

Reservoir Computing with Multi-Scale Attention for Efficient Speaker Verification

Luiz Silva and Edson Cataldo

Abstract—This paper proposes a lightweight architecture for speaker recognition and verification that combines Reservoir Computing (RC) with multi-scale and multi-head attention. A fixed recurrent reservoir models long-term temporal dynamics, while attention selectively aggregates speaker-discriminative information across temporal resolutions, restricting learning to compact layers and reducing the number of trainable parameters. Evaluations on VoxCeleb1, LibriSpeech, and CNCeleb1, including cross-corpus open-set scenarios, show competitive Equal Error Rate (EER) compared to state-of-the-art models with at least $3.5\times$ fewer parameters. Ablation results indicate that reservoir dimensionality and multi-scale aggregation drive performance, while attention improves robustness and feature integration. Strong generalization under domain mismatch further highlights the effectiveness of the proposed approach, making it suitable for resource-constrained and real-time applications.

Keywords—Speaker Recognition, Deep Machine Learning, Reservoir Computing, Attention Mechanism.

I. INTRODUCTION

The rapid growth of data-driven speech technologies has increased the demand for models capable of processing large volumes of audio data with high accuracy and low latency. In speaker recognition and verification, deep neural networks achieve state-of-the-art performance, but often at the cost of high computational complexity, large numbers of trainable parameters, and long training times. These limitations hinder their use in real-time and resource-constrained scenarios, motivating the development of lightweight architectures with improved efficiency-performance trade-offs [1].

Reservoir Computing (RC) has emerged as an efficient alternative for temporal signal processing [2]. By employing a fixed recurrent reservoir followed by a trainable readout layer, RC significantly reduces training complexity and computational cost [1]. However, its fixed dynamics may limit adaptability to input distributions, restricting the extraction of fine-grained temporal features [3]. Attention mechanisms offer a natural solution by enabling adaptive weighting of informative temporal or spectral regions, and have shown strong performance in speech-related tasks [4, 5, 6, 7]. When combined with RC, attention can compensate for static reservoir dynamics without sacrificing efficiency [8].

Speaker recognition and verification further require robustness to variations in phonetic content, speaking style, noise, and channel conditions [9, 10, 11]. These challenges demand models capable of capturing long-term dependencies while

remaining computationally efficient, especially in open-set scenarios.

In this work, we hypothesize that separating *temporal dynamic expansion* and *adaptive feature selection* provides an effective framework for speaker representation. We leverage RC as a fixed nonlinear dynamical system for temporal encoding, while attention mechanisms act as adaptive selectors that emphasize speaker-relevant patterns.

Based on this idea, we propose a hybrid RC-Attention architecture that integrates Reservoir Computing with multi-scale and multi-head attention. The multi-scale design captures temporal patterns at different resolutions, while attention selectively aggregates informative reservoir states, preserving efficiency while enhancing representational capacity.

Experimental results, including ablation studies and cross-corpus evaluations on VoxCeleb1, LibriSpeech, and CNCeleb1, demonstrate that the proposed approach achieves competitive speaker verification performance with significantly fewer trainable parameters than conventional deep models.

II. METHODOLOGY: RC ARCHITECTURE WITH MULTI-SCALE AND MULTI-HEAD ATTENTION

This section presents the proposed RC-Attention architecture for speaker representation, combining efficient temporal encoding with adaptive feature selection.

The model consists of two stages. First, a fixed recurrent reservoir projects the input speech signal into a high-dimensional dynamic space, providing nonlinear temporal expansion and capturing long-term dependencies without training recurrent weights. Second, the resulting reservoir states are processed by a multi-scale, multi-head attention module, where temporal pooling captures patterns at different resolutions and attention selectively emphasizes speaker-discriminative features.

By decoupling temporal encoding (reservoir) from adaptive feature selection (attention), the architecture improves discriminative capability while maintaining a reduced number of trainable parameters and low computational cost.

A. Overall System Flow

The proposed speaker verification pipeline is illustrated in Figure 1. Given an input audio signal $s(t)$, feature extraction is first performed, followed by processing through the proposed RC-Attention module.

