Proceedings of Machine Learning Research vol 278:1–9, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

DCRes2Net: Enhanced Res2Net with Dimensional Feature Fusion for Speaker Verification

Ya Li LEIA@MAIL.SCUEC.EDU.CN
College of Computer Science, South-Central Minzu University, Wuhan, China

Bin Zhou* BINZHOU@MAIL.SCUEC.EDU.CN
College of Computer Science, South-Central Minzu University, Wuhan, China

Bo Hu HUBO@ETAH-TECH.COM
Wuhan Dongxin Tongbang Information Technology Co., Ltd., Wuhan, China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Many classical convolutional architectures have been introduced to the field of speaker verification; however, solely employing one-dimensional or two-dimensional convolutions is insufficient for efficiently modeling speaker features. To address this limitation, this paper introduces a multi-dimensional feature fusion strategy and presents an enhanced Res2Net architecture based on dimensional feature fusion. Different feature modeling techniques complement each other, fully leveraging their respective advantages in temporal and spatial feature extraction to achieve comprehensive representation of multi-dimensional data. Experiments conducted on the VoxCeleb dataset demonstrate that the proposed architecture achieves competitive performance along with robust generalizability.

Keywords: Speaker Verification, Dimensional Feature Fusion, Res2Net

1. Introduction

Speaker Verification (SV) is a task aimed at verifying whether a given speaker originates from a registered source. The SV system extracts speaker characteristics from a segment of speech to compare different speaker identity information. The extracted speaker embeddings can also be utilized in downstream tasks, such as speech synthesis, speaker diarization and related areas in speech processing. Obviously, one of the great challenges of SV is how to extract distinguishable speaker features.

Due to the inherent sequential nature of speech, early researchers employed Time-delay Neural Networks (TDNN) (Desplanques et al., 2020; Yao et al., 2023) as a feature encoder to transform acoustic inputs into fixed-length speaker embeddings. TDNN performs 1D convolution along the time axis, using a controlled context window to capture temporal dependencies and effectively model speech across various time scales. However, its ability to model frequency information depends on the number of filter banks, which is only effective when the number is sufficiently large. The ECAPA-TDNN model (Desplanques et al., 2020), available in 512- and 1024-channel variants, demonstrates that increasing the number of channels enhances the time-frequency representation, but also leads to higher parameter counts and greater computational overhead.

Inspired by the success of 2D convolutions in image processing, researchers have applied 2D convolutional networks to speaker verification (Liu et al., 2022a; Chen et al., 2023; Liu et al., 2024) by treating speech spectrograms as images—where time and frequency correspond to the width and

