



A Subnetwork Approach for Spoofing Aware Speaker Verification

Alexander Alenin¹, Nikita Torgashov¹, Anton Okhotnikov¹,
Rostislav Makarov¹, Ivan Yakovlev¹

¹ID R&D Inc., New York, USA

{alenin,torgashov,okhotnikov,makarov,yakovlev}@idrnd.net

Abstract

This paper describes the ID R&D team submission to the Spoofing Aware Speaker Verification (SASV) challenge. Firstly, we present an approach that utilizes automatic speaker verification (ASV) system together with countermeasures (CM) subsystem in a single computational graph, called an anti-spoofing subnetwork. Subnetwork is a small network operating on feature maps from larger parent network, in this case trained for a speaker verification task. While requiring a small number of additionally trained parameters, subnetwork approach showed great performance in spoofing attacks detection task. Secondly, we cover training strategies for independently trained ASV and CM systems. In addition, we present a SASV-EER optimization approach using a fusion of multiple systems outputs and quality measurement functions (QMFs). Our best fusion achieves **0.136%** EER on SASV-2022 evaluation set, while the smallest single-model system with 11.6M parameters achieves **0.223%** EER.

Index Terms: Speaker recognition, Voice anti-spoofing, ASVSpooF 2019, SASV Challenge 2022

1. Introduction

Real-life applications of Automatic Speaker Verification (ASV) systems usually require a spoofing detection system operating over the same input audio. There are multiple types of spoofing attacks on a speech signal. One of them is a replay of a recorded target speaker's speech, called a replay attack or Physical Access (PA). The other type of attack implies the Text-to-speech engine or voice conversion system to synthesize the speech of a target speaker, called a synthesized attack or Logical Access (LA). A final type of spoofing attack involves the mimic of a target speaker's voice and is called an impersonation. While impersonation-like attacks might be rejected by a high-quality ASV system, PA and LA attacks require distinct algorithms trained for anti-spoofing task specifically.

There were organized multiple challenges targeting detection of a spoofed speech: [1], [2], [3], [4], however all of them are focused on spoofing detection task only, and not on the combination of speaker verification with spoofing detection task. SASV challenge aims to close this gap: it suggests participants to develop Automatic Speaker Verification (ASV) systems that can reject impostor access attempts while being robust against spoofing attacks at the same time. Participants are required to build a framework optimising the ASV systems operating in tandem with countermeasures (CM) systems.

In order to develop an integrated SASV solution researchers are encouraged to investigate the possibilities to train a single-model system operating with spoofing and verification embeddings and scores, or trained in a multi-task end-to-end fashion combining ASV and CM losses to minimize the SASV EER.

In Section 2 of a paper we present the models' architectures,

used loss function and input features hyper-parameters. Section 3 gives an overview of datasets used, training data augmentations and models' training stages. We also present our fusion scheme and used Quality Measurement Functions (QMFs) in this section. Sections 4 and 5 contain Results and Conclusions respectively.

2. System Setup

2.1. Input features

We extract 80-dimensional Mel filter bank log-energies with a 25 ms frame length and 10 ms step with an FFT size of 512 over the 20-7600 Hz frequency limits. After feature extraction, we subtract the mean along the time axis. To test models, we used 8-second input segments.

2.2. Architectures

Both models in our submission are based on the residual neural network [5] architecture. Deep ResNets have made a breakthrough in image classification task and have recently been efficiently applied to the speaker recognition task [6], [7]. As a baseline we used ResNet-34 architecture from [7]. Since the deeper models usually show performance improvements provided enough training data, we applied some modifications to the baseline ResNet-34. In particular, we have run a series of experiments and optimised a number of residual blocks and a number of filters in each residual block to increase the capacity of architectures. As a result, two modifications of the ResNet-34 model with 48 and 100 hidden layers were used.

Both detailed architectures are shown in the Table 1 and corresponding verification testing results on VoxCeleb1-test are presented in the Table 2, where C_{FA} and C_{Miss} equals to 1, and P_{target} equals to 0.01 for MinDCF metric.