The RC-Attention block operates as a unified embedding extractor, where temporal dynamic encoding is performed by the reservoir and adaptive feature selection is handled by the attention mechanism. A detailed schematic of this

Luiz Silva, Graduate Program in Electrical and Telecommunications Engineering, UFF, Niteroi-RJ, e-mail: luizalbertosilva@id.uff.br; Edson Cataldo, Graduate Program in Electrical and Telecommunications Engineering, UFF, Niteroi-RJ, e-mail: ecataldo@id.uff.br. This work was partially supported by CNPq (309803/2023-3 and 178711/2025-0).

module, including reservoir dynamics, multi-scale processing, and attention, is shown in Figure 2.

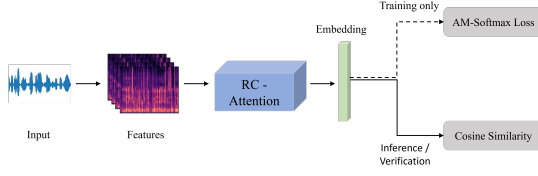


Fig. 1

HIGH-LEVEL SCHEMATIC OF THE SPEAKER VERIFICATION PIPELINE.

B. Preprocessing and Spectrogram Extraction

The raw speech signal $s(t)$ is transformed into a time–frequency representation using the Short-Time Fourier Transform (STFT), followed by a Mel-scale filterbank to obtain a Mel spectrogram. The resulting representation is

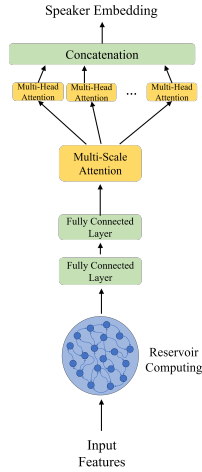


Fig. 2

INTERNAL STRUCTURE OF THE RC-ATTENTION BLOCK WITH MULTI-SCALE AND MULTI-HEAD ATTENTION.

adjusted to a fixed number of frames T through truncation or zero-padding. This yields an input tensor $x \in \mathbb{R}^{B \times T \times F}$, where B is the batch size, T is the number of temporal frames, and F is the number of Mel coefficients, preserving the sequential spectro-temporal structure of the signal.

C. Recurrent Reservoir

The input sequence is processed by a recurrent reservoir, where the internal state at time t evolves according to

$$r_t = \tanh(W_{\text{in}}x_t + W_{\text{rec}}r_{t-1}), \quad (1)$$

where $\tanh(\cdot)$ is applied element-wise, $W_{\text{in}} \in \mathbb{R}^{L \times F}$ and $W_{\text{rec}} \in \mathbb{R}^{L \times L}$ denote the input and recurrent weight matrices, respectively, and L is the number of reservoir neurons. For each time step, $x_t \in \mathbb{R}^F$ and $r_t \in \mathbb{R}^L$.

The sequence of reservoir states is given by

$$R = [r_1, r_2, \dots, r_T] \in \mathbb{R}^{T \times L}, \quad (2)$$

and, for a batch of inputs, the reservoir output is denoted as $r_{\text{Cout}} \in \mathbb{R}^{B \times T \times L}$.

The reservoir acts as a fixed nonlinear dynamical system that performs temporal expansion of the input signal. The weight matrices W_{in} and W_{rec} are randomly initialized from a uniform distribution $\mathcal{U}(-1/\sqrt{L}, +1/\sqrt{L})$ and remain fixed during training, ensuring low training complexity while preserving stable internal dynamics.

To adapt the reservoir representation for subsequent processing, r_{Cout} is passed through a lightweight projection module composed of two layers with ReLU activations, reducing the dimensionality to $L/2$. This stage introduces a trainable bottleneck that compresses the high-dimensional reservoir states while enhancing their discriminative structure.

Finally, layer normalization is applied across temporal frames, resulting in a stabilized representation $h \in \mathbb{R}^{B \times T \times L/2}$, which serves as input to the subsequent multi-scale attention modules.

D. Multi-Scale and Multi-Head Attention

Let $h \in \mathbb{R}^{B \times T \times D}$ denote the projected reservoir output, where $D = L/2$. Multi-scale representations are obtained via temporal average pooling:

$$h^{(k)} \in \mathbb{R}^{B \times T_k \times D}, \quad k = 1, \dots, K,$$

with $T_k < T$. This multi-scale decomposition enables the representation of temporal patterns at different resolutions.

