A DENOISING AUTOENCODER FOR SPEAKER RECOGNITION. RESULTS ON THE MCE 2018 CHALLENGE

Roberto Font
roberto.font@biometricvox.com

Introduction

Most state-of-the-art speaker recognition systems consist on a frontend that extracts a fixed and low dimensional representation of speech segments – like ivectors or, more recently, xvectors-, and a discriminative backend. This backend usually consists of Linear Discriminant Analisys (LDA) followed by Probabilistic Linear Discriminant Analysis (PLDA) to compute scores.
As an alternative to this approach, **we propose the use of a Denoising Autoencoder (DAE) which takes as input an ivector and tries to map it to the mean of all the ivectors of that particular speaker**. To this end, the DAE is trained to maximize the cosine distance between its output and the mean ivector for that speaker. Our proposed backend consists of: length normalization, DAE transformation, PLDA scoring and S-Norm for score normalization.

MCE 2018 challenge

The MCE 2018 1st Multi-target speaker detection and identification Challenge Evaluation presents a call-center fraud detection scenario: given a speech segment, detect if it belongs to any of the speakers in a blacklist. The data was generated from real call center user-agent telephone conversation and distributed to participants in the form of ivectors. The table below shows the distribution of each data partition.

	# speakers	#utterances
Training blacklist	3,631	10,893
Training background	5,000	30,952
Development blacklist	3,631	3,631
Development background	5,000	5,000
Evaluation	?	16,017

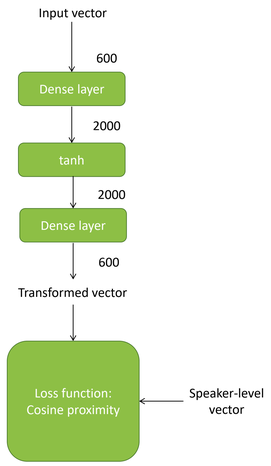
The challenge is divided into two related subtasks:

- Top-S detection: detect if the segment belongs to any of the blacklist speakers
- Top-1 detection: detect which specific blacklist speaker (if any) is speaking in the segment.

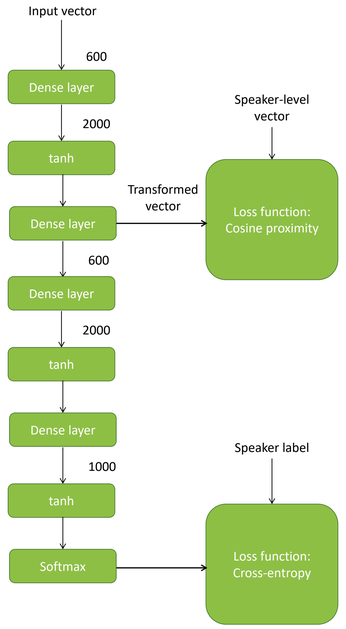
Proposed approach

The proposed system consists of a DAE that has as input an ivector and as output a vector with the same dimension. During training, the target is the mean of all ivectors from that speaker, and we try to minimize the cosine proximity between the output and the target.
We also test a possible extension which extends the DAE to try to discriminate between speakers. We add two additional hidden layers and a new output layer that makes a prediction about the speaker identity. During training, the loss is a linear combination of cosine proximity for the intermediate output and cross-entropy for the final output.

Simple DAE:



Discriminative DAE:



We train both neural networks with Adam optimizer, a learning rate of 0,001 and a batch size of 128 for 5 epochs using only the background training set. We used Keras with Tensorflow backend.

Results

The results on the development are shown in table below. As we can see, score normalization is extremely beneficial, particularly for the DAE system. We can also see that the discriminative DAE does not improve the results over the simple DAE.

Results on development set:

System	Top-S EER [%]	Top-1 EER [%]
Baseline	2,01	12,26
LDA + PLDA	1,82	6,96
LDA + PLDA + SNorm	1,26	6,72
DAE + PLDA	1,73	7,22
DAE + PLDA + Snorm	**1,25**	**6,52**
Disc. DAE + PLDA + SNorm	1,38	6,80

Results are shown in the table below. Again, **the DAE system achieves lower error rates than the LDA-PLDA system. It is worth noting that the system exhibits no overfitting.**

Results on evaluation set:

System	Top-S EER [%]	Top-1 EER [%]
Baseline	6,24	11,24
LDA + PLDA	4,63	6,81
LDA + PLDA + SNorm	4,42	6,56
DAE + PLDA	4,60	6,75
DAE + PLDA + Snorm	**4,33**	**6,11**

Code

A Python implementation of the proposed approach can be found at:

http://github.com/BiometricVox/DAE_SpeakerID

BIOMETRIC VOX