

# SMALL FOOTPRINT TEXT-INDEPENDENT SPEAKER VERIFICATION FOR EMBEDDED SYSTEMS

Julien Balian, Raffaele Tavarone, Mathieu Poumeyrol, Alice Coucke

{firstname.lastname}@sonos.com  
Sonos Inc., Paris, France

## ABSTRACT

Deep neural network approaches to speaker verification have proven successful, but typical computational requirements of State-Of-The-Art (SOTA) systems make them unsuited for embedded applications. In this work, we present a two-stage model architecture orders of magnitude smaller than common solutions (237.5K learning parameters, 11.5MFLOPS) reaching a competitive result of 3.31% Equal Error Rate (EER) on the well established VoxCeleb1 verification test set. We demonstrate the possibility of running our solution on small devices typical of IoT systems such as the Raspberry Pi 3B with a latency smaller than 200ms on a 5s long utterance. Additionally, we evaluate our model on the acoustically challenging VOiCES corpus. We report a limited increase in EER of 2.6 percentage points with respect to the best scoring model of the 2019 VOiCES from a Distance Challenge, against a reduction of 25.6 times in the number of learning parameters.

**Index Terms**— speaker verification, neural networks, text independent, small footprint

## 1. INTRODUCTION

Speaker verification refers to the task of verifying a user identity based on their voiceprint. This technology has received increasing attention in recent years, partially due to its application to voice assistants. Speaker verification enables the contextualisation of spoken queries and tailored assistant responses to personalised content (*e.g.* “add an event to my calendar”).

The aggregation of the variable-length sequential audio input into a fixed length embedding plays a crucial role for any practical application. In the first approaches to time aggregation, the embeddings are computed over fixed segments of audio [1] with additional steps required to aggregate the representations. More recently, with the advent of neural networks, end-to-end solutions have been proposed that directly handle variable-length input [2]. Using properly designed layers, the temporal statistics can be accumulated internally in the network [2, 3, 4]. While reaching good accuracy, these end-to-end, neural-based methods have typical computational

footprints that require offline or server-side execution. Although some speaker verification engines with low execution latency [5] or the ability to run on mobile devices [6] have been proposed, they remain too large for embedded applications where memory and computing power are further limited.

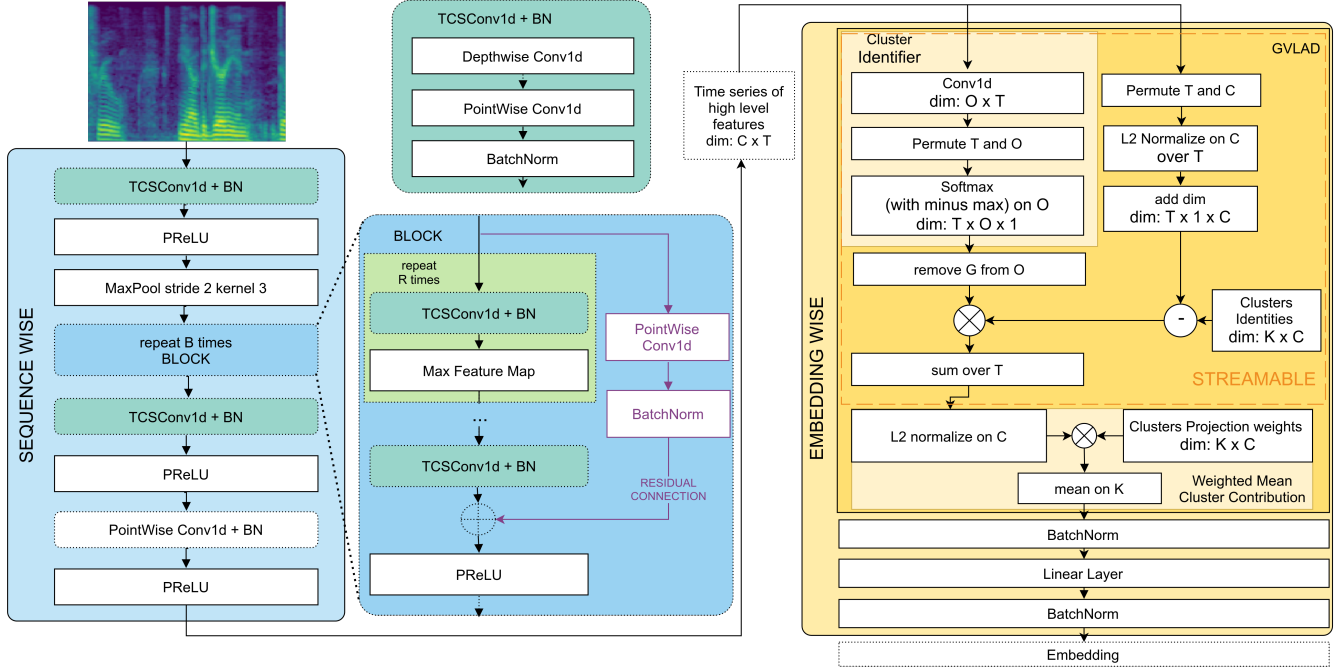
In this work, we propose a speaker verification system specifically tailored to embedded use cases. We budget CPU and memory resources to match that of typical keyword spotting systems [7] designed to run continuously and in real time on device. Our approach allows to decouple streamed time-series features extraction from aggregation, providing an optimal balance between representation quality and inference latency. The features extraction stage is based on the QuartzNet [8] model – never used in the context of speaker verification to our knowledge – and the aggregation stage on Ghost Vector of Locally Aggregated Descriptors (GVLAD) [9], with key modifications, *e.g.* the inclusion of Max Features Map [10] operations and a more computationally efficient method for descriptor aggregation.