For each scale k , a multi-head attention pooling mechanism with N heads is applied. Each head operates on $h^{(k)}$ using independent learnable projections, allowing the extraction of complementary temporal patterns. Unlike full self-attention, this formulation corresponds to a parametric attention pooling strategy [12].

For head n , attention scores are computed as

$$a_t^{(k,n)} = \langle h_t^{(k)}, w_n \rangle, \quad w_n \in \mathbb{R}^D, \quad (3)$$

where $\langle \cdot, \cdot \rangle$ denotes inner product. It is followed by temporal softmax normalization:

$$\alpha_t^{(k,n)} = \frac{\exp(a_t^{(k,n)})}{\sum_{t'=1}^{T_k} \exp(a_{t'}^{(k,n)})}. \quad (4)$$

The corresponding context vector is obtained by weighted aggregation:

$$c_n^{(k)} = \sum_{t=1}^{T_k} \alpha_t^{(k,n)} h_t^{(k)}, \quad c_n^{(k)} \in \mathbb{R}^{B \times D}. \quad (5)$$

The N heads are combined through a learnable gating mechanism:

$$z_n = \frac{\exp(g_n/\tau)}{\sum_{i=1}^N \exp(g_i/\tau)}, \quad (6)$$

where $g \in \mathbb{R}^N$ and τ is a temperature parameter controlling the sharpness of the distribution [13]. The aggregated representation for scale k is then given by

$$c^{(k)} = \sum_{n=1}^N z_n c_n^{(k)}, \quad c^{(k)} \in \mathbb{R}^{B \times D}. \quad (7)$$

Finally, representations from all scales are concatenated:

$$c = [c^{(1)}; c^{(2)}; \dots; c^{(K)}] \in \mathbb{R}^{B \times KD}. \quad (8)$$

This module performs adaptive feature selection over the reservoir states, where multi-scale pooling captures temporal structure at different resolutions and multi-head attention emphasizes complementary discriminative patterns.

E. Final Projection and ℓ_2 Normalization

The concatenated multi-scale representation is projected into the final embedding space through a linear layer, followed by layer normalization.

An ℓ_2 normalization is then applied to the resulting vector, constraining embeddings to lie on a unit hypersphere. This enables similarity computation using cosine distance or inner product, which is standard in speaker verification tasks [18].

F. Discriminative Loss Function

The model is trained using the Additive Margin Softmax (AM-Softmax) loss [14], which enhances class separability by introducing an angular margin between embeddings and class weights, along with a scaling factor.

III. EXPERIMENTAL SETUP

All experiments were conducted using a fixed configuration of the proposed RC-Attention architecture to ensure consistency and reproducibility across evaluations. Implemented in PyTorch; run on an AMD Ryzen 7 5700G CPU, NVIDIA RTX 3060 GPU, 32 GB RAM.

A detailed summary of the architectural configuration and hyperparameters is provided in Table I.

TABLE I
PARAMETERS OF THE PROPOSED MODEL

Parameter	Value
Embedding dimension	256
Sampling rate	16 kHz
Window length	25 ms
Window shift	10 ms
Number of Mel filters (F)	80
Audio duration	4 s
Batch size (B)	16
Loss function	AM-Softmax
Optimizer	Adam
Reservoir neurons (L)	512
Number of attention heads	4
Number of temporal scales	4
Temporal scales	1, 2, 4, 8

Adam, AdamW, and Adamax optimizers were evaluated, with Adam yielding the best performance. Similarly, among Softmax, AM-Softmax, and AAM-Softmax, the best results were obtained with AM-Softmax.

A. Evaluation Benchmark

The proposed model is evaluated on three widely used speaker recognition datasets: VoxCeleb1, LibriSpeech, and CNCeleb1. These datasets were selected to assess performance under both in-domain and cross-domain conditions.

VoxCeleb1 [15] consists of speech recordings collected from interview videos, exhibiting significant variability in noise, channel conditions, and speaking styles. It is used for in-domain open-set verification and comparison with existing methods.

