Text-independent Speaker Verification System by Short-Term Voice Phrases for SdSV Challenge

Yakov Filin[1,2], *Andrey Lependin*[2]

[1]Estesis, LLC
[2]Altai State University
j@estesis.tech, andrey.lependin@gmail.com

Abstract

The quality of work of voice verification systems (ASV) deteriorates with a decrease in the information that needs to be analyzed by voice verification models. Traditional state-of-the-art systems usually work with audio files lasting about 10 seconds to guarantee the quality of work in security systems. However, such a work scenario is not applicable, for example, to voice assistants and call centers, where the system user for various reasons cannot speak for such a long time and has a voice clip lasting about 2-3 seconds. In this case, the voice of people needs to be confirmed by very short speech fragments. To solve this problem, within the framework of the SDSV Challenge 2020 [1] competition, the task of verifying users by short phrases was proposed. A solution was proposed that increased the quality of the base model from 10.67% to 4.4% by the EER metric.

Index Terms: speaker verification, short-term automatic speaker verification, voice biometrics

1. Introduction

The speaker verification technology provides a high-quality, low-cost biometric solution for the financial and banking sectors, forensics, commercial security systems, corporate call centers.

Modern voice biometric verification systems, as a rule, are based on deep neural networks (DNNs) and already become traditional i-vector models on universal background models of Gaussian mixtures (GMM-UBM) [3]. The architecture of DNN systems, as a rule, contains three main modules:

- Feature extraction module with pre-processing, payload allocation using Voice Activity Detection (VAD) approaches;
- A model of voice verification that generates a low-level representation (template) of a speech fragment. Post-processing of low-level representations from the neural network, for example, using linear discriminant analysis (LDA), can also be used here;
- A module for comparing voice patterns for similarity items. As a rule, for this, an enrollment template of reference speaker samples is formed, with which the verified voice sample is then compared.

It is worth noting that voice verification systems are dependent on external environmental conditions, for example, the level of noise characteristics, the level of reverberation of the room, communication channel, and the quality of the recording microphone. In this paper, we will consider the development of a verification system for short phrases with a duration of about 3 s in various noise conditions on voice data sets Voxceleb1 / 2, Librispeech, DeepMine (text-independent data set of the SDSV Challenge 2020 competition).

2. Feature Extraction Description

Input audio files are presented in the form of sound files in WAV format with a sampling frequency of 16 kHz, in the case of files with a higher sampling frequency, they were resampled to 16 kHz.

To obtain the audio payload, the Voice Activity Detection approach of the WebRTC VAD library was used [4].

In the framework of the work, two types of signs were used:

- 40-dimensional MFCC [5] with a window length of 25 ms and an overlap of 10 ms.
- FFT-power spectrograms obtained with their 512 FFT coefficients, a window width of 25 ms and an overlap of 10 ms.

When generating the appropriate features from the audio file, a random segment of 3-5 seconds was selected.

3. Datasets

This section describes the source datasets.

3.1. Voxceleb1

The VoxCeleb1 data set [6] contains more than 100,000 audio recordings from 1251 speakers. They were obtained from the videos of the social network YouTube and represent the voices of famous people. The data set is gender balanced (55% of men and 45% of women).

Table 1: *Voxceleb1 database structure.*

	Train	**Test**
Speakers	1211	40
Audio	148642	4874

3.2. Voxceleb2

The VoxCeleb2 dataset [7] was a radical extension of the VoxCeleb1 dataset. The data in VoxCeleb2 is represented by audio recordings obtained from YouTube videos obtained under various conditions.

Table 2: *Voxceleb2 database structure.*

	Dev	Test
Speakers	5994	118
Audio	1092009	36237

3.3. DeepMine (Text-Independent)

The database [2] contains the voices of more than 1850 speakers with a total duration of more than 540 hours in Persian and English. This database has good coverage of accents, age of speakers and gender for text-dependent and text-independent verification of speakers.

3.4. Librispeech

Freely distributed dataset [8] of this speech, containing 100 hours of English speech, cut from the files of the audio books of the LibriVox project.

4. Neural Network Architecture And Training

In this paper, we used two types of neural network architectures: time-delay neural networks (TDNNs), which have become one of the qualitative methods for generating x-vectors and a modified resnet architecture.

The feature vector for TDNN architecture was melf-frequency cepstral coefficients (MFCC), for resnet-based architecture, the Fourier spectrum (FFT).

4.1. Baseline

As a basic model, the E-TDNN network model was considered [9]. A trained low-level performance reduced the dimension to 150 using LDA. Score was evaluated by the PLDA model, trained on the Voxceleb1 set without score normalization.