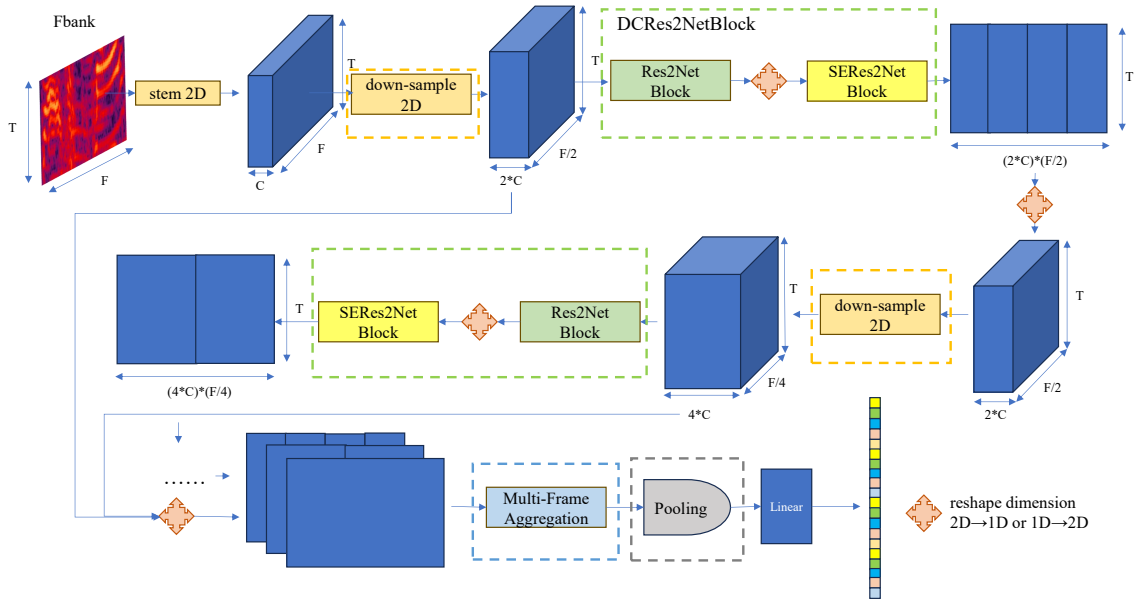

Figure 1: Overview of the proposed DCRes2Net architecture.

height, respectively—and expanding the channel dimension for feature extraction. This approach leverages the ability of 2D convolutions to capture local spatial features and model correlations between the temporal and spectral dimensions, thereby enhancing time-frequency representation learning. Nevertheless, relying exclusively on 2D convolutional modules for speaker representation modeling presents several limitations. First, high-performing 2D convolutional networks typically require a large number of basic modules, resulting in increased model complexity and higher parameter counts. Second, the need to slide convolutional kernels across both time and frequency dimensions leads to significantly higher computational costs. Third, as 2D convolutions were originally designed for image-based applications, their adaptability to speech-specific characteristics is limited. The focus of 2D convolutions on local feature extraction restricts the capacity to capture long-range temporal dependencies, often requiring deep network stacking to achieve competitive performance.

Different feature encoders contribute distinctively to speaker modeling. To fully exploit the advantages offered by convolution modules in multiple dimensions, this paper proposes a novel dimensional feature fusion architecture—DCRes2Net, as illustrated in Figure 1. The proposed architecture leverages Res2Net module—which concurrently supports both 2D and 1D convolution modalities—as its fundamental building block. These modules are integrated into a Dimensional Feature Fusion Module (DCRes2NetBlock) that facilitates multi-dimensional feature modeling by synergistically combining the merits of various feature extraction approaches. To hierarchically integrate these representations, a multi-frame aggregation strategy is employed to consolidate features across different abstraction levels. Subsequently, an attention-based statistics pooling mechanism (Okabe K, 2018) is applied to project the aggregated features into a fixed-dimensional space, followed by a linear transformation for dimensionality reduction to derive the speaker embedding. Experiments conducted on the open-source VoxCeleb dataset (Nagrani et al., 2017; Chung et al., 2018)

substantiate the efficacy of the proposed DCRes2Net architecture, demonstrating that it achieves competitively robust performance with a reduced parameter count and computational cost.

2. System Description

2.1. DCRes2NetBlock

To achieve multi-scale feature representation in the residual module, Gao et al. (2021) proposed a novel convolutional neural network module known as Res2Net. The core innovation of the Res2Net module lies in its utilization of grouped convolutions to realize multi-scale feature modeling. In this study, Res2Net modules of varying dimensions are integrated to model speaker features. The microstructure of the DCRes2NetBlock is illustrated in Figure 2. Initially, a grouped convolution operation within the two-dimensional Res2Net module is employed to model local time-frequency domain features. Subsequently, the feature dimensions are reshaped, and the processed features are finally fed into a one-dimensional SERes2Net module (Desplanques et al., 2020). This module integrates a Squeeze-Excitation (Hu et al., 2018) block following the one-dimensional Res2Net module, thereby further enhancing the representational capacity of the Res2Net module. Specifically, assuming the 2D Res2Net block produce feature maps with dimensions $[C, F, T]$ (ignoring the batch dimension), these features are first flattened along the channel-frequency axis to shape $[C * F, T]$, then propagated through the subsequent 1D Res2Net component.

