Improving Speaker Verification Model for Low-Resources Languages

Maksym Kizitskyi, Olena Turuta and Oleksii Turuta

Kharkiv National University of Radio Electronics, 14 Nauky Ave., Kharkiv, 61166, Ukraine

Abstract

Speaker verification is an essential task in speech processing. Implementation this task based on convolutional neural networks. Several key metrics were evaluated, including equal error rate and precision top-K, and were compared the performance of different architectures and loss functions. The experiments are conducted using a Ukrainian dataset and include comparisons of models trained on multilingual data, as well as models trained on clean and augmented data. The results are presented in tables and figures, showing that even for low-resource languages, the models can achieve good performance metrics. The authors also discuss the implications of their findings and the potential for transferring skills to other languages. The paper provides valuable insights for researchers working in the field of speaker verification.

Keywords

Speaker verification, ConvNext, DOLG Architecture, Multilingual Training

1. Introduction

In today's digital age, speech recognition and speaker verification techniques have become increasingly important for a variety of applications. These technologies have revolutionized the way we interact with machines, allowing for seamless communication and automation in various fields, from personal assistants to security systems. Speech recognition refers to the ability of machines to identify and transcribe human speech, while speaker verification focuses on verifying the identity of the person speaking. Both technologies have numerous practical applications, including improving accessibility for individuals with disabilities, enhancing the user experience of devices and applications, and enhancing security measures in industries such as banking and finance. Thus, understanding the importance and potential of speech recognition and speaker verification is crucial for those interested in the future of technology and its impact on society.

Despite the vast potential of speech recognition and speaker verification technologies, there are still significant challenges in implementing them for low-resource languages [1] like Ukrainian. Many of these languages lack the necessary data and resources to develop robust and accurate models. However, recent advancements in machine learning, particularly in deep learning techniques, have made it possible to overcome some of these limitations and enable the development of speech and speaker recognition models for these languages.

The potential impact of these technologies on low-resource languages is immense. Speech recognition can greatly improve accessibility for individuals who speak these languages, allowing them to communicate more effectively with technology and access a wider range of digital content. Speaker verification can also enhance security measures in industries like finance and government, enabling secure authentication of individuals who speak these languages.

Furthermore, the development of speech recognition and speaker verification models for low-resource languages can have broader socio-economic benefits. For example, it can improve the

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine
EMAIL: maksym.kizitskyi@nure.ua (M. Kizitskyi); olena.turuta@nure.ua (O.V. Turuta); oleksii.turuta@nure.ua (O.P. Turuta)
ORCID: 0000-0001-9771-5771 (M. Kizitskyi); 0000-0002-1089-3055 (O.V. Turuta); 0000-0002-0970-8617 (O.P. Turuta)

efficiency and accuracy of customer service for businesses operating in these regions, increasing customer satisfaction and loyalty. It can also facilitate the development of new tools and applications that are specifically tailored to the needs of these populations, enhancing their digital literacy and participation in the global digital economy.

The aim of the work is to develop an approach to perform highly accurate (comparable with performance of SOTA models for resource-rich language) speaker verification for low-resource languages like Ukrainian.

So the goal of this work is to:

- Develop a robust speaker verification system
- Study the effectiveness of transferring skills of speaker verification from other languages
- Compare the effectiveness of different approaches and algorithms

2. Related works

Speaker verification is the process of verifying the identity of a person based on their voice. This process is often used in security systems, access control and other applications where identification is required. However, speaker verification systems are typically designed for high resource languages, leaving low resource languages with limited options. In this literature review we will explore the state-of-the-art research on speaker verification for low resource languages.

In recent years, researchers have attempted to address the issue of speaker verification for low resource languages by developing systems that are capable of identifying individuals who speak less common languages. These efforts have been driven by the need to ensure that all people, regardless of their language, can have access to secure and reliable identification systems.

One approach that has been used to overcome the lack of resources for low resource languages is data augmentation. This technique involves creating new data from existing data by applying various transformations such as pitch shifting, noise addition and speed variation. In a study by Chen et al. (2021) [2], the authors proposed a data augmentation method for speaker verification in low resource languages using a combination of noise addition, reverberation and pitch shifting. The authors reported that their proposed method outperformed the baseline approach, which only used the original data.

Another approach that has been explored is transfer learning, which involves training a model on a resource-rich language and then fine-tuning it for a low resource language. In a study by Sigtia et al. (2018) [3], the authors proposed a transfer learning method for speech recognition in Swahili, a low resource language spoken in East Africa. The authors trained a deep neural network (DNN) on a large dataset of English speech and then fine-tuned the model on a smaller dataset of Swahili speech. The authors reported that their proposed method outperformed the baseline approach, which only used the small dataset of Swahili speech.

In addition to data augmentation and transfer learning, other approaches have also been explored, such as unsupervised speaker adaptation and speaker diarization. Unsupervised speaker adaptation involves adapting a pre-trained model to a new speaker without requiring any labeled data. In a study by Gautam et al. (2019) [4], the authors proposed an unsupervised speaker adaptation method for speaker verification in Hindi, a low resource language spoken in India. The authors reported that their proposed method outperformed the baseline approach, which required labeled data.

In conclusion, the research on speaker verification for low resource languages is an emerging area of study, and several approaches have been proposed to address this issue. Data augmentation, transfer learning, unsupervised speaker adaptation and speaker diarization are some of the approaches that have been explored. While these approaches have shown promise, there is still much work to be done to develop accurate and reliable speaker verification systems for low resource languages.

3. Methods and materials

Consider the data that will be used in further experiments, some other materials and methods proposed to solve the problem under consideration.

3.1. Dataset Description

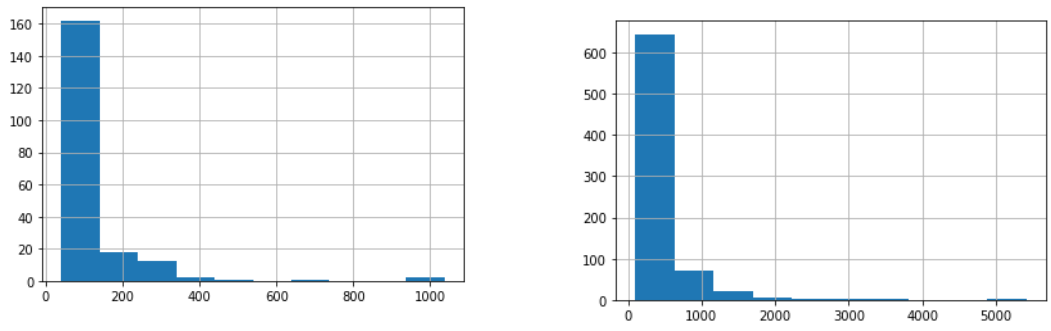
As a base dataset we have chosen Common Voice dataset [5]. It`s a crowd source dataset that contains a lot of audio recordings of different speakers even for low resource languages like Ukrainian. Large number of speakers is essential to build robust speaker verification system. The Ukrainian dataset contains 73 hours of recording of 120 speakers in train split and 14 hours of 639 speakers in test split.