4.2. TDNN

The neural network was based on the architecture of the basic x-vector network [10] with modified contexts of TDNN blocks and the adaptive cosine scaling loss function [11] for calculating the error. This approach promotes rapid convergence of the neural network and works commensurate with Angular Softmax in quality. The size of the low-level embedding layer in this work was 512. The metric of similarity of the verified samples was the cosine distance.

Table 2: *TDNN Neural Network Architecture.*

Layer	Layer context	Layer Output Size
TDNN-1	[t-2, t+2]	512
TDNN-2	{t-3, t, t+3}	512
TDNN-3	{t-3, t, t+3}	512
TDNN-4	{t}	512
TDNN-5	{t}	1500
StatPool	[0, T)	3000
FC layer	{0}	512
Batch-Norm	{0}	512
Adaptive Scaling Cosine Logits ([11])	{0}	N_speakers

4.3. Resnet-Based (SE-Resnet)

As a resnet network, the classic 34-layer ResNet neural network with a modified base unit was used. The base block was SE (Squeeze-Execution block), borrowed from the squeezenet architecture [11].

This neural network was trained on the grounds of FFT.

The embedding layer was 512 in size. Adaptive Scaling Cosine Logits was also used as a loss functional.

The general architecture of the SENet module [13] - block is shown in Figure 2.

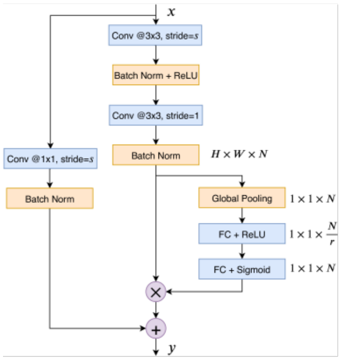

Figure 1: *SE-block*

4.4. Training and hyperparameters

The training of current neural architectures was done in Python using the Pytorch framework. For training, a workstation with 2 1080TI graphics cards was used.

As a neural network optimizer, the DiffGrad optimizer with an initial learning speed of 1e-3 was used. Learning speed changed on a schedule using ReduceLrOnPlateau.

It is worth noting that in order to achieve the current level of quality, the neural network SE-Resnet studied 2 times longer compared to TDNN architecture (5 days compared to 2.5 days).

5. Results

The assessment of the basic speaker verification system is presented in Table 2 and amounted to 10.67% according to the EER metric on the Deepmine dataset. Its improvements were the proposed TDNN and Se-Resnet models.

5.1. Quality assessment of single-systems

Table 3: *Baseline model quality*

EER	MinDCF	Set
10.67	**0.4319**	**progress**
8.31	0.3216	progress-male
11.73	0.4419	progress-female
9.61	0.4111	progress-EN
6.17	0.2949	progress-FA
7.04	0.2759	progress-EN-male
10.55	0.4387	progress-EN-female
4.24	0.2053	progress-FA-male

6.59	0.2950	progress-FA-female
10.67	0.4324	evaluation
8.26	0.3239	evaluation-male
11.71	0.4430	evaluation-female
9.58	0.4118	evaluation-EN
6.14	0.2962	evaluation-FA
7.11	0.2770	evaluation-EN-male
10.58	0.4374	evaluation-EN-female
4.25	0.2062	evaluation-FA-male
6.53	0.2953	evaluation-FA-female

Each of the proposed models was trained at time intervals of 3-5 seconds for the TDNN system and 3 seconds for the SE-Resnet system, which was controlled during the formation of the mini-training batch. If the length of the audio file was more than 3-5 seconds, then a random audio fragment was selected. If the file was shorter, the audio file was mirrored relative to the end of the audio file and the necessary time periods were allocated.
Each mini-batch in training consists of 256 examples.
The quality of work from each proposed system is separately presented in tables 3-4.

Table 3: *SE-Resnet model quality*

EER	MinDCF	Set
5.25	**0.2270**	**evaluation**
5.25	0.2281	**progress**

Detailed information on the TDNN-based approach is presented in Table 4.

Table 4: *TDNN model quality*

EER	MinDCF	Set
5.40	**0.2332**	**progress**
4.06	0.1883	progress-male
6.21	0.2602	progress-female
6.51	0.2967	progress-EN
4.06	0.1774	progress-FA
5.23	0.2443	progress-EN-male
7.27	0.3267	progress-EN-female
2.66	0.1369	progress-FA-male
4.92	0.2022	progress-FA-female
5.39	**0.2328**	**evaluation**
4.04	0.1876	evaluation-male
6.20	0.2599	evaluation-female
6.46	0.2955	evaluation-EN
4.07	0.1773	evaluation-FA
5.23	0.2457	evaluation-EN-male
7.17	0.3241	evaluation-EN-female
4.25	0.2062	evaluation-FA-male
6.53	0.2953	evaluation-FA-female

5.1. Score Fusion

For fusion the score of the proposed systems, the method of averaging these two values was used.
The results are presented in Table 5.

