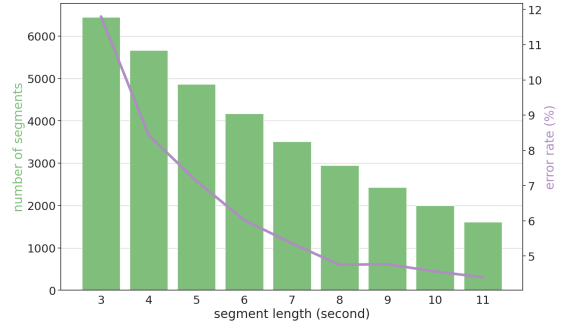

Figure 3: *The error rate of TitaNet-LID-3x5x1024 on VoxLingua107 evaluation set with different segment length. Generally the larger the input length, the better the model performs.*

R	C	Params (M)	val macro acc (%)	eval error rate (%)
3	256	7.3	89.8	9.6
5	256	7.7	90.7	8.4
7	256	8.1	92.1	9.2
5	512	12.3	93.5	7.0
5	1024	28.9	95.0	5.6
5	2048	92.1	95.9	6.8

Table 4: *TitaNet-LID: scalablity study. We use the same number of blocks $B = 3$ for all experiments and change model depth using number of sub-blocks R and model width with number of channels C respectively.*

C. In this section, we report both validation macro accuracy and evaluation error rate because as shown in Figure 2, the evaluation set is quite small and the distribution of it is imbalanced. As seen in Table 4, the validation macro accuracy would increase alongside the model size. The inconsistent trend of error rate could possibly be due to the limitation of evaluation set. TitaNet-LID-3x5x512 with 12.3M parameters outperforms the baseline model ECAPA-TDNN-LID with 21M parameters. The model could be easily scaled down for faster inference speed for constrained deployment conditions or scaled with for better performance.

5. Conclusion

In this paper, we present TitaNet-LID, an end-to-end model for language identification. It combines 1D depth-wise separable convolutions and Squeeze-and-Excite mechanism with global context to extract intermediary dimensional vector representations. The representations are then sent to a classifier to identify the language of the input audio.

TitaNet-LID achieves SOTA accuracy on the VoxLingua107 dataset among models trained from scratch. It is close to SOTA for SSL-pretrained models while being 10x fewer parameters. This makes TitaNet-LID an attractive model for memory or compute-constrained environments. We also show that TitaNet-LID can be adapted to unseen languages and new acoustic conditions with simple fine-tuning. It achieves new SOTA result on FLEURS.

The model's implementation and checkpoint are made available through NVIDIA NeMo.

6. References

[1] M. Van Segbroeck, R. Travadi, and S. S. Narayanan, "Rapid language identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1118–1129, 2015.

[2] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, pp. 1136–1159, 05 2013.

[3] H. Liu, L. P. G. Perera, A. W. H. Khong, S. J. Styles, and S. Khudanpur, "Pho-lid: A unified model incorporating acoustic-phonetic and phonotactic information for language identification," in *Interspeech*, 2022.

[4] H. Mazzawi, X. Gonzalvo, A. Kracun, P. Sridhar, N. Subrahmanya, I. Lopez-Moreno, H.-J. Park, and P. Violette, "Improving keyword spotting and language identification via neural architecture search at scale." in *Interspeech*, 2019, pp. 1278–1282.

[5] P. Shen, X. Lu, S. Li, and H. Kawai, "Interactive learning of teacher-student model for short utterance spoken language identification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5981–5985.

[6] P. Shen, X. Lu, and H. Kawai, "Transducer-based language embedding for spoken language identification," in *Interspeech*, 2022.

[7] T. M. Bartley, F. Jia, K. C. Puvvada, S. Kriman, and B. Ginsburg, "Accidental learners: Spoken language identification in multilingual self-supervised models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] W. Cai, D. Cai, S. Huang, and M. Li, "Utterance-level end-to-end language identification using attention-based cnn-blstm," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 5991–5995.

[9] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Speaker Odyssey*, 2018.

[10] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *ICASSP*, 2014.

[11] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. González-Rodríguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *INTERSPEECH*, 2014.

[12] S. Ganapathy, K. Han, S. Thomas, M. Omar, M. Van Segbroeck, and S. Narayanan, "Robust language identification using convolutional neural network features," in *INTERSPEECH*, 2014.

[13] K. Walker and S. M. Strassel, "The RATS radio traffic collection system," in *Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore, June 25-28, 2012*, H. Li, B. Ma, and K. Lee, Eds. ISCA, 2012, pp. 291–297. [Online]. Available: http://www.isca-speech.org/archive/odyssey_2012/od12_291.html

[14] X. Miao, I. McLoughlin, and Y. Yan, "A New Time-Frequency Attention Mechanism for TDNN and CNN-LSTM-TDNN, with Application to Language Identification," in *Proc. Interspeech*, 2019.

[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *21st Annual conference of the International Speech Communication Association (INTERSPEECH 2020)*. International Speech Communication Association (ISCA), 2020, pp. 3830–3834.

[16] 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.

[17] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *IEEE SLT*, 2021.

[18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

[19] M. H. Radfar, R. Barnwal, R. Swaminathan, F.-J. Chang, G. P. Strimel, N. Susanj, and A. Mouchtaris, "Convrnn-t: Convolutional augmented recurrent neural network transducers for streaming speech recognition," in *Interspeech*, 2022.

[20] 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 *Proc. Interspeech 2020*, 2020, pp. 3610–3614. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2059

[21] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP*, 2020.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE CVPR*, 2018.

[23] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP*, 2022.

[24] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. M. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech*, 2021.

[25] A. Conneau, A. Bapna, Y. Zhang, M. Ma, P. von Platen, A. Lozhkov, C. Cherry, Y. Jia, C. Rivera, M. Kale, D. van Esch, V. Axelrod, S. Khanuja, J. Clark, O. Firat, M. Auli, S. Ruder, J. Riesa, and M. Johnson, "Xtreme-s: Evaluating cross-lingual speech representations," in *Interspeech*, 2022.

[26] K. Kukk and T. Alumäe, "Improving language identification of accented speech," in *Interspeech*, 2022.

[27] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification," in *Proc. Interspeech*, 2018.

[28] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building ai applications using neural modules," *arXiv:1909.09577*, 2019.

[29] S. Majumdar, J. Balam, O. Hrinchuk, V. Lavrukhin, V. Noroozi, and B. Ginsburg, "Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition," *arXiv:2104.01721*, 2021.

[30] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*. IEEE, 2018.

[31] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017.

[32] 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," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[33] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2022.