In some experiments we additionally use datasets in other languages (language 1 and language 2) in training process. The data about duration and number of unique speakers is presented in Table 1.

Table 1
Number recoded hours and speakers per language

Language	Ukrainian	Language 1	Language 2
Duration, hours	87	215	1217
Number speakers	759	2731	6965

In order not to overfit on speakers with a few number of recordings we dropped speakers with less than 40 recordings from training dataset. On figure 1 shown the histograms of number of recordings per speaker.



a) b)

Figure 1: Distribution of number of recordings per speaker a) language 1 b) language 2

Also in some experiments we limit the number of recordings by speaker for Ukrainian recordings in order to make dataset no balanced.

As a step of feature extraction we split each audio into 3 second chunks and extracted spectrogram from them. After it we normalized them and from this step we could process them like images.

During the training process in some experiments, we applied Mell spectrogram augmentations such as time and frequency masking. This was done in order to prevent overfitting and make model robust to real-world data.

In order to evaluate model on real-world data we additionally collected recordings of interviews, department meeting in Google Meet, etc.

3.2. Methods

We have chosen as key metrics:

1) Equal error rate – it is one of the most widely used metric to evaluate speaker verification models.

2) Precision Top-K – spends for fraction of examples in Top-K most similar data points with the original one. In our experiments we used K equal to (3, 5, 10).

3) Mean and standard deviation of positive similarity – mean and standard deviation of cosine similarity between examples of the same class.

4) Mean and standard deviation of negative similarity – mean and standard deviation of cosine similarity between examples that do not belong to the same class as an original data point.

4. Experiment

We have chosen as a beck bone a ConvNext [6] because it`s one of the best performing convolutional neural network architectures in computer vision tasks, such as an ImageNet. We used randomly initialized weights, because Mel spectrograms are completely different, from datasets network was trained on so it`s unlikely that pre training will give an advantage in the task of speaker verification. Because of limitations in computation resources we have only tried to use small and tiny version of it.

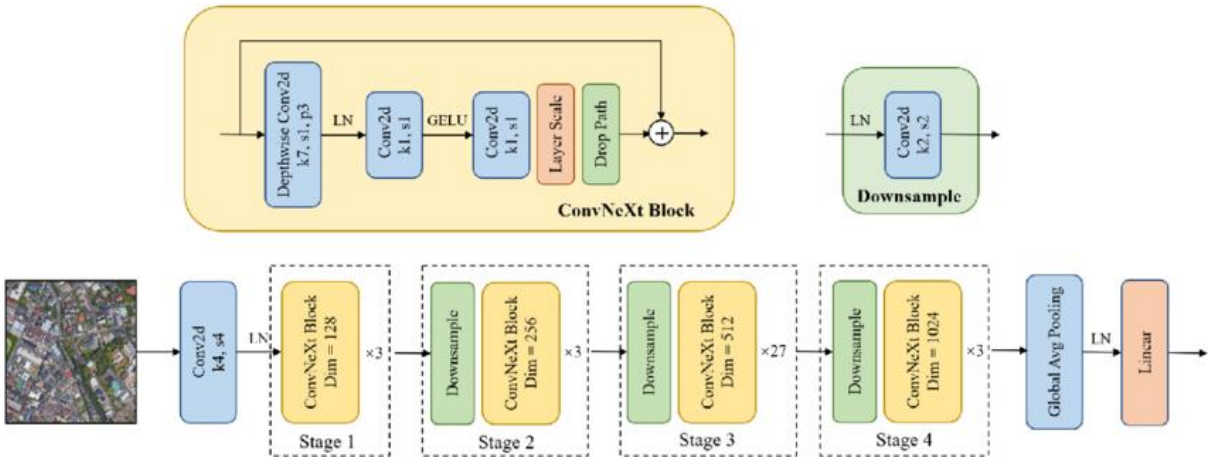


Figure 2: ConcNext architecture [6]

In order to improve the model we were experimenting with the DOLG architecture [7] which showed SOTA results in face recognition and image retrieval. It originally used ResNet as a backbone, so we have to adapt it to our task and ConvNext as a backbone.

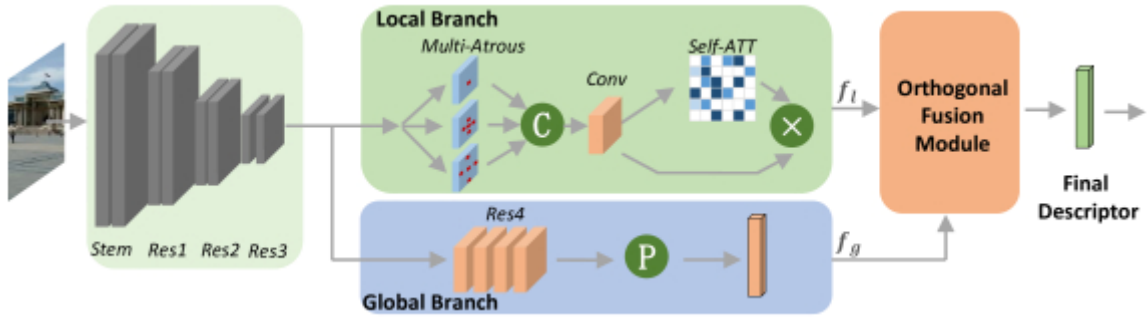


Figure 3: DOLG architecture [7]

As loss functions TripletLoss[8], ArcFace [9] and Sub center ArcFace [10]. Both of this metrics show the SOTA results in other metric learning tasks. The implementation of this losses was taken from python library Pytorch Metric learning [11]. The hyper parameter scale for both losses was chosen according to formula proposed in paper AdaCos[12]. As a sampling strategy for Triplet Loss we have chosen sampling semi hard negatives.

All networks were trained with the same initial learning rate and Cosine scheduler.

We performed several experiments.

1) Compared performance of ConvNext small and ConvNext tiny after training for 12 epochs. Only Ukrainian dataset was used as a training data. ArcFace was used as a loss function. This experiment will help to determine the best architecture to continue experiments.

2) Compared performance of ConvNext with size selected from previous experiment training for 12 epochs. Only Ukrainian dataset was used as a training data. In this experiment we compared

different loss function: Triplet loss, ArcFace loss, Sub center Arcface loss, ArcFace loss + Triplet loss, Sub center Arcface loss + Triplet loss.