Table 5: *Fusion (Primary) model quality*

EER	MinDCF	Set
4.46	**0.1942**	**progress**
3.44	0.1580	progress-male
5.06	0.2163	progress-female
5.31	0.2471	progress-EN
3.21	0.1402	progress-FA
4.30	0.2057	progress-EN-male
5.91	0.2715	progress-EN-female
2.14	0.1076	progress-FA-male
3.85	0.1605	progress-FA-female
4.45	**0.1945**	**evaluation**
3.42	0.1586	evaluation-male
5.04	0.2163	evaluation-female
5.28	0.2471	evaluation-EN
3.21	0.1402	evaluation-FA
4.34	0.2083	evaluation-EN-male
5.82	0.2700	evaluation-EN-female
2.10	0.1074	evaluation-FA-male
3.89	0.1604	evaluation-FA-female

Figure 2 shows the DET curve of the developed system.

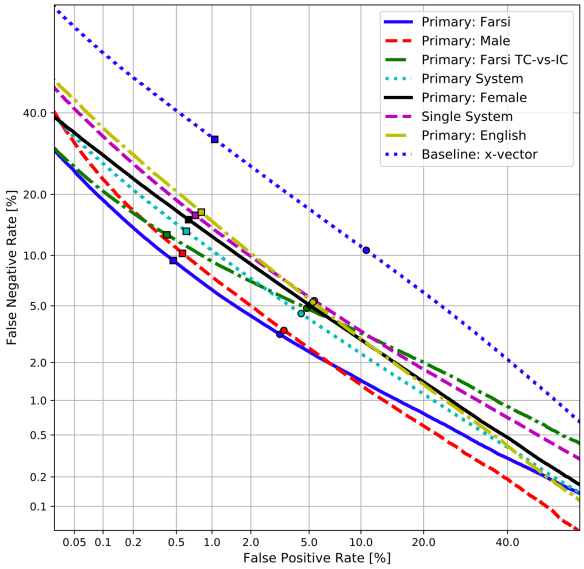

Figure 2: *DET-curve.*

According to estimates on the evaluation subset of the DeepMine dataset, the fused Primary system showed better results compared to the base model and improved it from 10.67% to 4.46% by the EER metric.

6. Acknowledgements

This article demonstrates the efficiency of deep neural networks based on TDNN and SE-Resnet architectures for the task of text-independent verification of speakers. It should be noted that the system adapts well to different languages, as She studied mainly in English, and also works steadily at Farsi. The result of a simple fusion averaging score allowed us to get high-quality predictions for the SDSV Challenge.

7. References

[1] Zeinali H. et al. Short-duration Speaker Verification (SdSV) Challenge 2020: the Challenge Evaluation Plan // arXiv preprint arXiv:1912.06311. – 2019.

[2] H. Zeinali, H. Sameti, and T. Stafylakis, "DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English." in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.

[3] Campbell W. M., Sturim D. E., Reynolds D. A. Support vector machines using GMM supervectors for speaker verification //IEEE signal processing letters. – 2006. – T. 13. – №. 5. – C. 308-311.

[4] Google WebRTC. https://webrtc.org/. Accessed 20 Mar 2016

[5] Logan B. et al. Mel frequency cepstral coefficients for music modeling //Ismir. – 2000. – T. 270. – C. 1-11.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a largescale speaker identification dataset," in INTERSPEECH, 2017.

[7] Chung J. S., Nagrani A., Zisserman A. Voxceleb2: Deep speaker recognition //arXiv preprint arXiv:1806.05622. – 2018.

[8] Panayotov V. et al. Librispeech: an asr corpus based on public domain audio books //2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2015. – C. 5206-5210.

[9] Liu Y. et al. THUEE system description for NIST 2019 SRE CTS Challenge //arXiv preprint arXiv:1912.11585. – 2019.

[10] Snyder D. et al. X-vectors: Robust dnn embeddings for speaker recognition //2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2018. – C. 5329-5333.

[11] Zhang X. et al. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2019. – C. 10823-10832.

[12] Iandola F. N. et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size //arXiv preprint arXiv:1602.07360. – 2016.

[13] Hu J., Shen L., Sun G. Squeeze-and-excitation networks //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – C. 7132-7141.