SFC: Backpropagation Neural Network - Classification

$JánProfant^1$

[1]BUT FIT, Brno

xprofa00@stud.fit.vutbr.cz

1. Introduction

In this paper, we propose classification approach to text dependent speaker recognition using Backpropagation neural network [1]. We used popular technique used in speaker recognition field in the last years - i-vectors [2]. We trained feed forward backpropagation neural network using stochastic gradient descent [3] to classify closed set of speakers using two text dependent speaker recognition datasets - small and large. This approach can be used for example in bank institutions - voice as password, restricted access to building or other voice biometric cases.

1.1. System Description

We used Mel Frequency Cepstral Coefficients (MFCCs) [4] together with Stacked bottlenecks [5] as input features for i-vector extraction. Universal Background model [6] and i-vector extractor itself were trained on subset of the PRISM set [7]. System scheme is shown in Figure 1.

We used 250 dimensional i-vector as input with one hidden layer. Neural network was trained to classify input i-vector - assign it one speaker from closed set of speakers.

2. Datasets

For our experiments we used two datasets - small and large. Small dataset was recorded during Hackhathon at Phonexia s.r.o. [8] - it consist from 158 recordings from 27 speakers. Large dataset consists from 1794 recordings from 200 speakers and is part of RSR2015 corpus [9]. RSR2015 is corpus developed for text dependent speaker recognition with multiple phrases recorded on various devices - from our previous experiments we chose one of the best performing phrases. Metadata about datasets are summarized in Table 1.

We split our datasets into train and test, where test set contained always 2 recordings from one speaker.

3. Experiments

We ran multiple experiments with different size of hidden layer. Furthermore, we normalized all input i-vector using l^2 normalization [10]. Experiments for both of our datasets are shown in Table 2.

From our experiments, we can clearly see that l^2 normalization helps neural network to better classify speakers. We tried to run experiments also with similar size of training and test set and trend was similar. In case of small dataset, accuracy reached 100% and also for large dataset the best result in terms of accuracy was 99.75%. We can compare our results to [11] where PLDA model was used achieving Equal Error Rate (EER) [12] around 1.5% - it is important to note, that it was verification task, not classification.

3.1. Application Documentation

Application used in this paper was written in C++ and is fully compatible with *merlin.fit.vutbr.cz* server setup.

3.1.1. Compilation

```
#!/bin/bash
make
```

3.1.2. Data

Data for small set are part of attached *zip* file stored in directories

- **data/** - small dataset stored in structure expected by application
- **data_l2-norm/** - l^2 normalized small dataset stored in structure expected by application

3.1.3. Application

Application can be run with parameters

```
#!/bin/bash
./main
-l input_list
-d input_directory
-i ivector_size
-t num_test_ivectors
-e eps
-h hidden_layer_neurons
```

where

- *input_list* - path to input list (string)
- *input_directory* - path to input directory (string)
- *ivector_size* - size of input i-vector (int)
- *num_test_ivectors* - number of test i-vectors per speaker (int)
- *eps* - learning rate (float)
- *hidden_layer_neurons* - number of hidden layer neurons (int)

or just by running attached script

```
#!/bin/bash
./run.sh
```

4. Conclusions

We proposed feed forward neural network architecture for classification of input i-vectors to their speakers. This paper does not consider case, when input i-vector does not come from speaker out of classification set - this extension requires further experiments, for example in setting analytical threshold

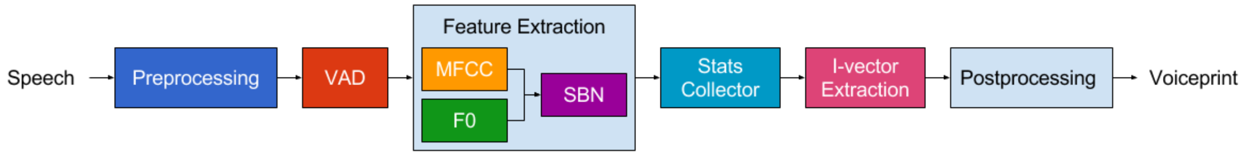

Figure 1: *General block scheme of i-vector extractor.*

Table 1: *Datasets with specified phrases used for text dependent speaker verification.*

Dataset	Recordings	Speakers	Average rec/spk	Min rec/spk	Max rec/spk	Phrase
Small	158	27	5.85	5	12	Abracadabra, open Sesame. I forgot my keys.
Large	1794	200	8.97	8	9	My dress needs some work on it.

Table 2: *Accuracy on test data for small dataset and large dataset.*

Dataset	HL Neurons	Accuracy	Epochs
Small	40	83.33%	13000
Small	70	90.74%	40000
Small	100	**96.30%**	80000
Small ($l^2 norm$)	100	94.44%	8000
Small ($l^2 norm$)	70	96.30%	5000
Small ($l^2 norm$)	90	**100%**	4000
Large	100	**95.00%**	10000
Large	150	94.00%	9000
Large	200	85.25%	6000
Large ($l^2 norm$)	20	95.25%	15000
Large ($l^2 norm$)	30	98.5%	30000
Large ($l^2 norm$)	70	**99.75%**	14000

when using soft max layer. These datasets can be considered relatively easy - similar recording conditions, no background noise and small number of speakers. Despite this fact, experiments shows surprisingly good results - **100%** accuracy for small dataset and **99.75%** for large dataset. Our experiments shows that l^2 normalization increased accuracy. Also, training on larger number of speakers does not result in significant decrease in terms of accuracy and needs further experiments on large datasets with more than 1000 speakers, so it can be sufficient for real environment.

5. References

[1] R. Hecht-Nielsen *et al.*, "Theory of the backpropagation neural network." *Neural Networks*, vol. 1, no. Supplement-1, pp. 445–448, 1988.

[2] O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 4516–4519.

[3] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010.* Springer, 2010, pp. 177–186.

[4] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[5] S. Yaman, J. Pelecanos, and R. Sarikaya, "Bottleneck features for speaker recognition," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[7] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the prism evaluation set," in *Proceedings of NIST 2011 workshop.* Citeseer, 2011.

[8] Phonexia, "Phonexia s.r.o." https://www.phonexia.com/en/.

[9] Exploit Technologies Pte Ltd, "RSR2015 Overview & Specifications," https://www.etpl.sg/innovation-offerings/ready-to-sign-licenses/rsr2015-overview-n-specifications.

[10] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, vol. 2011, 2011, pp. 249–252.

[11] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using plda with uncertainty propagation," *matrix*, vol. 500, p. 1, 2013.

[12] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Second international conference on spoken language processing*, 1992.