3) Compare the performance of the best architecture from previous experiments training on large datasets, which includes other languages. Validation is performed only on Ukrainian dataset. This experiment will help to determine possibility of transferring skills from other languages in the task of speaker verification. Since the size of dataset is increased network is trained only for 4 epochs.

4) Compare the performance of the best model from previous experiment with the same model as a backbone in DOLG architecture. This experiment will help to identify the possibility of applying DOLG architecture in the task of speaker verification.

5) Compare the performance of the model from previous experiments trained on clean data and augmented data. This experiment is aimed to determine how the usage of augmented data effects the training proses.

After all of this experiments we perfumed speaker diarization on our dataset using the best model from previous experiments. To achieve it we split audio into parts of 3 seconds, transformed it to Mell spectrogram and got embedding by our model. Then they were clustered using KMeans algorithm.

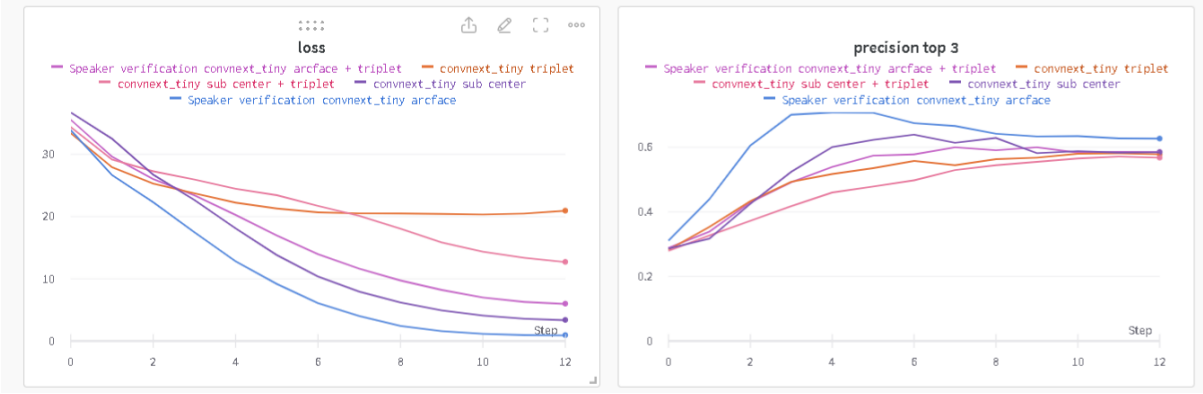
Training will be carried out in the Kaggle environment using P100 GPU.

5. Results

The results of the experiments are shown in Figures 4 – Figure 6 and in Tables 2. All the graphs are shown in the appendix A.

5.1. ML Results

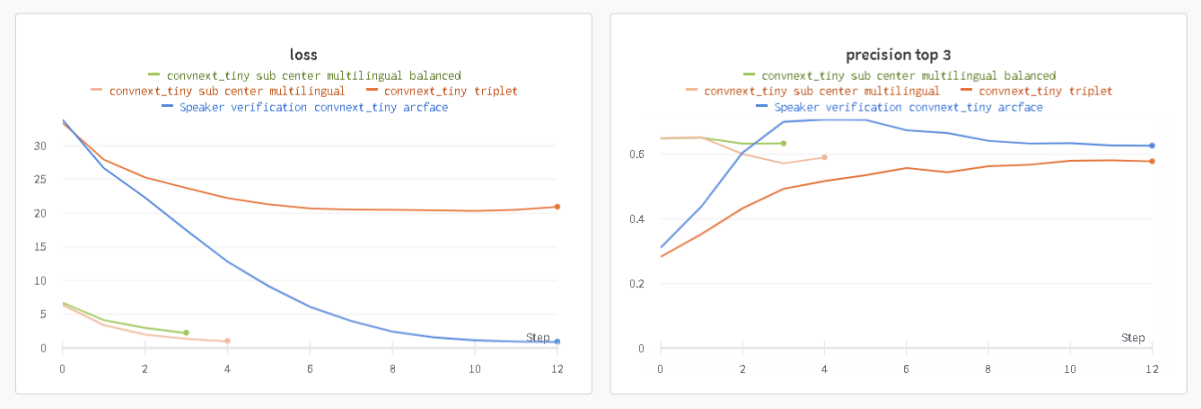
Figure 4 shows the change of loss and precision in the top 3 during the second experiment. Both metrics improve over the training process, bate after 6 epoch reaching the plateau.



a) b)

Figure 4: Change of metric during the second experiment a) loss b) precision at top 3

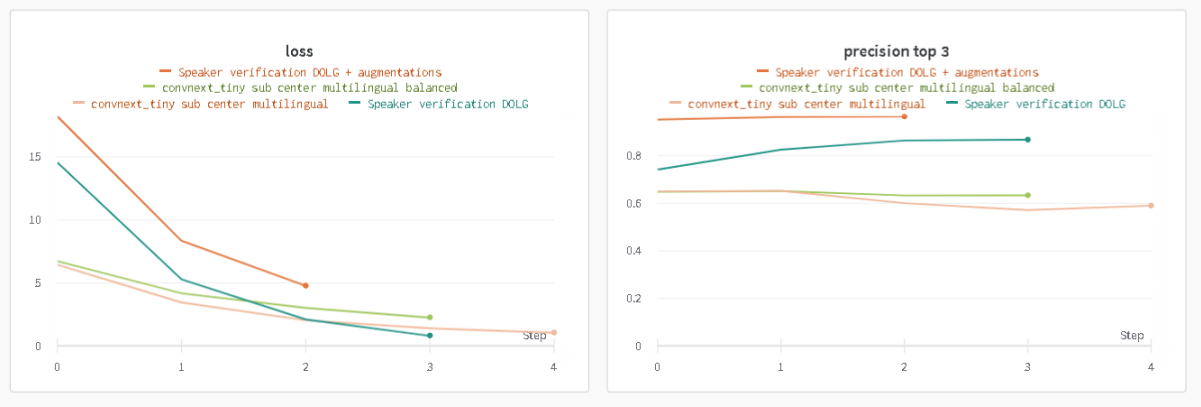
Figure 5 shows the change of loss and precision in the top 3 during the third experiment. Models trained on multilingual datasets achieved significantly better results, even after less training time.



a) b)

Figure 5: Change of metric during the third experiment a) loss b) precision at top 3

Figure 6 shows the change of loss and precision in the top 3 during the fourth and fifth experiments. DOLG architecture shows better initial performance and better performance in general. Also data augmentations slightly improved the model`s performance and robustness to new data.



a) b)

Figure 6: Change of metric during the fourth and fifth experiments a) loss b) precision at top 3

Table 2 shows all final performance metrics for all experiments.