We demonstrate that our approach reaches performances comparable with the state of the art (3.31% EER on the VoxCeleb1 verification test set) with a number of learning parameters orders of magnitude smaller, making it fit for embedded applications. Voice being a highly sensitive biometric identifier, our lightweight approach grants speaker verification abilities to small devices typical of IoT systems, while fully respecting the privacy of users.

The paper is structured as follows. In Section 2, we detail our system in terms of neural network architectures and training method. In Section 3, we describe our experimental settings including datasets and hyper-parameters selection, and computational performances. In Section 4, we compare our solution with other SOTA approaches. We give our conclusions in Section 5.

## 2. SPEAKER EMBEDDING MODEL

We propose a fixed-size speaker embedding model performed in two stages. In the first stage, a streaming neural network, inspired by the QuartzNet [8] architecture, takes as input an arbitrarily long time series of acoustic features and outputs



**Fig. 1:** Network architecture. Left panel: sequence-wise network (blue).  $B$  is the number of stacked blocks. In each block,  $R$  is the number of TCSCConv1d+BN modules followed by Max Feature Map activation stacked before a last TCSCConv1d+BN and residual connection. Right panel: embedding-wise network (yellow).  $T$  is the time dimension and  $C$  is the channel dimension. In the GV-LAD module,  $O = K + G$  is the total number of clusters, defined as the sum of the number of contributing clusters  $K$  and the number of Ghost clusters  $G$ .

another series of higher-level features. This network is referred to as sequence-wise network. The second stage consists of an aggregator neural network, called embedding-wise network, built upon the GV-LAD [8] architecture, that aggregates the outputs of the streaming stage along the time dimension to build a fixed-size embedding of the audio signal.

## 2.1. Model Architecture

The sequence-wise network of our solution is pictured in blue on the leftmost part of Figure 1. One basic component, taken from the QuartzNet [8] model, is a Time Channel Separable 1-dimensional Convolution (TCSCConv1d) module. It is composed of a 1-dimensional depthwise convolution, where each kernel operates only across the time axis, and a pointwise convolution acting on all filters but independently on each frame. The first TCSCConv1d module is followed by Batch Normalisation (BN) and Parametric ReLU (PReLU) [11]. The next layer performs max-pooling to halve subsequent computations. Successive blocks are also composed of TCSCConv1d followed by BN and a Max Features Map (MFM) operation [10] that, at each location, optimally selects the output of distinct filters. As shown in Figure 1, the TCSCConv1d+BN followed by MFM constitutes a block that is repeated  $R$  times with a residual connection and followed by a PReLU activation.

Time aggregation techniques are crucial for creating ef-

ficient embeddings. Historically, aggregation was performed using a simple mean pooling mechanism [12], later refined by statistical pooling [3] and attention mechanisms [4, 13], or using that last output of the recurrent cells [2]. Statistical pooling being too demanding for our constrained budget, we compared Self Attentive Pooling (SAP), recurrent cells, and GLVAD and got better results with the latter. To further limit the computations, we replace the last linear projection in GV-LAD by a simple cluster-wise projection averaged on the cluster dimension  $K$ . This comes with almost no performance drop, while making the cost of adding a new cluster linear instead of quadratic in the number of learning weights in the projection layer. The rightmost part of Figure 1 (in yellow) displays a detailed view of the embedding-wise network.  $O = K + G$  is the total number of clusters and is defined as the sum of the numbers of clusters  $K$  contributing to the embedding and the number of "Ghost" clusters  $G$ , named that way because they are not included in the final concatenation (see [9] for more details).