2.3. Subnetwork Approach

Our headliner in SASV-2022 Challenge is a subnetwork. Subnetwork is a novel technique for training neural network models for downstream tasks. The method is similar to a transfer learning approach, however subnetwork is mainly suitable for training the subtask models related to general task. In our approach, we firstly trained a backbone for the main speaker verification task. Further, we trained a small model on top of the frozen verification backbone as a second classification head to solve the anti-spoofing problem. Unlike transfer learning methods, our approach is far ahead in terms of accuracy and size. In addition it also preserves the backbone embedding for dealing with the main speaker verification task. The subnetwork itself is a small neural network that leverages the performance of the main backbone: it utilizes multiple verification ResNet backbone layers' outputs by using the low- and high-level features that appear to be good speech representations when processed all together.

Table 1: Models architectures

Layer name	Output (C × F × T)	ResNet48	ResNet100
Conv2D	C × 80 × T	96, 3×3, stride=1	128, 3×3, stride=1
ResBlock-1	C × 80 × T	$\begin{bmatrix} 3 \times 3, 96 \\ 3 \times 3, 96 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 6$
ResBlock-2	C × 40 × T/2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 8$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 16$
ResBlock-3	C × 20 × T/4	$\begin{bmatrix} 3 \times 3, 160 \\ 3 \times 3, 160 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 24$
ResBlock-4	C × 10 × T/8	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$
Flatten (C, F)	2560 × T/8	—	
Pooling	5120	StatsPooling	
Dense	256	—	
AM-Softmax	Num. of speakers	—	

Table 2: Results on VoxCeleb1-O Standard protocol

Model	EER [%]	MinDCF
ResNet48	1.09	0.102
ResNet100	0.90	0.069

Table 3: Results on ASVspoof 2019 LA protocol

Model	EER [%]	MinDCF
AASIST [8]	0.83	0.0275
r48-cm-cls	0.53	0.0156

The architecture of a proposed subnetwork is similar to the one described in [9]. Detailed subnetwork scheme for ResNet48 is shown on Figure 1. We downsample outputs of first convolutional layer and outputs of all four consequent stages of a ResNet backbone using the point-wise convolutions followed by the BatchNorm layer and ReLU activation. We then process each of downsampled feature maps by concatenating them along the channels axis to a current subnetwork’s feature map. After that, we apply a Depthwise separable 2D convolution with a 3x3 kernel, BatchNorm, ReLU and MaxPool layers to fit the spatial dimensions of tensor, matching the height and width of the next downsampled feature map of a backbone. The proposed approach allows to obtain a compact neural network with a total of 600K trainable subnetwork parameters only. In Table 3 we present ASVspoof 2019 LA testing results for ResNet48 anti-spoofing subnetwork, where min t-DCF metric was calculated similarly to [8].

2.4. Loss function

To train the models for the SASV-2022 Challenge we used an Additive Margin Softmax (AM-Softmax) loss [10]. It has proved to be effective in the face recognition task and has been

successfully applied to a speaker recognition task as well. This loss function reduces an interclass variance with the help of margin penalty, which is applied to the target class logit. We used the scale parameter of AM-Softmax equal to 40, and the maximum value of margin was set to 0.3 according to [7].

3. Experiments

3.1. Datasets

For speaker recognition (ASV) systems training the VoxCeleb2-dev (5994 speakers) dataset [11] was used, and a training subset of ASVspoof 2019 Logical Access (LA) [3] dataset was used for voice anti-spoofing (CM) systems training. To evaluate the systems’ performance the eval subset of ASVspoof 2019 LA dataset was used. Besides, a dev subset of ASVspoof 2019 LA was used for development purposes, such as best training epoch weights selection for anti-spoofing models and an optimization of linear fusion weights for combined SASV system.

3.2. Data augmentation

To augment the verification training data we used the MUSAN corpus [12] and a real room impulse responses (RIRs) database [13]. We applied various on-the-fly augmentations during the training process. For each training utterance we utilized 6 different augmentation strategies:

- **Music:** A single music file was randomly selected from MUSAN and added to the original audio (5-15dB SNR). The duration of additive noise was matched to the duration of the original signal.
- **Noise:** Randomly selected noise from MUSAN was added to the original recording (0-15dB SNR).
- **Speech:** Three to seven speakers were randomly picked, summed together, and then added to the original signal (13-20dB SNR).
- **Reverb:** Artificially reverberated a signal via convolution with real RIRs.