Moreover, pertinent optimizations have been implemented within the Res2Net module, as expressed in the following equation, where the input features are assumed to be partitioned into s groups along the channel dimension.

$$y_i = \begin{cases} x_i & i = 1 \\ K2U_i\left(x_i + y_{i-1}\right) & 2 \leq i \leq s \end{cases} \tag{1}$$

$$y^{'}{}_i = \begin{cases} x^{'}{}_i & i = 1 \\ K1U_i\left(x^{'}{}_i + y^{'}{}_{i-1}\right) & 2 \leq i \leq s \end{cases} \tag{2}$$

Where x_i and y_i denote the input and corresponding output of the i-th sub-group, respectively. The operator $K2U_i(\cdot)$ represents a 3×3 convolution layer followed by batch normalization (Ioffe and Szegedy, 2015) and a non-linear activation operation, whereas $K1U_i(\cdot)$ corresponds to a one-dimensional convolution with a kernel size of 3, also succeeded by batch normalization and an activation operation. In both cases, the Gaussian Error Linear Unit (GELU) (Hendrycks D, 2016) is employed as the activation function.

In the original Res2Net module, the residual output from the first group was not transmitted laterally, with lateral information transfer commencing only from the subsequent groups. This design inherently results in underutilization of the first group's features. To fully exploit the information from all groups, the proposed method incorporates the output of the first group into the lateral transmission pathway, thereby enhancing both information transfer efficiency and overall model performance, as illustrated in Figure 2. This optimization enables a more cohesive fusion of multi-scale features within the module, thereby augmenting the representational capacity. Additionally, by dynamically adjusting the dilation factor within the one-dimensional Res2Net module, the model broadens its contextual modeling scope, enhancing its ability to capture long-range contextual dependencies.

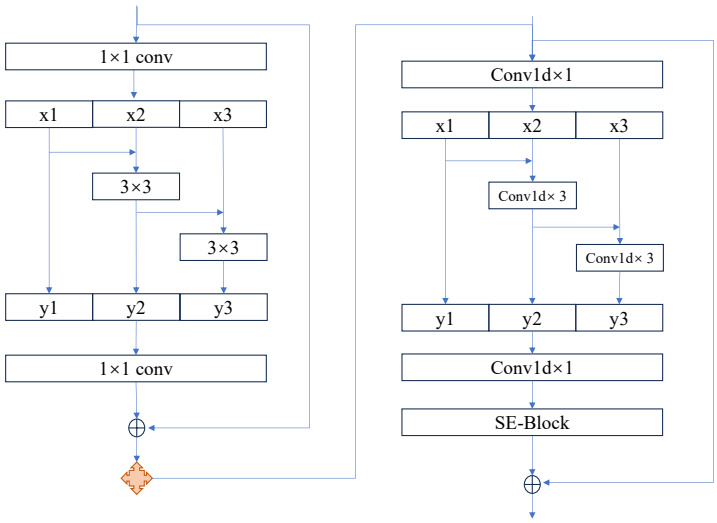

Figure 2: Illustration of DCRes2NetBlock.

2.2. Multi-Frame Aggregation

The multi-frame aggregation (MFA) layer is designed to integrate contextual information from sequential data. In this study, feature maps from different hierarchical levels are concatenated to combine features with varying receptive fields, and the resulting high-dimensional concatenated features are subsequently fused into a compact low-dimensional representation. Specifically, the downsampled feature maps are first extracted and reshaped from three dimensions to a two-dimensional representation with the shape $[F * C, T]$ (ignoring the batch dimension), and then these features are aggregated. In the temporal domain, the consistency of temporal context among the features is preserved by retaining the original temporal dimension; in the feature domain, feature sequences from different abstraction levels are consolidated to yield a final dimension of $[F * C * n, T]$.