LibriSpeech [16] is a large-scale corpus of read English speech recorded under relatively controlled conditions. In this work, the “other” subset is used to introduce additional variability, enabling evaluation under moderately challenging but structured conditions.

CNCeleb1 [17] is a multilingual dataset with substantial variability in acoustic environments, recording conditions, and speaking styles. Due to its domain mismatch, it is used to evaluate cross-corpus generalization and robustness of the learned embeddings.

B. Evaluation Metrics

Performance is evaluated using Equal Error Rate (EER), consistent with the baseline methods used for comparison.

Model complexity is assessed in terms of the number of trainable parameters, providing a hardware-independent measure that facilitates comparison with existing lightweight architectures.

IV. EXPERIMENTAL RESULTS

This section presents the experimental evaluation of the proposed RC-Attention architecture. Due to the stochastic initialization of the reservoir, all experiments were repeated with multiple random seeds, and results are reported as mean and standard deviation of the Equal Error Rate (EER).

An early stopping criterion was applied during training to prevent overfitting, with patience of 20 epochs and validation loss as monitored metric.

The evaluation is structured into three scenarios. First, an ablation study is conducted to assess the contribution of each architectural component. Second, in-domain open-set verification is performed on VoxCeleb1, including comparisons with existing speaker verification models [18]. Third, cross-corpus experiments are carried out by training on VoxCeleb1 or CNCeleb1 and evaluating on multiple datasets, including LibriSpeech, VoxCeleb1, and CNCeleb1, enabling comparison with the ECAPA-TDNN framework [19].

Performance–Cost Contribution of Network Components

An ablation study was conducted to quantify the contribution of each architectural component, modifying one element at a time under a fixed training and evaluation protocol. The evaluated variants include: (i) removal of multi-scale aggregation, (ii) single-head attention, (iii) reduced reservoir size, (iv) reduced projection dimensionality, and (v) removal

TABLE II

IMPACT OF INDIVIDUAL ARCHITECTURAL AND HYPERPARAMETER MODIFICATIONS ON THE PERFORMANCE OF THE RC-ATTENTION SYSTEM. THE FIRST ROW CORRESPONDS TO THE BASELINE CONFIGURATION, WHILE ROWS 2–6 REPORT THE ABLATION RESULTS OBTAINED BY MODIFYING A SINGLE COMPONENT AT A TIME. THE NUMBER OF PARAMETERS IS REPORTED AS TRAINABLE / TOTAL (IN MILLIONS).

Methods	Trainable / Total (M)	EER (%)
Baseline	1.8 / 2.1	3.21 ± 0.03
Multi-scale disabled	1.2 / 1.5	4.78 ± 0.03
Single-head attention	1.8 / 2.1	3.98 ± 0.02
Reduced reservoir size (<i>reservoir_size</i> = 256)	1.6 / 1.7	7.55 ± 0.02
Reduced reservoir size (<i>reservoir_size</i> = 128)	1.4 / 1.5	12.30 ± 0.02
Reduced hidden dimension (<i>hidden_dim</i> = 512)	1.0 / 1.3	6.34 ± 0.04
Head gating disabled (frozen gating module)	1.8 / 2.1	3.99 ± 0.02

of head-gating. Results are reported in Table II in terms of Equal Error Rate (EER) and parameter count.

Disabling multi-scale processing significantly degrades performance (3.21% to 4.78% EER), highlighting the importance of multi-resolution temporal representations. Reducing the reservoir size leads to the largest drop (up to 12.30% EER), confirming its central role in providing rich temporal dynamics.

In contrast, single-head attention and removal of head-gating cause moderate degradation (approximately +0.7% EER), indicating their role in improving robustness and feature aggregation. Reducing projection dimensionality introduces a clear trade-off between efficiency and performance (EER up to 6.34%).

Overall, performance is most sensitive to reservoir dimensionality and temporal scale diversity, supporting the use of fixed reservoirs combined with multi-scale attention for efficient and discriminative speaker modeling.

A. Computational Efficiency and Model Complexity

Table III compares the proposed RC-Attention architecture with representative speaker verification models in terms of parameter count and Equal Error Rate (EER) on VoxCeleb1. Notably, the comparison includes widely adopted architectures reported in recent literature [18], which serve as strong baselines for evaluating lightweight speaker verification models.