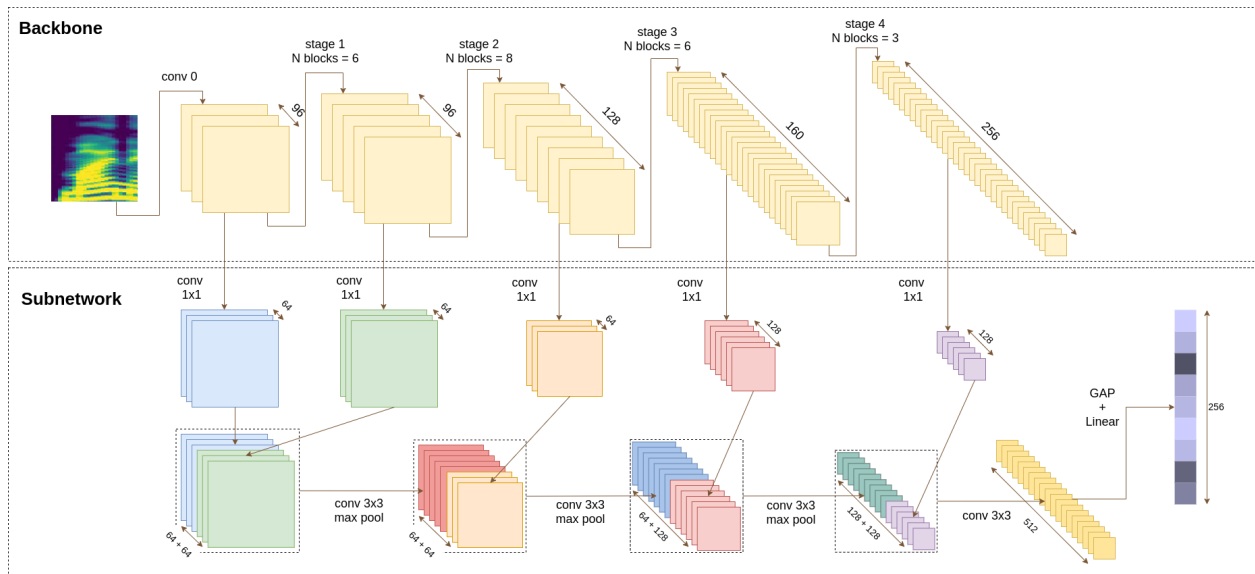


Figure 1: Scheme of subnetwork for ResNet48 backbone

- **Speed:** We applied speed augmentations (via FFT resampling), where pitch of each file has been increased or decreased by a 10%. As a result, we increased the number of speakers by a factor of 3.
- **Spectral augmentation:** We also applied SpecAugment[14] to the input log Mel-spectrograms and randomly masked 0 to 5 frames in the time domain and 0 to 10 frequency bins.

3.3. Implementation details

All the described models were trained using TensorFlow 2 framework [15] and SGD optimizer with momentum (set to 0.9). To train verification backbones faster, we used Google Cloud TPUs, while anti-spoofing subnetwork was trained on a single NVIDIA RTX2080Ti GPU. We used the following two-stages scheme to train the models.

3.3.1. Verification backbone

We trained speaker recognition models for 50 epochs, where each epoch consists of 5000 steps, with a batch size of 256. To form a batch, we sampled 256 unique speakers and took a single utterance for each of them. We randomly cropped a 2-second audio segment from each utterance in the training batch. During the training we updated the learning rate and a margin of AM-Softmax loss function. Learning rate was linearly increased from minimum ($1e-4$) to maximum value (0.1), while margin was kept zero for the first 3 epochs. We then fixed learning rate at the maximum value and linearly increased the value of margin from zero to its maximum value (0.3) for the following 10 epochs. For the rest of training, we fixed the margin of AM-Softmax loss and applied an exponential decay to the learning rate every 4 epochs with a rate of 0.5. We also applied L2-norm regularization of $1e-5$ for all the model's weights except the AM-Softmax head, for which we increased the regularization value to $1e-4$.

3.3.2. Anti-spoofing subnetwork

In the second stage, we froze the verification backbone and trained an anti-spoofing subnetwork on top of its features using the AM-Softmax loss function. We trained the subnetwork for 50 epochs, 2000 steps each, with a batch size of 32. To form a batch, we randomly sampled 16 speakers and took a pair of utterances (bonafide and spoof) for each of them. The size of input was extended to 4 seconds while training, and no augmentations were applied to the training data. Cosine decay learning rate scheduling [16] was used with the following parameters: the initial learning rate is 0.1, the minimum learning rate is $1e-4$ and the length of decaying cycle is 5 epochs. Scheduling of AM-Softmax margin is similar to the backbone training. For the first two epochs it equals to zero, then it was linearly increased during the next three epochs, and finally it was fixed to its maximum value of 0.3 for the rest of training. We also applied L2-norm regularization of $1e-3$ for all the model's weights in order to prevent an overfitting.