Due to the typically high redundancy in such high-dimensional features, a learnable multi-frame aggregation layer is introduced to project the high-dimensional representation into a lower-dimensional space, as shown in Equation (3):

$$MFA = BN\left(GELU\left(W_t \cdot MR + b_1\right)\right) \tag{3}$$

MR denotes the concatenated features, W_t represents the weights used to fuse the features via a time-delay neural network, and b_1 is the bias term. $GELU$ refers to the Gaussian Error Linear Unit activation function, and BN denotes batch normalization. The resulting MFA features possess dimensions $[F_m, T]$. This dimensionality reduction effectively alleviates the computational complexity and mitigates overfitting risks associated with processing high-dimensional features. By employing a learnable mapping, the MFA layer retains the key information in the high-dimensional features while discarding redundant or noisy components, thereby enhancing the training efficiency of subsequent modules and improving the overall generalization capability of the model.

Moreover, this multi-frame aggregation approach embodies the principle of feature reuse and sharing inherent in residual networks. Through cross-layer feature concatenation, the model flexibly reuses hidden features across layers, ultimately maximizing the utilization of available information.

3. Experiments and Analysis

3.1. Datasets and Evaluation Metrics

Experiments are conducted on the open-source speaker verification dataset, VoxCeleb (Nagrani et al., 2017; Chung et al., 2018). Specifically, the development set of VoxCeleb2 (Chung et al., 2018) is utilized for training, which includes 5,994 speakers and a total of 1,092,009 speech segments. VoxCeleb1 (Nagrani et al., 2017)'s development and test sets are employed for evaluation. The dataset includes three sets of trials with varying difficulty levels: VoxCeleb-O, VoxCeleb-H, and VoxCeleb-E. Given the complexity of acoustic environments, the training data is enhanced using noise datasets MUSAN (Snyder et al., 2015) and RIR (Ko et al., 2017).

The model performance is assessed through two industry-standard evaluation criteria: equal error rate (EER) and the minimum detection cost function (minDCF) with 0.01 target probability.

3.2. Implementation Details

The proposed DCRes2Net model is implemented using the 3D-Speaker toolkit (Chen et al., 2024). For input features, we employ 80-dimensional log mel filterbank (FBank) features with a window length of 25 ms and an offset of 10 ms as input features. In addition to augmenting with noise datasets, speed perturbation is applied to the audio, with randomly sampling at rates of 0.9, 1.0, and 1.1 to triple the number of speakers.

In our experiments, the stochastic gradient descent (SGD) optimizer is employed with an initial learning rate of 0.1, a momentum of 0.9, and weight decay set to 0.0001. We also incorporate a cosine annealing scheduler and linear warm-up scheduler for learning rate scheduling, with a minimum learning rate of 0.0001. The angular additive margin softmax (AAM-Softmax) loss (Deng et al., 2022) is employed, with margin and scaling factors set to 0.3 and 32, respectively. The final fully connected layer outputs speaker features of dimension 192. To enhance training efficiency and ensure robust model generalization, we perform randomized extraction of 3-second audio clips from each audio sample during training data preparation.

Due to the high CPU demands and extended training time associated with dynamic data augmentation, some experiments in this study were conducted using a single-round augmentation approach. In this method, the augmented features are saved to disk, and subsequent training rounds load these precomputed features directly from the file. This strategy may adversely affect overall model performance. To ensure fairness across experiments, all other models were also reproduced using this same training protocol. To facilitate the retrieval of literature corresponding to each model, the following citations are provided: ECAPA-TDNN-1024 (Desplanques et al., 2020), ERes2Net (Chen et al., 2023), CAM++ (Wang et al., 2023), Branch-ECAPA-TDNN (Yao et al., 2023), DS-TDNN (Li et al., 2024), Gemini Res2Net34 (Liu et al., 2024), DF-ResNet56 (Liu et al., 2022a), and MFA-TDNN(Lite) (Liu et al., 2022b).

3.3. Results and Analysis

The comparative experimental results between the proposed DCRes2Net model and baseline systems are presented in Table 1. The upper section of the table reports the performance of various models reproduced using single-round data augmentation, while the lower section provides evaluation metrics obtained under dynamic data augmentation strategies. When dynamic augmentation is not employed during training, the performance of baseline models exhibits limited improvement.