## 2.2. Training loss function

Speaker recognition can be seen as a metric learning or a classification problem; both approaches have been shown to be successful. We explored triplet loss techniques for metric learning (like soft triplet loss based on softmax [14]), and cross-entropy based methods for classification, like angular

prototypical [6] or ArcFace [15] losses. We found that ArcFace was the most efficient, especially when coupled with focal loss [16].

### 3. EXPERIMENTS

#### 3.1. Datasets

The proposed approach is evaluated on two datasets. We first train our network on the VoxCeleb2 [17] dev split only. This dataset contains 5994 speakers of 145 different nationalities, with over 1,092,009 utterances. We compared this model with previously published papers in Section 4.1 using the VoxCeleb1 verification test, following the protocol proposed by [17].

We also train an additional model using the same architecture, but augmenting the training data with the MUSAN [18] dataset, a corpus of music, speech, and noise. Reverberation effects are obtained by applying rooms simulation from Pyroomacoustics [19]. We simulate over 5,000 rooms and apply the augmentation randomly eight times for each original audio sample with variation of gain, SNR, noise type, and noise source location. We also include a small amount of speed augmentation as a approximate means of accounting for within-speaker speech tempo variability. The model trained with augmented data is analysed in challenging acoustic conditions in Section 4.2 on the Voices Obscured in Complex Environmental Settings (VOiCES) corpus [20]. The corpus addresses challenging noisy and far-field acoustic environments known to strongly impact the final performance of speaker verification systems.

#### 3.2. Experimental setup

##### 3.2.1. Model implementation

The acoustic features are 64-dimensional Mel filterbank energies, extracted from the input audio every 10ms over a window of 20ms. Mean and variance normalisation (MVN) is applied but no Voice Activity Detection (VAD) pre-processing is done. We experimented with several combinations of depth and width of the architecture (while keeping the total number of parameters fixed) and converged to a configuration with a total of 22 TCSCov1d each with 96 filters ( $B = 5$  and  $R = 3$  in Figure 1). All the kernels have a constant size of 15, tuned to reach a good balance between performance, computational load and final size of the receptive field.

Compared with the original GVLAD work [9], we select a higher number of clusters  $K$  at the aggregation stage (see Section 2.1) as we find it yields better results. The number of Ghost clusters  $G$  seems to have little impact as soon as there are at least 3. We therefore set  $K = 32$  and  $G = 3$  for the following experiments. The embedding network has 237,499 learning parameters and an output dimension of 96. Despite our constrained budget, we do not apply compression

Network Part	FMA*	Div*	FLOPS	params
Sequence wise	10850.4	0	10.8M	211.6K
Embedding wise	659.7	8.0	0.7M	25.8K
TOTAL	11509.4	8.0	11.5M	237.5K

**Table 1:** Computational Inference Cost (\* K over 1s). *FMA*: fused multiply-add operation. *Div*: div operation. *FLOPS*: floating point operations per second. *params*: number of learning parameters.

techniques such as quantization, teacher-student, or pruning. We of course expect these refinements to further improve our model, but we choose to focus solely on optimizing the architecture in this paper.

##### 3.2.2. Training details

Each batch contains  $S$  speakers, each with  $N$  utterances of duration  $D$  in seconds. Best results are obtained with:  $S = 75$ ,  $N = 5$ , and  $D = 2\text{to}5$ .  $2\text{to}5$  means a random uniform sampling of each sample from 2 to 5 seconds. The same definition of training epoch as in [6] is used: every speaker in the dataset is seen once. Given the memory constraints, we train with mixed precision and auto scaling loss which we find useful to stabilise the training and avoid exploding gradients. We use the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.0005. A scheduler is applied to reduce the learning rate on loss plateau while monitoring the EER on the VoxCeleb2 test set. The scheduler patience is set to 250 epochs. The scaling and margin parameters of the ArcFace [15] loss function are respectively set to  $s = 15$  and  $m = 0.5$ . Masking on the time-dimension is applied on the sequence-wise part and during the first stages of the GVLAD process to account for utterances of varying lengths in a batch.