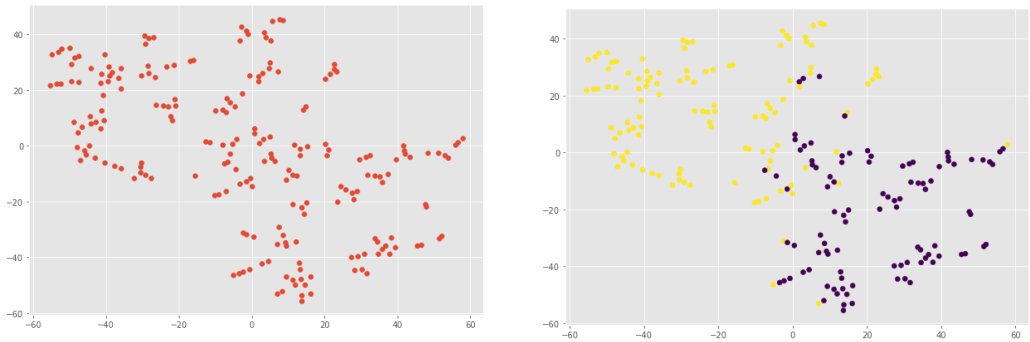
Table 2
Performance metrics of all experiments

Name	loss	mean_neg	mean_pos	std_neg	std_pos	eer_mean
Speaker verification DOLG + augmentations	4,779827	0,006445	0,512067	0,06615	0,191957	0,05745
convnext_tiny sub center multilingual balanced	2,266273	0,143715	0,621793	0,136854	0,188838	0,120028
convnext_tiny sub center multilingual	1,056141	0,098119	0,543737	0,112699	0,226304	0,130431
Speaker verification DOLG	0,819421	0,012565	0,559775	0,122927	0,183735	0,068941
convnext_tiny triplet	20,94974	0,064958	0,634587	0,18114	0,201048	0,116371
convnext_tiny sub center + triplet	12,71726	0,049018	0,682394	0,205036	0,192486	0,105805

convnext_tiny sub center Speaker verification	3,396142	0,05256	0,529312	0,126081	0,233282	0,115455
convnext_small Speaker verification	1,904608	0,183038	0,611239	0,102753	0,195134	0,127895
convnext_tiny arcface	0,952009	0,134064	0,55084	0,086243	0,219322	0,131675

5.2. Testing Results

In order to test model performance on the real-world data we performed speaker diarization of Google Meet call between 2 speakers. First of all, we split the audio into windows of 3 seconds each. Next we transformed the raw audio into mel spectrograms and extracted embeddings using our model. These embeddings were clustered using KMeans algorithm. The results of clusterization are shown on figure 7.



a) b)

Figure 7: TSNE progections of voice embeddings and clusterization result

As we can see from the plots, there are 2 large clusters which represents speakers. The boundary region between clusters represents fragments, where both speakers are active.

As a next step each embedding was matched with the corresponding timestamp. The result was formatted according to srt format and is shown on figure 8.

```
1
00:00:00,000 --> 00:00:09,000
Speaker 0

2
00:00:09,000 --> 00:00:24,000
Speaker 1

3
00:00:24,000 --> 00:00:30,000
Speaker 0

4
00:00:30,000 --> 00:00:54,000
Speaker 1

5
00:00:54,000 --> 00:01:03,000
Speaker 0
```

Figure 8: Speaker diarization example

In conclusion, model trained for speaker verification showed good results in the task of speaker diarization on real-world data.

6. Discussions

As a result of the first experiment it was shown, that even for low-resource languages models can achieve quite good performance metrics. Also results of both convnext tiny and convnext small are quite similar. For both networks we can see that after the 6th epoch the precision at n starts to decrees or stay approximately the same. That may indicate the overfitting of the networks. Also after the 6th epoch negative std reached plateau and don't decrease as fast as before. On the other hand, std of positive examples is constantly increasing over training. So it was decided to use convnext tiny, because it has less parameters, so following experiments can be performed faster. The question of performance of large networks (like base, or large) is still open, so probably they can perform better in the task of speaker verification.

In the second experiment we compared different loss functions. In the end all networks performed approximately the same. But losses that contain triplet loss performed a bit worse than ArcFace and sub center ArcFace. These losses reached plateau faster and convergence slower. In general, all the metrics follow the same trend like in previous one. We have chosen sub center arcface because it shows more robustness to a new data, while keeping good performance metrics.

In the third experiment we compared the model trained only on one language with trained on multilingual dataset. Multilingual models show a superior metrics on test Ukrainian set and achieve better results in general. But the model that was trained on fully multilingual dataset reached plateau faster than one trained one balanced (where number of recordings per speaker is approximately the same as in a target language), which may indicate overfitting to languages with more speakers. So transferring of skills for low resource languages, like Ukrainian, from other languages is quite effective, but in order to achieve better results, dataset should be balanced.

In the fourth experiment we compared ConvNext with DOLG pipeline with ConvNext as a backbone on balanced multilingual dataset. DOLG shows superior results, and pretty much achieved SOTA result in the task of speaker verification. In addition, it was trained only for 6 epochs, so it may possible achieve better results with further training.

In the fifth experiment we applied augmentations to spectrogram and repeated previous experiment. As a result, the model achieved even better level of performance and robustness.

Next we tried to analyze with the help of the model from last experiment real-word data – Google Meet call of 2 people. So it performed quite well. However, sometimes if there was no sound, the model can treat it as a separate speaker. So in conclusion we recommend to use ConvNext tiny as a backbone in DOLG pipeline to achieve SOTA results.

7. Conclusions

The paper presents the study on speaker verification using deep learning models. The study used four key metrics to evaluate the performance of the models: equal error rate, precision top-K, mean and standard deviation of positive similarity, mean and standard deviation of negative similarity. The study compared the performance of different network architectures, such as ConvNext and DOLG, and different loss functions, such as TripletLoss, ArcFace, and Sub center ArcFace. The study also compared the performance of models trained on single language datasets and those trained on multilingual datasets.

The experiments showed that even for low-resource languages, the models can achieve quite good performance metrics. The results indicated that the ConvNext tiny model performed better than the ConvNext small model. The study also found that Sub center ArcFace loss showed more robustness to new data while maintaining good performance metrics. Furthermore, the study showed that transferring skills from other languages to low-resource languages was quite effective in achieving better performance metrics. Finally, the study performed speaker diarization on the dataset using the best model from previous experiments, achieving good results.

