TITANET: NEURAL MODEL FOR SPEAKER REPRESENTATION WITH 1D DEPTH-WISE SEPARABLE CONVOLUTIONS AND GLOBAL CONTEXT

Nithin Rao Koluguri, Taejin Park, Boris Ginsburg

NVIDIA, USA

ABSTRACT

In this paper, we propose TitaNet, a novel neural network architecture for extracting speaker representations. We employ 1D depth-wise separable convolutions with Squeeze-and-Excitation (SE) layers with global context followed by channel attention based statistics pooling layer to map variable-length utterances to a fixed-length embedding (t-vector). TitaNet is a scalable architecture and achieves state-of-the-art performance on speaker verification task with an equal error rate (EER) of 0.68% on the VoxCeleb1 trial file and also on speaker diarization tasks with diarization error rate (DER) of 1.73% on AMI-MixHeadset, 1.99% on AMI-Lapel and 1.11% on CH109. Furthermore, we investigate various sizes of TitaNet and present a light TitaNet-S model with only 6M parameters that achieve near state-of-the-art results in diarization tasks.

Index Terms— speaker verification, speaker embedding, t-vectors, context, diarization

1. INTRODUCTION

Speaker recognition is a broad research area that solves two major tasks based on characteristics of voices: speaker identification and speaker verification. Speaker identification is about identifying a person and speaker verification is about verifying whether the speaker is who they claim to be. Speaker diarization is a task of partitioning audio recordings into speaker-homogeneous segments belonging to each individual speaker. Typically, speaker recognition and diarization systems operate on unconstrained speech utterances, which are converted to a vector of fixed length, called speaker embeddings. These speaker embeddings represent the identity of each speaker and are used for speaker recognition and speaker diarization tasks.

In recent years, deep neural networks (DNNs) have been actively employed for speaker embedding extractors since d-vector [1] was proposed. Subsequently, x-vector [2] was widely used because of the superior performance achieved by employing statistical pooling and time delay neural network (TDNN). Other architectures such as ResNet-based convolutional neural networks (CNNs) [3] and CNNs with cross convolutional layers [4] were employed for capturing the traits of speech. In addition, to cope with the variable-length

inputs, Transformer [5], CNN-LSTM [6] and a slew of variants of TDNN [7, 8, 9] were applied for DNN-based speaker embedding extractors. In this work, we develop a speaker embedding extractor model that shows superior performance on both speaker verification and diarization tasks. We adopt the architecture from ContextNet [10], a state-of-the-art automatic speech recognition (ASR) model, which combines local features from 1D depth-wise separable convolutions and global context from Squeeze and Excitation layers. We train text-independent speaker recognition models where the identity of the speaker is based on how speech is spoken, not necessarily on what is being said. The following are the main contributions of this paper:

- We propose use of 1D separable depth-wise convolutions compared to full 1D convolutions.

- We bring in global context to speaker embedding models by introducing global average pooling after the squeeze and excitation module. This contrasts with the embedding extractors based on TDNN, such as x-vector [2] or most recently ECAPA-TDNN [11].

- TitaNet-M is half the size of comparable speaker embedding extractors like ECAPA-TDNN or Conformer-based baselines and achieves superior performance in speaker diarization.

- We train our networks end-to-end using angular softmax margin loss and use cosine similarity as a backend for speaker representations. Such approach leads us to avoid the burden of training external models like probabilistic linear discriminant analysis (PLDA) [12] and agglomerative hierarchical clustering (AHC) [13] as in well-known speaker verification systems [2, 7], or speaker diarization systems [13, 14].