The proposed model achieves an EER of 3.21% with approximately 2.1M parameters, whereas most competing approaches operate in the range of 7–10M parameters. This corresponds to a reduction of at least 3.5× in model size while maintaining comparable verification performance.

In particular, the performance gap relative to high-capacity architectures such as ResNet- and Res2Net-based systems remains below 1% absolute EER, indicating that a substantial portion of the discriminative capability can be retained with a significantly more compact model.

TABLE III

COMPARISON OF THE NUMBER OF PARAMETERS AND VERIFICATION PERFORMANCE (EER) OF DIFFERENT METHODS ON THE VOXCELEB1 DATABASE [18].

Methods	Embedding size	Features	Parameters	EER(%)
i-vector+PLDA	400	80	-	10.73
d-vector	1024	80	7.5M	6.44
x-vector	512	80	8.5M	3.98
x-vector (E-TDNN)	512	80	10.0M	3.76
x-vector (F-TDNN)	512	80	9.0M	3.53
r-vector (ResNet)	256	80	8.0M	3.18
r-vector (Res2Net)	256	80	8.5M	2.71
RC-Attention	256	80	2.1M	3.21 ± 0.03

TABLE IV

OPEN-SET SPEAKER VERIFICATION PERFORMANCE COMPARISON (EER %) BETWEEN THE PROPOSED METHOD AND ECAPA-TDNN FRAMEWORK [19], FOR MODELS TRAINED ON VOXCELEB1.

Method	VoxCeleb1 test	CNCeleb1 test	LibriSpeech test-other	LibriSpeech test-clean
ECAPA-TDNN	2.99	22.04	2.70	2.70
RC-Attention (proposed)	3.21 ± 0.03	12.82 ± 0.08	1.24 ± 0.02	1.06 ± 0.02

TABLE V

OPEN-SET SPEAKER VERIFICATION PERFORMANCE COMPARISON (EER %) BETWEEN THE PROPOSED METHOD AND ECAPA-TDNN FRAMEWORK [19], FOR MODELS TRAINED ON CNCELEB1.

Method	VoxCeleb1 test	CNCeleb1 test	LibriSpeech test-other	LibriSpeech test-clean
ECAPA-TDNN	13.22	18.66	4.33	5.90
RC-Attention (proposed)	15.33 ± 0.02	11.54 ± 0.05	2.88 ± 0.02	2.18 ± 0.02

B. Cross-Corpus and Generalization Behavior

In this experiment, the proposed RC-Attention model is trained on either VoxCeleb1 or CNCeleb1 and evaluated on multiple datasets, including LibriSpeech, VoxCeleb1, and CNCeleb1, introducing significant domain mismatch in terms of language, recording conditions, and speaking style. This setup follows a challenging and realistic evaluation protocol for speaker verification systems [19]. The results are summarized in Tables IV and V.

The proposed model exhibits a distinct generalization behavior compared to ECAPA-TDNN. When evaluated on LibriSpeech, substantial improvements are observed across all conditions. For instance, when trained on VoxCeleb1, the EER decreases from 2.70% (ECAPA-TDNN) to 1.24% and 1.06% on the test-other and test-clean subsets, respectively, corresponding to a relative reduction of more than 50%. Similar gains are observed when training on CNCeleb1, where the proposed model consistently outperforms ECAPA-TDNN.

Improvements are also observed when evaluating on CNCeleb1, indicating enhanced robustness under domain and language mismatch. These results suggest that the combination of reservoir-based temporal dynamics and attention-based aggregation promotes representations that are less sensitive to phonetic and linguistic variability.

On VoxCeleb1 test, the proposed model achieves slightly

higher EER compared to ECAPA-TDNN. However, this difference remains limited while the model operates with a significantly reduced number of parameters, highlighting a different trade-off between in-domain optimization and cross-domain generalization.

1) *Attention Map Analysis Across Evaluation Conditions:* Attention maps (Figure 3) show more localized attention for the VoxCeleb1-trained model and more distributed attention for the CNCeleb1-trained model, consistent with its greater variability and with the improved cross-corpus generalization reported in Tables IV and V.