In conclusion, the study demonstrated the effectiveness of deep learning models in the task of speaker verification, even for low-resource languages. The study provides insights into the best-performing network architectures and loss functions for this task and shows the potential for transferring skills from other languages to low-resource languages. The findings of this study could have significant implications for developing better speaker verification systems.

The perspective of future studding includes comparison of large amount of convolutional neural network architectures (especially with large number of parameters), different loss functions and their combinations. Also it`s quite important to study transfer learning from other languages and perform multilingual speaker verification.

8. References

[1] E. Erdem et al., 'Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning'. 06-Apr-2022.

[2] R. Zevallos, 'Text-To-Speech Data Augmentation for Low Resource Speech Recognition'. arXiv, 2022.

[3] H. Gelas, L. Besacier, and F. Pellegrino, 'Developments of Swahili resources for an automatic speech recognition system', in Workshop on Spoken Language Technologies for Under-resourced Languages, 2012.

[4] N. Brummer, A. Mccree, S. Shum, D. Garcia-Romero, and C. Vaquero, 'Unsupervised Domain Adaptation for I-Vector Speaker Recognition', in Proc. The Speaker and Language Recognition Workshop (Odyssey 2014), 2014, pp. 260–264.

[5] R. Ardila et al., 'Common Voice: A Massively-Multilingual Speech Corpus', in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, 'A ConvNet for the 2020s', in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976.

[7] M. Yang et al., 'DOLG: Single-Stage Image Retrieval with Deep Orthogonal Fusion of Local and Global Features', 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11752–11761, 2021.

[8] E. Hoffer and N. Ailon, 'Deep Metric Learning Using Triplet Network', in Similarity-Based Pattern Recognition, 2015, pp. 84–92.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, 'ArcFace: Additive Angular Margin Loss for Deep Face Recognition', in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4685–4694.

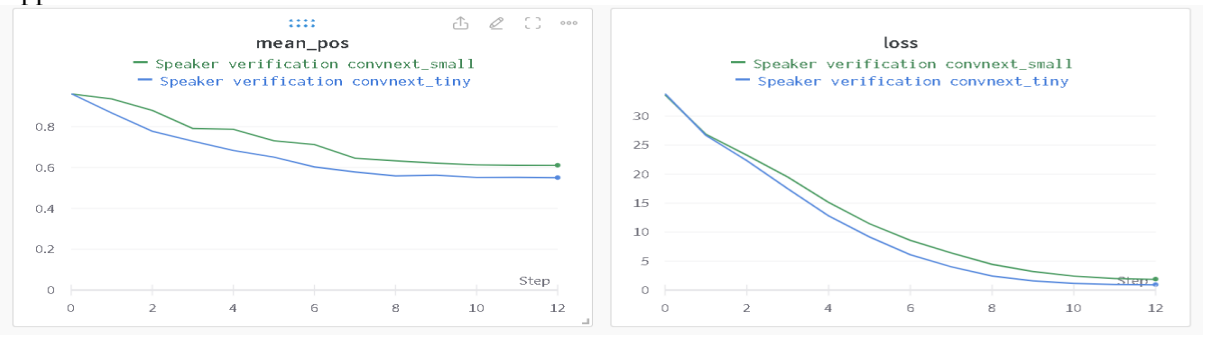
[10] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, 'Sub-center ArcFace: Boosting Face Recognition by Large-Scale Noisy Web Faces', in Computer Vision -- ECCV 2020, 2020, pp. 741–757.

[11] K. Musgrave, S. Belongie, and S.-N. Lim, 'PyTorch Metric Learning'. arXiv, 2020.

[12] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, 'AdaCos: Adaptively Scaling Cosine Logits for Effectively Learning Deep Face Representations', 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10815–10824, 2019.

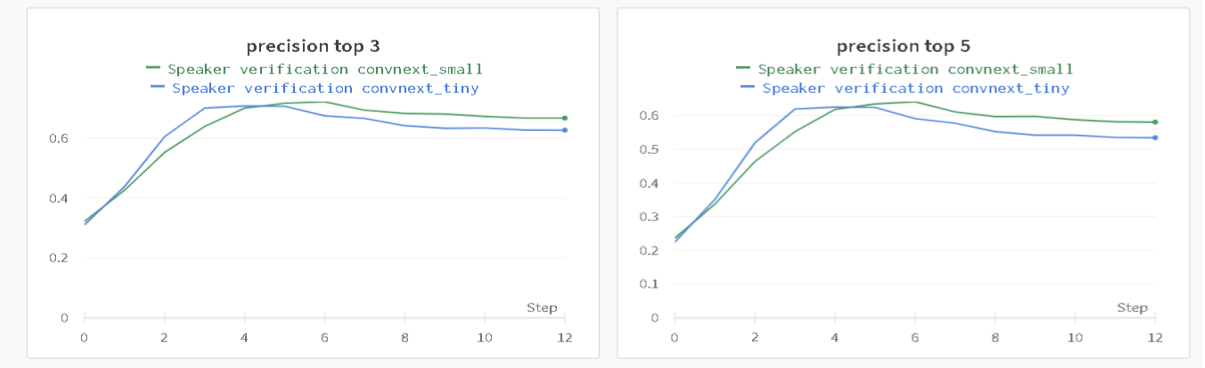
[13] A. L. Yerokhin, A. S. Babii, A. S. Nechyporenko, O. P. Turuta, A Lars-Based Method of the Construction of a Fuzzy Regression Model for the Selection of Significant Features, Cybernetics and Systems Analysis, Vol. 52, Issue 4, (2016), 641–646. https://doi.org/10.1007/s10559-016-9867-5

Appendix A



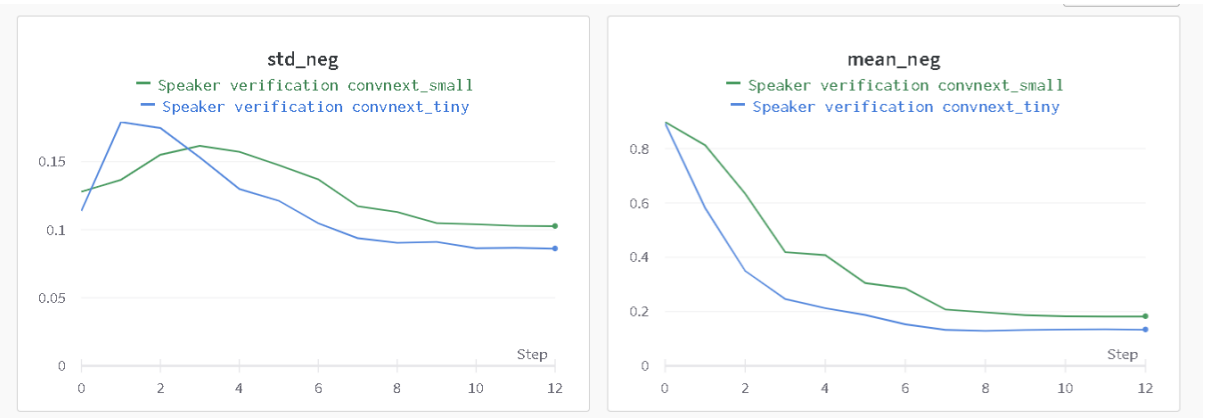
a) b)