#### 3.3. Inference & Streaming

Separating the model in two distinct stages allows to set a decoupled memory and latency budget. Thanks to `tract`<sup>1</sup>, an open source neural network inference library, we are able to run the first stage in streaming mode and drastically reduce the delay after end-pointing. Some numbers for each stage of the proposed approach are displayed in Table 1.

Latency is a cornerstone to real-time applications of speaker verification systems. It depends on the audio duration to embed and the computational power given a fixed model. Both dimensions are displayed in Table 2. These results show that the system can run on a single core of Raspberry Pi3 B. This latency could be further improved by streaming part of the embedding-wise network (dashed orange box in the rightmost part of Figure 1). This will be the object of a future work.

<sup>1</sup>available at <https://github.com/sonos/tract> version used: 0.11.1

CPU & inference	1s	5s	10s
Intel i7-8750H - Batch	14.8	33.7	58.40
Intel i7-8750H - Stream	1.11	5.09	9.97
Raspberry Pi 3B - Batch	198.8	438.2	733.8
Raspberry Pi 3B - Stream	45.1	184.4	221.5

**Table 2:** Mean embedding latency in milliseconds on a mono-core CPU (Raspberry Pi running on 64-bit Ubuntu) for various audio lengths, *Batch* is the latency if we apply the whole embedding process once end-pointing is triggered, *Stream* is the latency when the first stage is performed in streaming with a Real Time Factor (RTF) lower than 1.

## 4. RESULTS

The proposed approach is evaluated by computing a cosine similarity score between embeddings on two datasets.

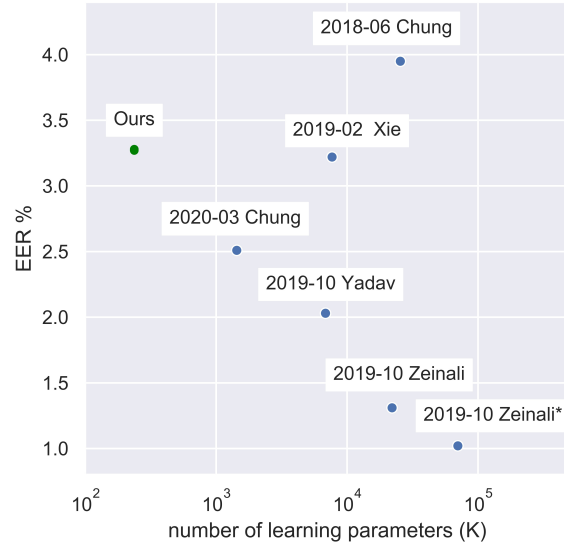
### 4.1. VoxCeleb1 verification test

The proposed model (trained on the VoxCeleb2 dev split) is compared to other existing works in terms of EER and number of parameters, following the VoxCeleb1 verification test protocol proposed in [17]. Figure 2 shows that despite its constrained budget, our model is almost on par with the original GVLAD approach [9] with 32 times less number of parameters. Other solutions have lower EER, but require computational capabilities incompatible with embedded systems. Contrary to the number of parameters, the computation cost is more rarely reported in the literature, but critical to real life applications. The second smallest model displayed on Figure 2, denoted 2020-03 Chung, has only 1.437M learning parameters (6 times more than the proposed approach) but actually requires 353.3 MFLOPS, or 30 times more operations at processing time than our solution<sup>2</sup>.

### 4.2. VOICES 2019 challenge

We compare the proposed approach with existing works following the *fixed* training conditions of the VOICES benchmark (see [23] for more details). The results reported in Table 3 for our approach refer to a model trained on VoxCeleb2 augmented as detailed in Section 3.1. Even in challenging acoustic conditions, the increase in EER compared to other approaches remains limited (+2.57 percentage points from the best performing approach) while the number of parameters is drastically reduced (26 times smaller). It should be noted that the best scoring systems in Table 3 employ additional VAD and scoring mechanisms that are expected to

<sup>2</sup>To compute these numbers, we built the ONNX model from the source code available at [https://github.com/clovaai/voxceleb\\_trainer](https://github.com/clovaai/voxceleb_trainer)



**Fig. 2:** EER% comparison on the VoxCeleb1 verification test with several previous approaches: 2018-06 Chung [17], 2019-02 Xie [9], 2020-03 Chung [6], 2019-10 Yadav [21], 2019-10 Zeinali [22], 2019-10 Zeinali\* [22] (models fusion). EER for 2020-03 Chung has been recomputed by building an embedding model from the provided source code, adding a cosine distance scorer and running the benchmark, all others EERs are reported from the corresponding papers.