We introduce three models with different sizes. The architecture is easily scalable in both depth and width by design, and we show that the scaling in width is very effective in reducing the model size with a small change in performance. In the experimental section, we demonstrate the performance of the model on a VoxCeleb1 cleaned test speaker verification trial file. In addition, we evaluate diarization performance on popular evaluation datasets like AMI (Lapel), AMI (Mix-Headset) [15], NIST-SRE-2000 [16] and CH109 [17].
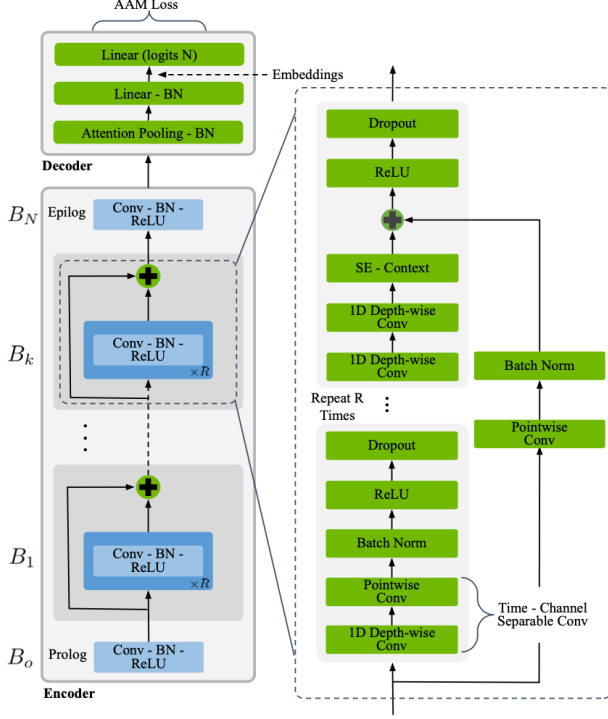
Fig. 1. TitaNet Encoder and Decoder Architecture

2. MODEL ARCHITECTURE

2.1. Encoder

The model is based on the ContextNet ASR architecture [10] comprising of an encoder and decoder structure. We use the encoder of the ContextNet model as a top-level feature extractor, and feed the output to the attentive pooling layer. This layer computes attention features across channel dimensions to capture the time-independent utterance-level speaker representations.

TitaNet is a 1D time depth-wise channel separable convolutional model with ContextNet-like architecture combined with channel attention pooling. Fig. 1 describes the ContextNet-$B \times R \times C$ model encoder and attention pooling decoder, where B is the number of blocks, R is the number of repeated sub-blocks per block, and C is the number of filters in the convolution layers of each block. The encoder starts with a prologue block B_0, followed by mega blocks B_1 ... B_{N-1} and ends with epilogue block B_N. Prologue and epilogue blocks differ from mega blocks, they both have the same convolution module (Conv), batchnorm and relu layers and have fixed kernel sizes of 3 in prologue and 1 in epilogue for all the network architectures we propose. They do not contain residual connections and dropout layers. Each mega block begins with time-channel separable convolutional[18] layer with stride 1 and dilation 1, followed by batchnorm, relu and dropout.

Each time-channel separable convolution module is made

up of two parts: a depth-wise convolutional layer and a pointwise convolutional layer. Depth-wise convolutions apply a single filter per input channel (input depth). Pointwise convolutions are 1×1 convolutions, used to create a linear combination of the outputs of the depth-wise layer. These layers are repeated R times, which can be modified to vary the depth of the network. These repeated layers are residually connected with Squeeze and Excitation layers with global average pooling for context inclusion. By using global context, the SE layer squeezes a sequence of local feature vectors into a single global context vector, broadcasts this context back to each local feature vector, and merges the two via multiplications. The width of the network can be increased or decreased by varying output channel filter sizes of each mega block. For TitaNet models, width and depth are changed by varying these filter sizes, C and the number of repeated layers, R respectively.

2.2. Decoder and Embeddings

The top level acoustic features obtained from the output of encoder are used to compute intermediate features that are passed to the decoder for getting utterance level speaker embeddings. The intermediate time-independent features are computed using an attentive statistics pooling layer [9], where the channel attention features E are computed across time-channels to get a time-independent feature representation S of size $B \times 3072$.

The intermediate features S are passed through the Decoder consisting of two linear layers, one of output size 192 and another for a linear transformation from 192 to the final number of classes N, to compute the probability that the current segment belongs to a speaker from the training set. In this fashion, the network extracts fixed-length representation from variable length speech segments. We extract t-vectors before the final logits linear layer of fixed size 192.