Figure A.1: Change of metric during the first experiment a) mean positive similarity b) loss



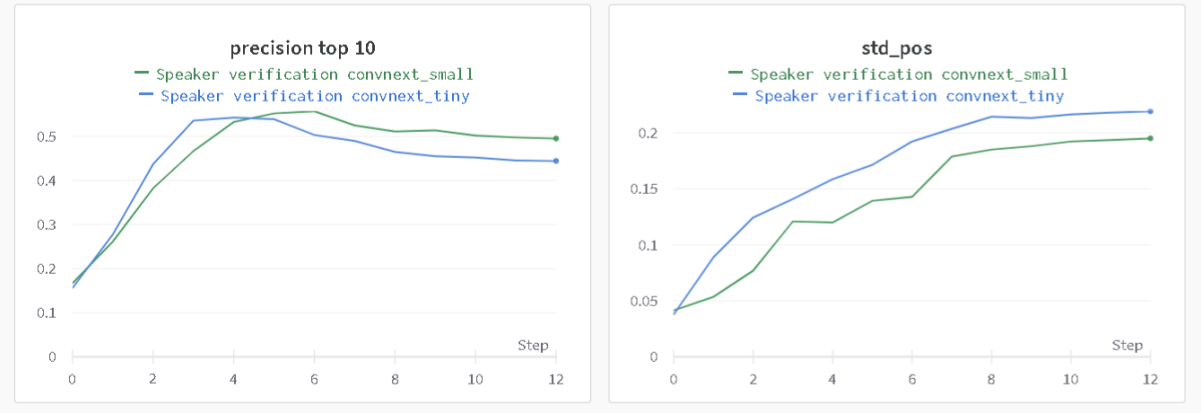
a) b)

Figure A.2: Change of metric during the first experiment a) precision at top 3 b) precision at top 3



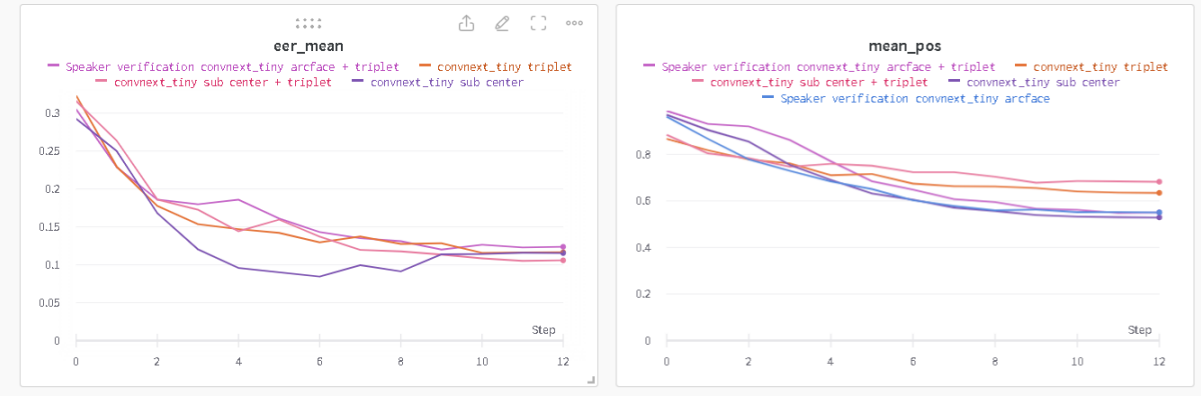
a) b)

Figure A.3: Change of metric during the first experiment a) negative standard deviation b) mean negative similarity



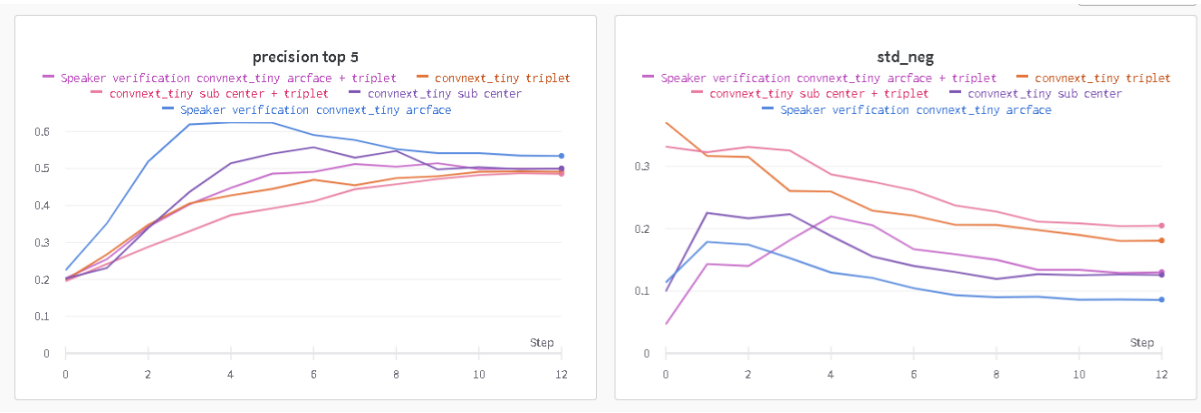
a) b)

Figure A.4: Change of metric during the first experiment a) precision at top 10 b) positive standard deviation



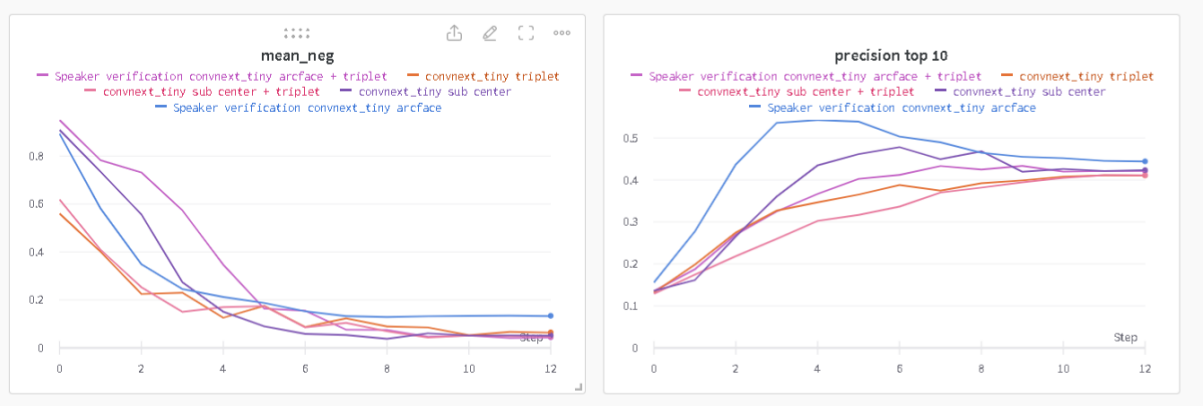
a) b)