3.4. Fusion description

The output of our integrated SASV system includes fusion of cosine similarity scoring of backbone and anti-spoofing subnetwork embeddings, and an anti-spoofing subnetwork spoofing probability output score as follows:

1. ASV cosine similarity score between mean enrollment model backbone embedding and a verification file backbone embedding
2. CM cosine similarity score between mean enrollment model anti-spoofing subnetwork embedding and a verification file anti-spoofing subnetwork embedding
3. CM spoof probability of a verification file from the 2-class head of anti-spoofing subnetwork
4. Same as 1 with additionally applied ASNorm backend

ASNorm cohort included 1200 random files from ASVSpooof 2019 LA train set with a $top N = 300$ trials used to estimate mean and std of scores distribution for normalization.

Table 4: SV, SPF and SASV protocols EER (%) for the SASV 2022 development and evaluation partitions
E - enrollment utterances, *V* - verification utterance

Group	Name	Description	SV-EER [%]		SPF-EER [%]		SASV-EER [%]	
			Dev	Eval	Dev	Eval	Dev	Eval
Challenge Baseline	ECAPA-asv	ECAPA-TDNN ASV score [17]	1.88	1.63	20.30	30.75	17.38	23.83
	AASIST-cm	AASIST CM score [8]	46.02	49.24	0.07	0.67	15.85	24.37
	Baseline2	Ensemble of ECAPA-TDNN and ASIST	12.87	11.48	0.13	0.78	4.85	6.37
ASV	r48-asv	ResNet48 ASV cosine score (<i>E</i> vs <i>V</i>)	0.000	0.151	15.230	24.417	12.263	18.123
	r100-asv	ResNet100 ASV cosine score (<i>E</i> vs <i>V</i>)	0.051	0.111	14.676	22.942	11.921	17.277
CM	r48-cm	ResNet48 CM cosine score (<i>E</i> vs <i>V</i>)	49.265	48.394	0.067	1.400	15.556	24.224
	r48-cm-clc	ResNet48 CM classification score (<i>V</i>)	36.124	50.045	0.135	0.520	13.274	25.381
Single Model System	SF1	r48-asv + r48-cm	0.205	0.522	0.146	0.916	0.199	0.743
	SF2	+ r48-cm-clc	0.068	0.377	0.067	0.431	0.068	0.406
	SF3	++ ResNet48 ASV ASNorm score (<i>E</i> vs <i>V</i>)	0.077	0.278	0.076	0.433	0.076	0.339
	SF4	+++ QMF: ResNet48 CM class. score (<i>E</i>)	0.068	0.238	0.067	0.279	0.068	0.260
	SF5	++++ QMFs: speech lengths (<i>E</i> and <i>V</i>)	0.068	0.186	0.067	0.245	0.068	0.223
Models Ensemble	F1	SF2 + r100-asv	0.068	0.279	0.072	0.448	0.068	0.354
	F2	+ AASIST-cm	0.128	0.261	0.007	0.226	0.052	0.242
	F3	++ ASNorm for ResNet48 and ResNet100	0.137	0.258	0.016	0.224	0.062	0.242
	F4	+++ QMF: ResNet48 CM class. score (<i>E</i>)	0.009	0.150	0.007	0.172	0.007	0.153
	Submission	++++ QMFs: speech lengths (<i>E</i> and <i>V</i>)	0.000	0.105	0.004	0.168	0.004	0.136

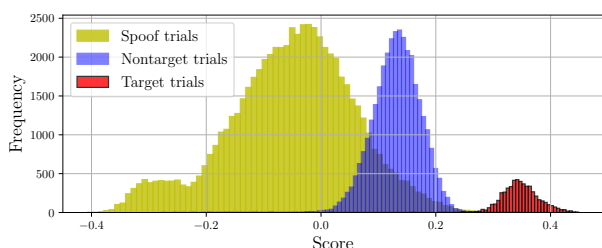


Figure 2: Histogram of Fusion scores

3.4.1. Quality Measurement Functions

To further improve the target metrics, QMF [9] correcting terms were used in addition to ASV and CM scores to shift each trial. The following factors were used in a final submission:

- **Enrollment model speech length** - sum of speech lengths across all enrollment files in a model:
 $\Delta = -\sum_{i=1}^{|E|} L(E_i)/200$, where $L(x)$ - speech length in seconds, E - enrollment model, E_i - enrollment utterance. File speech length was extracted using a simple energy-based VAD from Kaldi toolkit [18].
- **Verification file speech length:** $\Delta = -L(V)/10$, where $L(V)$ - speech length of a verification utterance.
- **Enrollment model inverted CM score** - mean value of inverted sign CM system output for all enrollment files in a model: $\Delta = -\sum_{i=1}^{|E|} CM(E_i)/|E|$, where $CM(x)$ - CM spoof probability. This factor is considered as a feature describing the mean quality of enrollment model.

3.4.2. Fusion scheme

The SASV system output is a linear fusion of ASV and CM scores and QMF values for enrollment and verification files. The optimal weights are estimated using COBYLA toolkit [19] minimizing the EER metric on SASV development set.

3.5. Evaluation

Evaluation of systems' performance is done using Equal Error Rate (EER) metric, corresponding to the operating point of equal False Acceptance and False Rejection error rates.

4. Results

The testing results on SASV dev and eval data are presented in the Table 4. It reflects our ASV backbones and CM anti-spoofing subnetworks quality for ResNet48 and ResNet100 models. From this table we can see how our single-model system SASV-EER was improved by using a fusion of ASV and CM embedding-based scores (SF1). Moreover, extending such fusion with CM spoof probability scores and various QMFs (SF5) leads us to a significant x3 SASV-EER reduction on eval subset. Finally, the SF5 system consists of ResNet48 model only with QMFs in fusion and reaches **0.223%** SASV-EER on eval set, while being relatively compact and fast.

In a similar to the single-model system fashion we have built an ensemble system containing both ResNet48 and ResNet100 models together with an open-sourced AASIST [8] CM scores. By exploiting the previously showed fusion improvement strategy with various QMFs we were able to achieve the final metric of **0.136%** SASV-EER on a challenge evaluation set with our ensemble submission. Histogram of scores for our final system is presented on the Figure 2.

5. Conclusions

In our paper we presented a subnetwork approach that combines verification and anti-spoofing models in a single computational graph and shows good results for both SV and SPF protocols independently. Additionally, we proposed a novel scoring strategy for SASV protocols, which includes the usage of embedding-based similarity scores and anti-spoofing classification head output spoof probability. Furthermore, we have found out that the usage of QMF factors could be very profitable, especially if applied to both enrollment and verification utterances.

6. References

- [1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth annual conference of the international speech communication association*, 2015.
- [2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.
- [3] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.
- [4] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [6] J. Thienpondt, B. Desplanques, and K. Demuynck, "Tackling the score shift in cross-lingual speaker verification by exploiting language information," *arXiv preprint arXiv:2110.09150*, 2021.
- [7] D. Garcia-Romero, G. Sell, and A. Mccree, "MagNetO: X-vector Magnitude Estimation Network plus Offset for Improved Speaker Recognition," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 1–8. [Online]. Available: <http://dx.doi.org/10.21437/Odyssey.2020-1>
- [8] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," *arXiv preprint arXiv:2110.01200*, 2021.
- [9] A. Alenin, A. Okhotnikov, R. Makarov, N. Torgashov, I. Shigabeev, and K. Simonchik, "The id r&d system description for short-duration speaker verification challenge 2021," in *Interspeech 2021*, 2021, pp. 2297–2301.
- [10] 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.
- [11] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.
- [12] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [13] I. Szoke, M. Skacel, L. Mosner, J. Paliesek, and J. Cernocky, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, p. 863–876, Aug 2019. [Online]. Available: <http://dx.doi.org/10.1109/JSTSP.2019.2917582>
- [14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, Sep 2019. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-2680>
- [15] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.
- [16] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.
- [17] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapadnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.
- [18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldı speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [19] M. J. Powell, "A view of algorithms for optimization without derivatives," *Mathematics Today-Bulletin of the Institute of Mathematics and its Applications*, vol. 43, no. 5, pp. 170–174, 2007.