This limitation primarily stems from insufficient simulation of noisy environments—specifically in terms of noise diversity. Consequently, models fail to effectively learn the diverse acoustic characteristics present in noisy conditions, resulting in suboptimal performance on the test set. In contrast, the proposed DCRes2Net model, which incorporates dimensional feature fusion, demonstrates robust performance even under constrained training conditions. By leveraging dimension transformation and scaling mechanisms, the model is able to more effectively capture speaker-discriminative features. Compared to ERes2Net, the DCRes2Net model achieves relative reductions in EER of 15.5%, 11.3%, and 9.5% on the VoxCeleb1 test set, respectively, thereby validating the effectiveness of the proposed approach in speaker verification tasks.

Table 1: Comparison of EER and minDCF Metrics for Various Architectures on the VoxCeleb1 Test Set

Architecture	DA	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	MinDCF
ECAPA-TDNN-1024	×	1.04	0.138	1.35	0.155	2.61	0.249
ERes2Net	×	1.16	0.141	1.24	0.137	2.31	0.219
CAM++	×	1.10	0.154	1.11	0.138	2.20	0.223
DCRes2Net	×	**0.98**	**0.108**	**1.10**	**0.131**	**2.09**	**0.211**
ECAPA-TDNN-1024	✓	0.89	0.092	1.07	0.119	1.98	0.196
Branch-ECAPA-TDNN	✓	0.90	0.094	1.13	0.126	2.13	0.214
DS-TDNN	✓	0.90	0.118	1.15	0.140	2.11	0.199
Gemini Res2Net34	✓	1.05	0.100	1.09	0.115	1.91	0.176
DF-ResNet56	✓	0.96	0.103	1.09	0.122	1.99	0.184
ERes2Net	✓	0.92	0.094	0.99	0.111	1.92	0.181
MFA-TDNN(Lite)	✓	0.97	**0.091**	1.14	0.121	2.17	0.199
DCRes2Net	✓	**0.82**	0.099	**0.92**	**0.105**	**1.74**	**0.170**

With the integration of dynamic data augmentation techniques, various models exhibit significant improvements on the test sets. On the most challenging evaluation benchmark, VoxCeleb1-H, the proposed DCRes2Net model achieves a substantial performance breakthrough, reducing the EER from 1.90 to 1.74. Compared to the ECAPA-TDNN-1024 model, DCRes2Net demonstrates clear advantages across most evaluation metrics, achieving relative EER reductions of 7.9%, 14.0%, and 17.5%, respectively. Notably, this performance improvement is accomplished with only approximately one-third of the parameter count and three-fifths of the floating-point operations (FLOPs) required by ECAPA-TDNN-1024, highlighting the model's computational efficiency. In contrast, the Gemini Res2Net34 model, which relies solely on a two-dimensional Res2Net architecture, exhibits inferior performance. This reveals its limitations in modeling global feature dependencies and its inability to effectively capture the complex interactions among global representations. The advantages of DCRes2Net become even more pronounced under more demanding evaluation scenarios. For instance, on the VoxCeleb1-H test subset, DCRes2Net achieves an additional relative EER reduction of 9.0%, further demonstrating its robustness and superiority in challenging acoustic environments.

Table 2: Ablation Study of Key Components in DCRes2Net

Model	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
w/o 1D	1.97	0.185	1.92	0.206	3.27	0.299
w/o 2D	1.23	0.129	1.34	0.16	2.55	0.242
w/o dilation	1.12	0.112	1.19	0.138	2.18	0.210
DCRes2Net	0.98	0.108	1.10	0.131	2.09	0.211