Figure A.5: Change of metric during the second experiment a) equal error rate b) mean positive similarity



a) b)

Figure A.6: Change of metric during the second experiment a) precision at top 5 b) negative standard deviation



a) b)

Figure A.7: Change of metric during the second experiment a) mean negative similarity b) precision at top 10

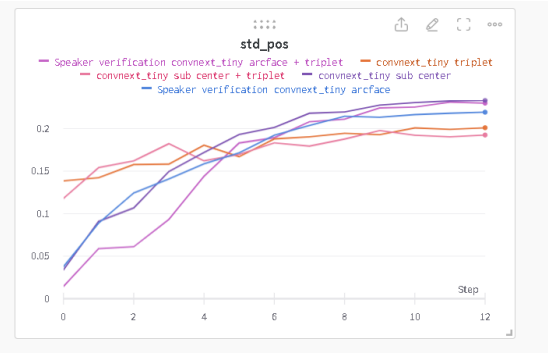
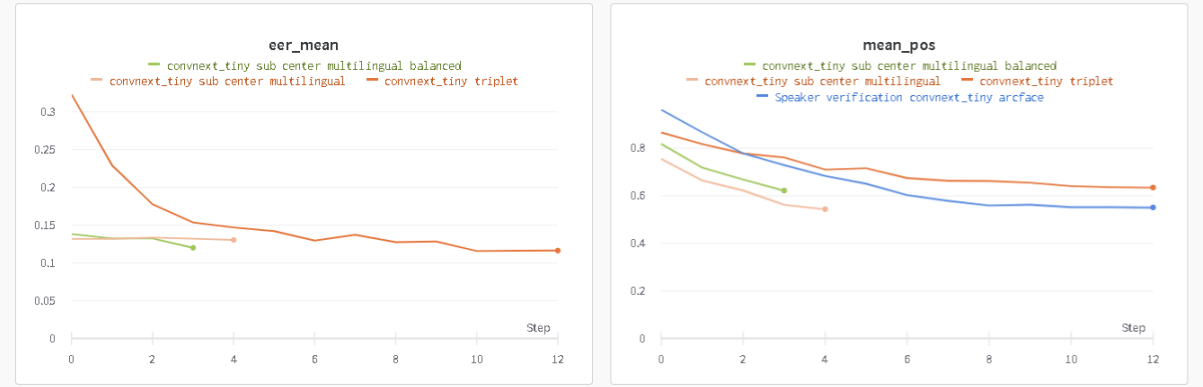
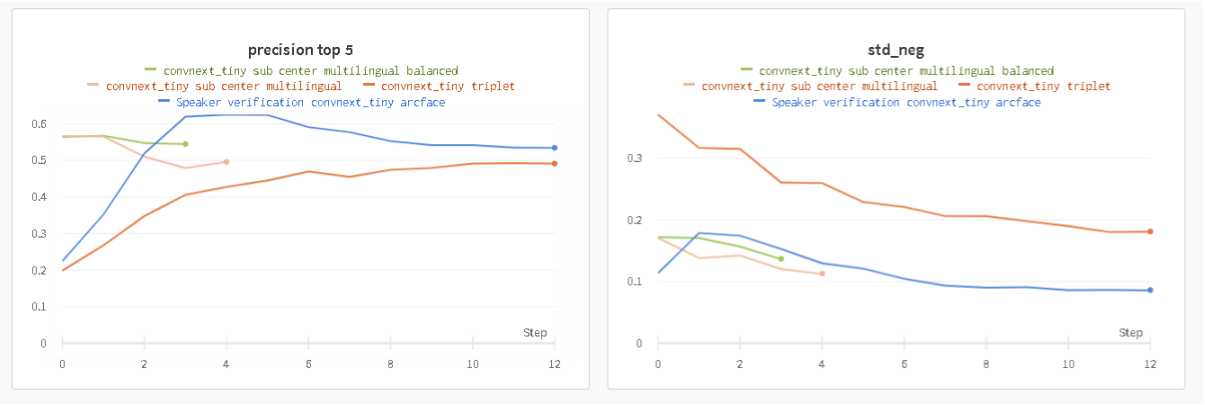


Figure A.8: Change of positive standard deviation during the second experiment



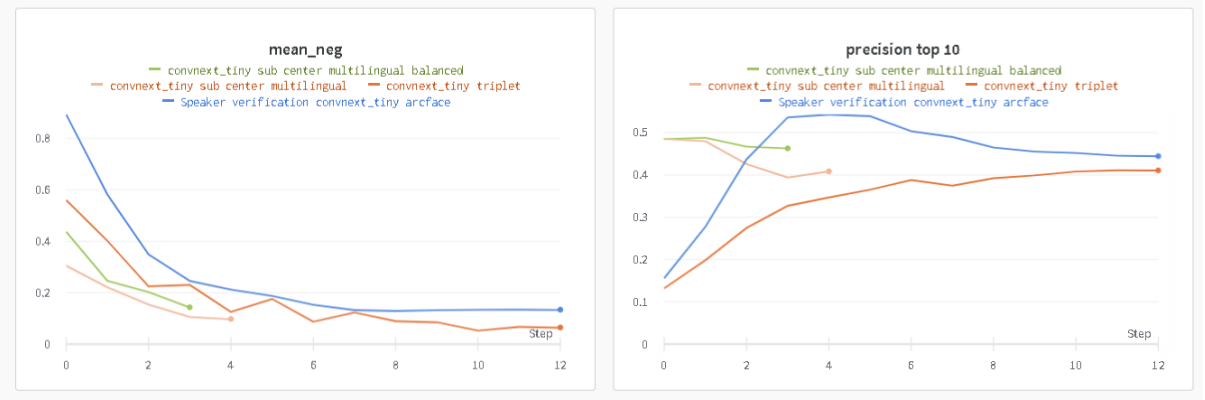
a) b)

Figure A.9: Change of metric during the third experiment a) equal error rate b) mean positive similarity



a) b)