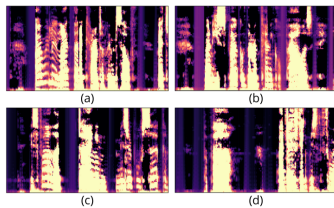


Fig. 3

(A,B) VOXCELEB1-TRAINED; (C,D) CNCELEB1-TRAINED. LIGHTER REGIONS = HIGHER ATTENTION WEIGHT.

V. CONCLUSION

This work presented a hybrid RC-Attention architecture for speaker verification, combining fixed reservoir dynamics with multi-scale and multi-head attention. Experimental results show competitive performance with significantly fewer trainable parameters, with ablation analysis identifying reservoir dimensionality and multi-scale aggregation as the main contributors to discriminative performance, while attention improves robustness and feature integration. Cross-corpus evaluations reveal strong generalization under domain mismatch, particularly on LibriSpeech, indicating reduced sensitivity to phonetic and domain variability and suitability for resource-constrained, real-time scenarios.

Limitations include variability from stochastic reservoir initialization and the added computational cost of multi-scale attention. Future work will explore more efficient attention mechanisms, adaptive reservoir designs, and self-supervised strategies for cross-domain generalization.

REFERENCES

- [1] M. Yan, C. Huang, P. Bienstman, P. Tino, W. Lin, and J. Sun, "Emerging opportunities and challenges for the future of reservoir computing," *Nature Communications*, vol. 15, no. 2056, pp. 1–18, 2024.
- [2] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nature Computational Science*, vol. 2, pp. 10–19, 2022.
- [3] H. Jaeger, "The echo state approach to analysing and training recurrent neural networks – with an erratum note," *GMD Report*, vol. 148, pp. 1–47, 2010.
- [4] X. Xie, H. Cai, C. Li, Y. Wu, and F. Ding, "A voice disease detection method based on MFCCs and shallow CNN," *Journal of Voice*, 2023.
- [5] M. Hireš, M. Gazda, P. Drotár, N. D. Pah, M. A. Motin, and D. K. Kumar, "Convolutional neural network ensemble for Parkinson's disease detection from voice recordings," *Computers in Biology and Medicine*, vol. 141, p. 105021, 2022.
- [6] C.-Y. Park *et al.*, "Harnessing the power of voice: A deep neural network model for Alzheimer's disease detection," *Dementia Neurocognitive Disorder*, pp. 1–10, 2024.
- [7] D. C. Shubhangi and A. K. Pratibha, "Asthma, Alzheimer's and dementia disease detection based on voice recognition using multi-layer perceptron algorithm," in *Proc. ICSES*, 2021, pp. 1–7.
- [8] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
- [9] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.
- [10] N. Shome, A. Sarkar, A. K. Ghosh, R. H. Laskar, and R. Kashyap, "Speaker recognition through deep learning techniques: A comprehensive review and research challenges," *Periodica Polytechnica Electrical Engineering and Computer Science*, vol. 67, no. 3, pp. 300–336, 2023.
- [11] A. Q. Ohi, M. F. Mridha, M. A. Hamid, and M. M. Monowar, "Deep speaker recognition: Process, progress, and challenges," *IEEE Access*, vol. 9, pp. 89619–89643, 2021.
- [12] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," in *Proc. INTERSPEECH*, 2019, pp. 4305–4309.
- [13] S. A. Balanya, J. Maroñas, and D. Ramos, "Adaptive temperature scaling for robust calibration of deep neural networks," *Neural Computing and Applications*, vol. 36, no. 7, 2024.
- [14] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.
- [15] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. INTERSPEECH*, pp. 2616–2620, 2017.
- [16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, pp. 5206–5210, 2015.
- [17] Y. Fan *et al.*, "CN-Celeb: A challenging Chinese speaker recognition dataset," in *Proc. ICASSP*, pp. 7604–7608, 2020.
- [18] M. Jakubec, R. Jarina, E. Lieskovska, and P. Kasak, "Deep speaker embeddings for speaker verification: Review and experimental comparison," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107232, 2024.
- [19] J. Zhang, J. Ma, X. Guo, L. Li, and L. He, "A speaker recognition method based on stable learning," in *Proc. ICASSP*, 2024, pp. 10221–10225.