2.3. Loss function

The TitaNet model was trained end-to-end with additive angular margin (AAM) loss [19]. The AAM helps to optimize the cosine distance between speaker embeddings. For all the verification and diarization experiments presented in this paper, we use cosine similarity as the back-end:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)} + \sum_{j=1, j \neq y_i}^{n} e^{s \cos \theta_j}} \quad (1)$$

where m is margin, s is scale and θ_j is the angle between the final linear layer weight W_j and incoming feature x_i. Here m and s are predefined hyper parameters.

3. EXPERIMENTS

We designed three TitaNet models: TitaNet-S with 256 channels, TitaNet-M with 512 channels, and TitaNet-L with 1024

Dataset	# of Speakers	Duration (in Hrs)	# Utterances (in K)
VoxCeleb1	1211	227	332
VoxCeleb2	5994	1895	2274
SRE	3787	503	944
Fisher	951	162	278
Switchboard	2400	247	425
LibriSpeech	2338	336	634
Total	16681	3373	4890

Table 1. Statistics of each dataset used for training TitaNet

channels. All models have the same number of repeating layers R as 3, and the same kernel filter sizes as 3,7,11 and 15. TitaNet-S has 6.4M , TitaNet-M – 13.4M, and TitaNet-L – 25.3M parameters.

3.1. Datasets

3.1.1. Training Data

We use following datasets to datasets to train TitaNet: VoxCeleb1 and VoxCeleb2 dev[20], NIST SRE portion of datasets from 2004-2008 (LDC2009E100), Switchboard-Cellular1 and Switchboard-Cellular2 [21], Fisher [22], and Librispeech [23] (see Table 1). Combined, these datasets consist of about 4.8M utterances from 16.6K speakers. We augment the training data with RIR [24] impulse corpora, speed pertubation with 0.95x & 1.05x speeds and also spec augment [25].

3.1.2. Evaluation Data

We use the VoxCeleb1 cleaned test trial file to evaluate EER for speaker verification. We use the following three datasets for the evaluating speaker diarization system:
- NIST-SRE-2000 [16]: all sessions from LDC2001S97.
- AMI Corpus [15]: Lapel and MixHeadset audio subsets from partition set [26].
- CH109 [17]: we use a subset of CALLHOME American English speech (CHAES), which contains only two speakers. There are 109 sessions in this subset. The remaining 11 sessions in CHAES are used as a dev set for CH109 and NIST-SRE-2000.

3.2. Experiment Setup

Every speaker recognition experiment consists of common data pre-processing steps for training, development, and evaluation steps. During the pre-processing, we do not use a speech activity detector (SAD) to avoid dependence on an additional model. Instead, we split speech segments longer than 3 sec into random chunks of 1.5, 2, and 3 sec. We compute acoustic features for every 25 ms frame window shifted over 10 ms. The acoustic features are 80-dimensional mel spectrograms computed using a 512 FFT and a Hann window. Next, Mel-spectrogram features are normalized over the frequency axis. Every utterance fed to the encoder has size
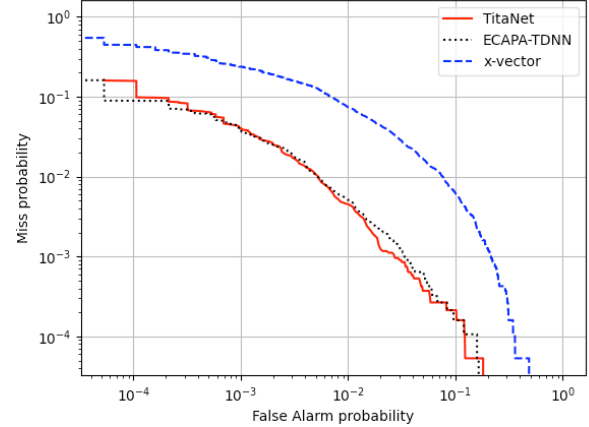

Fig. 2. DET curve for VoxCeleb1 cleaned trial comparing with previous studies

of $T \times 80$, where T is the number of frames in a given speech utterance file. The accuracy of speaker verification systems is measured using EER and minimum normalized detection cost (MinDCF) with $P_{target} = 10^{-2}$ and $C_{FA} = C_{Miss} = 1$. Both the EER for verification and the DER - for diarization are done using cosine similarity (CS) back-end.