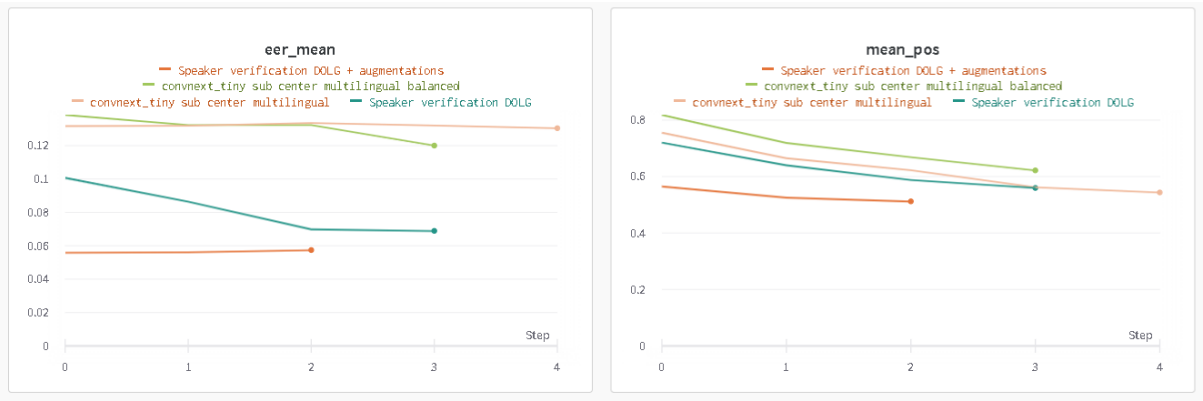
Figure A.10: Change of metric during the third experiment a) precision at top 5 b) negative standard deviation



a) b)

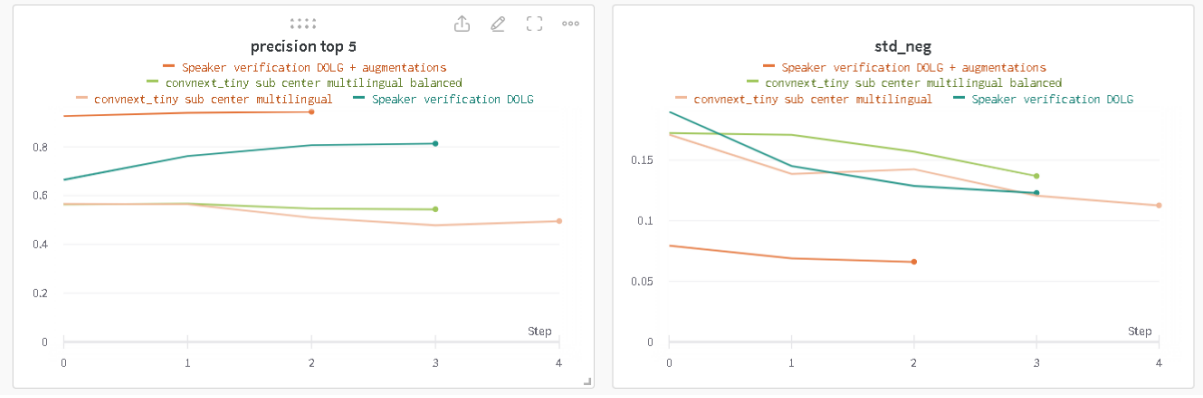
Figure A.11: Change of metric during the third experiment a) mean negative similarity b) precision at top 10

Figure A.12: Change of positive standard deviation during the third experiment



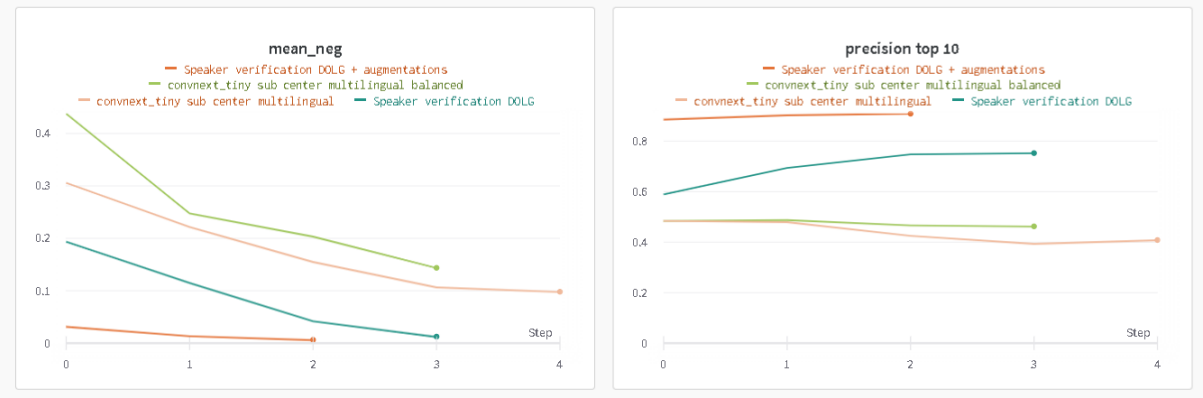
a) b)

Figure A.13: Change of metric during the fourth and fifth experiments a) equal error rate b) mean positive similarity



a) b)

Figure A.14: Change of metric during the fourth and fifth experiments a) precision at top 5 b) negative standard deviation



a) b)

Figure A.15: Change of metric during the fourth and fifth experiments a) mean negative similarity b) precision at top 10

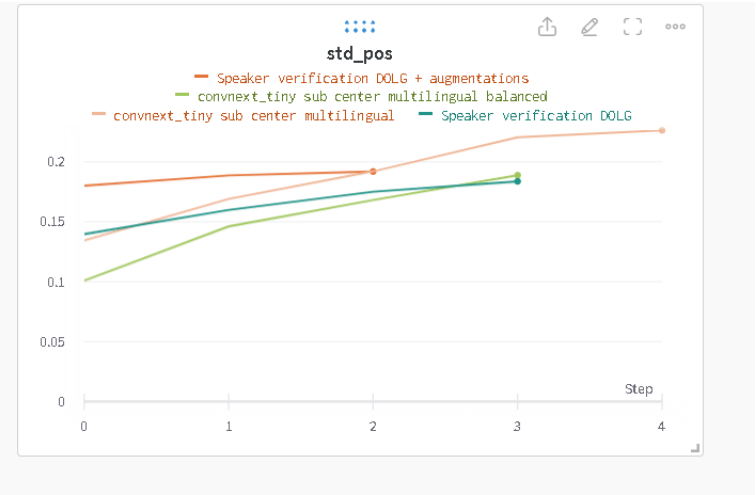


Figure A.16: Change of positive standard deviation during the fourth and fifth experiments