Model	EER%	params (K)
STC-Innovations Ltd. [24]	5.04	5953.5
BUT from Brno University [25]	<b>4.90</b>	6083.0
Ours	7.47	<b>237.5</b>

**Table 3:** Comparison in the *fixed* training conditions [23] of EER% and learning parameters to the best reported single models on the VOICES from a Distance 2019 Challenge.

significantly improve accuracy especially in challenging conditions.

## 5. CONCLUSION

We propose an efficient model architecture for speaker verification suited for embedded systems. Our results demonstrate that our solution yields a limited increase of EER on well established benchmarks, while drastically reducing the number of parameters and operations. Rarely reported in the literature, the inference properties of the model have been studied, highlighting a good level of responsiveness. Future work will be centered on further improving the accuracy in challenging acoustic environments, first by integrating the proposed solution with a low-resource VAD and advanced post-embedding scoring techniques as in [25].

## 6. REFERENCES

- [1] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, “Front-End Factor Analysis for Speaker Verification,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, Jun 2011.
- [2] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, “End-to-end text-dependent speaker verification,” in *ICASSP*. IEEE, 2016, pp. 5115–5119.
- [3] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in *Interspeech*, 2017, pp. 999–1003.
- [4] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, “Attentive statistics pooling for deep speaker embedding,” *arXiv:1803.10963*, 2018.
- [5] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, “Generalized end-to-end loss for speaker verification,” in *ICASSP*. IEEE, 2018, pp. 4879–4883.
- [6] Joon Son Chung, , et al., “In defence of metric learning for speaker recognition,” *arXiv:2003.11982*, 2020.
- [7] Alice Coucke, Mohammed Chlieh, Thibault Gisselbrecht, David Leroy, Mathieu Poumeyrol, and Thibaut Lavril, “Efficient keyword spotting using dilated convolutions and gating,” in *ICASSP*. IEEE, 2019, pp. 6351–6355.
- [8] Samuel Krizan et al., “Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions,” in *ICASSP*. IEEE, 2020, pp. 6124–6128.
- [9] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Senior, “Utterance-level aggregation for speaker recognition in the wild,” in *ICASSP*. IEEE, 2019, pp. 5791–5795.
- [10] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, “A light cnn for deep face representation with noisy labels,” *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2884–2896, 2018.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proc IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
- [12] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” in *SLT*. IEEE, 2016, pp. 165–170.
- [13] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey, “Self-attentive speaker embeddings for text-independent speaker verification,” in *Interspeech*, 2018, pp. 3573–3577.
- [14] Alexander Hermans, Lucas Beyer, and Bastian Leibe, “In defense of the triplet loss for person re-identification,” *arXiv:1703.07737*, 2017.
- [15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *CVPR*, 2019, pp. 4690–4699.
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
- [17] Joon Son Chung, Arsha Nagrani, and Andrew Senior, “Voxceleb2: Deep speaker recognition,” *arXiv:1806.05622*, 2018.
- [18] David Snyder, Guoguo Chen, and Daniel Povey, “Musan: A music, speech, and noise corpus,” *arXiv:1510.08484*, 2015.
- [19] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *ICASSP*. IEEE, 2018, pp. 351–355.
- [20] Colleen Richey et al., “Voices obscured in complex environmental settings (voices) corpus,” *arXiv:1804.05053*, 2018.
- [21] Sarthak Yadav and Atul Rai, “Frequency and temporal convolutional attention for text-independent speaker recognition,” in *ICASSP*. IEEE, 2020, pp. 6794–6798.
- [22] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, “But system description to voxceleb speaker recognition challenge 2019,” *arXiv:1910.12592*, 2019.
- [23] Mahesh Kumar Nandwana, Julien van Hout, Mitchell McLaren, Colleen Richey, Aaron Lawson, and Maria Alejandra Barrios, “The VOICES from a Distance Challenge 2019 Evaluation Plan,” *arXiv:1902.10828*, 2019.
- [24] Sergey Novoselov et al., “Stc speaker recognition systems for the voices from a distance challenge,” *arXiv:1904.06093*, 2019.
- [25] Hossein Zeinali et al., “But voices 2019 system description,” *arXiv:1907.06112*, 2019.