In our diarization experiments, the evaluation datasets are divided into two groups: telephonic and non-telephonic speech. Based on experiments with the dev sets, we found window size of 1.5 sec with a shift of 0.75 sec works best for telephonic speech. For non-telephonic speech, the best settings were 3 sec and 1.75 sec for window and shift respectively. In the evaluation datasets used in this paper, AMI Lapel and MixHeadset fall under non-telephonic speech, and the rest of them are in the telephonic speech group. Unlike the previous studies [9, 13], we do not use external dev data to tune the clustering parameters by relying on an auto-tuning approach [28]. Similar to the previous systems [9], we use collar 0.25 sec and ignore overlap speech regions for speaker error rate calculation. All TitaNet models in Table 2 are trained for 250 epochs with SGD optimizer, with initial learning rate (LR) 0.08 using cosine annealing LR scheduler on 4 nodes with 8 V100 GPUs per node.

Models (Backend)	# Params (M)	VoxCeleb1	
		EER (%)	MinDCF
x-vector (PLDA) [2]	9	2.97	0.323
ECAPA (CS) [9]	22.3	0.69	**0.082**
Conformer (CS)[27]	26.4	2.43	0.264
TitaNet-S (CS)	6.4	1.15	0.131
TitaNet-M (CS)	13.4	0.81	0.106
TitaNet-L (CS)	25.3	**0.68**	0.087

Table 2. TitaNet comparison with other models for speaker verification task. All models has been evaluated with Cosine Similarity (CS) backend except x-vector which used PLDA.

Models (Backend)	NIST-SRE 2000	AMI Lapel	AMI MixHeadset	CH109
x-vector (PLDA + AHC) [2]	8.39	-	-	9.72
ECAPA (SC) [9]	-	2.36	1.78	-
x-vector (MCGAN) [29]	**5.73**	-	-	-
TitaNet-S (NME-SC)	6.37	2.00	2.22	**1.11**
TitaNet-M (NME-SC)	6.47	**1.99**	1.79	1.13
TitaNet-L (NME-SC)	6.73	2.03	**1.73**	1.19

Table 3. TitaNet comparison with other models for speaker diarization with oracle SAD known speakers number, DER(%).

3.3. Evaluation Results

3.3.1. Speaker Verification

In the speaker verification experiments, we train the model on the datasets shown in the Table 1. We train these systems initially as a speaker identification model with 10 percent of audio files of each speaker set aside as validation data from training sets. With this setup, we trained TitaNet models end-to-end using additive margin angular loss. Table 2 shows the performance of TitaNet models on the VoxCeleb cleaned trial file. We observed a high-degree of sensitivity on validation curves with slight variations in the margin (m) and scale (s) for angular loss. With m as 30 and s as 0.2 TitaNet-L showed state of the art performance with EER of 0.68% on VoxCeleb1 cleaned test trial file outperforming previously reported results in [9, 30].

As it can be noticed from Table 2, Titanet models are easily scalable to achieve very competitive performances even with relatively few parameters around 6M. These models show a direct relationship on accuracy in contrast to the number of parameters. We show the Detection Error Trade-off (DET) curves to compare TitaNet-L model with other previously stated CNN based models.

3.3.2. Speaker Diarization

We employ our proposed speaker embedding extractor models for speaker diarization tasks. Cosine similarity is used for measuring the distance between speaker embeddings and Normalized Maximum Eigengap Spectral Clustering (NME-SC) [28] on the extracted embeddings to obtain the clustering result. We show the performance of each TitaNet model on popular evaluation datasets as shown in Tables 3 and 4. The diarization experiments are based on oracle SAD to evaluate the SAD-independent performance. In Table 3, we show the results for known number of speakers case and in Table 4, we present the results for unknown number of speakers for which the speaker count is estimated using NME-SC clustering algorithm.