To evaluate the effectiveness of the proposed approach and to investigate the impact of key components or mechanisms within DCRes2Net on the overall model performance, ablation studies were conducted on the respective components. The first and second rows in Table 2 report ablation experiments on different dimensional Res2Net modules. The results demonstrate that removing the 1D module causes a dramatic increase in the error rate, whereas the exclusion of the 2D module has a comparatively minor impact on performance. Analysis of the training logs reveals that the removal of the 1D module severely impedes model convergence and drastically diminishes its ability to model speaker features. The 2D module is primarily responsible for capturing local time-frequency domain characteristics; however, its capacity to model contextual dependencies is limited. In contrast, the 1D module effectively compensates for this deficiency by dynamically adjusting the dilation coefficients, thereby enabling feature modeling of varying contextual lengths at different network layers. This approach significantly improves the model's capacity to effectively model long-range contextual interactions across sequential data. The integrated and complementary utilization of both modules achieves robust modeling of speaker embedding features by leveraging their respective strengths, which in turn improves the model's performance under complex conditions. The third row illustrates the scenario in which the dilation coefficients are not expanded—that is, the dilation coefficients in the 1D SERes2Net module are uniformly set to 1. Under these circumstances, the contextual information learned by the model is limited. In DCRes2Net, however, the dilation coefficients incrementally increase across layers, systematically broadening the model's receptive field for contextual information and thereby enhancing its global feature modeling capability, which significantly improves overall performance.

4. Conclusion

In this paper, we propose an enhanced Res2Net architecture based on dimension-wise feature fusion. DCRes2Net achieves both local and global modeling of speaker characteristics by integrating outputs from feature extraction modules operating in different dimensions. Subsequently, a multi-frame aggregation layer concatenates feature maps with varying receptive field sizes into a compact, low-dimensional feature representation. Experiments on the VoxCeleb dataset demonstrate that DCRes2Net delivers superior performance with impressive robustness.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities of South-Central Minzu University (Grant Number: CZY23006).

References

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification. In *Proc. Interspeech 2023*, pages 2228–2232, Dublin, Ireland, August 2023.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Tinglong Zhu, Changhe Song, et al. 3d-speaker-toolkit: An open source toolkit for multi-modal speaker verification and diarization. *arXiv preprint arXiv:2403.19971*, 2024.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018*, pages 1086–1090, Hyderabad, India, September 2018.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834, Virtual Event, Shanghai, China, October 2020.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, February 2021. ISSN 1939-3539.

Gimpel K Hendrycks D. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, Salt Lake City, UT, USA, June 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pages 448–456, Lille, France, July 2015.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224, New Orleans, LA, USA, March 2017.

Yangfu Li, Jiapan Gan, Xiaodan Lin, Yingqiang Qiu, Hongjian Zhan, and Hui Tian. Ds-tdnn: Dual-stream time-delay neural network with global-aware filter for speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2814–2827, 2024.

Bei Liu, Zhengyang Chen, Shuai Wang, Haoyu Wang, Bing Han, and Yanmin Qian. Df-resnet: Boosting speaker verification performance with depth-first design. In *Proc. Interspeech 2022*, pages 296–300, Incheon, Korea, September 2022a.

Tianchi Liu, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li. MFA: TDNN with Multi-Scale Frequency-Channel Attention for Text-Independent Speaker Verification with Short Utterances. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7517–7521, Singapore, Singapore, May 2022b.

Tianchi Liu, Kong Aik Lee, Qiongqiong Wang, and Haizhou Li. Golden gemini is all you need: Finding the sweet spots for speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2324–2337, 2024.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proc. Interspeech 2017*, pages 2616–2620, Stockholm, Sweden, August 2017. ISCA.

Shinoda K Okabe K, Koshinaka T. Attentive statistics pooling for deep speaker embedding. In *Proc. Interspeech 2018*, pages 2252–2256, Hyderabad, India, September 2018.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A Music, Speech, and Noise Corpus. *arXiv preprint arXiv:1510.08484*, 2015.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. In *Proc. Interspeech 2023*, pages 5301–5305, Dublin, Ireland, August 2023.

Jiadi Yao, Chengdong Liang, Zhendong Peng, Binbin Zhang, and Xiao-Lei Zhang. Branch-ECAPA-TDNN: A Parallel Branch Architecture to Capture Local and Global Features for Speaker Verification. In *Proc. Interspeech 2023*, pages 1943–1947, Dublin, Ireland, August 2023.