TitaNet models outperform the previous state-of-the-art models on the AMI-Lapel, AMI-MixHeadset and CH109 evaluation datasets. It is worth noting that the performance of the small and medium TitaNet models show minor differences even if we reduce their model parameters by $2x$ and $4x$ respectively, compared to the largest model. In di-

Models (Backend)	NIST-SRE 2000	AMI Lapel	AMI MixHeadset	CH109
x-vector (PLDA + AHC) [2]	7.12	-	-	-
x-vector (VBx) [14]	**4.42**	-	2.17	-
ECAPA (SC)[9]	-	2.13	2.17	-
x-vector (MCGAN) [29]	6.76	-	4.92*	-
TitaNet-S (NME-SC)	5.49	2.3	1.97	**1.42**
TitaNet-M (NME-SC)	5.75	2.59	1.89	1.51
TitaNet-L (NME-SC)	5.38	**2.03**	**1.89**	1.63

∗This number is reported based on Kaldi evaluation partition [14]

Table 4. TitaNet comparison with other models for speaker diarization with oracle SAD estimated number of speakers, DER(%).

arization systems we believe there is no major performance improvement using larger TitaNet models when compared to verification. We hypothesize that this is related to the fact that separability in the embedding space does not require a higher level of precision since the clustering process only involves relatively few speakers. The TitaNet models show very good improvement on all datasets, except on the NIST-SRE-2000 evaluation set. But note that the clustering approaches in [14, 2] involves additional training for Hidden Markov Model or PLDA.

4. CONCLUSION

In this paper, we present TitaNet, a new speaker representation learning model that utilizes the global context of squeeze-and-excitation layers combined with channel attention pooling for extracting fixed length speaker embeddings. The model employs 1D depth-wise separable convolutions for speaker embedding models that showed state-of-the-art performance in ASR tasks. The TitaNet-M model, which is half the size of previous state-of-the-art systems outperforms them in speaker diarization tasks while achieving competitive numbers on verification tasks. The TitaNet-L model significantly outperforms existing models in speaker verification and diarization tasks.

The models' implementation and pre-trained checkpoints are made available through NVIDIA NeMo toolkit [31].[1]

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] E. Variani, X. Lei, E. McDermott, I. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *ICASSP*, 2014.

[2] D. Snyder et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

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

[3] Y. Yu, L. Fan, and W. Li, "Ensemble additive margin softmax for speaker verification," in *ICASSP*, 2019.

[4] Z. Gao, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An improved deep embedding learning method for short duration speaker verification," in *Interspeech*, 2018.

[5] P. Safari, M. India, and J. Hernando, "Self-attention encoding and pooling for speaker recognition," in *Interspeech*, 2020.

[6] J. Jung, H. Heo, et al., "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *ICASSP*, 2018.

[7] J. Villalba, N. Chen, D. Snyder, et al., "State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18.," in *Interspeech*, 2019.

[8] Y. Zhu and B. Mak, "Orthogonal training for text-independent speaker verification," in *ICASSP*, 2020.

[9] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "ECAPA-TDNN embedding for speaker diarization," *arXiv:2104.01466*, 2021.

[10] W. Han, Z. Zhang, Y. Zhang, et al., "ContextNet: Improving convolutional neural networks for automatic speech recognition with global context," *arXiv:2005.03191*, 2020.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[12] P. Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, 2010.

[13] G. Sell, D. Snyder, et al., "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge.," in *Interspeech*, 2018.

[14] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, pp. 101254, 2022.

[15] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005.

[16] A. Martin and M. Przybocki, "The NIST speaker recognition evaluations: 1996-2001," in *Odyssey*, 2001.

[17] A. Canavan, D. Graff, and G. Zipperlen, "CallHome american english speech," *Linguistic Data Consortium*, 1997.

[18] S. Kriman, S. Beliaev, B. Ginsburg, et al., "QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions," in *ICASSP*, 2020.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[20] J. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018.

[21] J. Godfrey and E. Holliman, "Switchboard-1 release 2 LDC97S62," *Linguistic Data Consortium*, 1993.

[22] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher english training speech part 1 transcripts," *Linguistic Data Consortium*, 2004.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

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

[25] D. Park, W. Chan, et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv:1904.08779*, 2019.

[26] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP*, 2020.

[27] A. Gulati, J. Qin, C. Chiu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020.

[28] T. Park, K. Han, et al., "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 38–385, 2019.

[29] M. Pal, M. Kumar, et al., "Meta-learning with latent space clustering in generative adversarial network for speaker diarization," *IEEE/ACM TASLP*, 2021.

[30] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "BUT system description to voxceleb speaker recognition challenge 2019," *arXiv:1910.12592*, 2019.

[31] 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.