THE BIOMETRIC VOX SYSTEM DESCRIPTION FOR THE NIST SRE20 CTS CHALLENGE

Roberto Font

Biometric Vox S.L.

roberto.font@biometricvox.com

ABSTRACT

This paper provides a detailed description of the Biometric Vox submission to the NIST SRE 20 CTS Challenge. Our system is based on Deep Neural Network x–vector embeddings and a PLDA backend. We obtained $EER = 3.40\%$, $C_{min} = 0.140$, $C_{act} = 0.207$ on the SRE 20 CTS Challenge set.

1. DATA RESOURCES

1.1. Training data

Our training material consists of the following datasets:

- NIST SRE 04-10
- MIXER6 as prepared by the Kadi sre16 recipe.
- Switchboard phase1-3 and cellular1-2.
- NIST SRE 18 Evaluation set.
- NIST SRE 19 CTS Challenge test set.
- Voxceleb 1 and 2.

We augment training data by generating four additional perturbed versions of each file by adding:

- Reverberation
- Musan noise.
- Musan music.
- Musan speech.

For embedding extractor training, we consider a training set consisting on: NIST SRE 04-10, MIXER6, Switchboard, NIST SRE 18 Evaluation set, VoxcelebCat and all augmented data from these datasets. VoxcelebCat is the result of concatenating all excerpts from the same video into one longer file and combining Voxceleb 1 train, Voxceleb 2 dev and Voxceleb 2 test.

The total number of utterances is $1.220.872$ from a total of 13.015 speakers.

1.2. Development data

In addition to the NIST SRE 2020 CTS Challenge leaderboard, we used the NIST SRE 18 Development set as our internal development set.

2. SYSTEM DESCRIPTION

2.1. Pre-processing and feature extraction

The input of this system are 23-dimensional MFCCs which are extracted from 25 ms windows with 15 ms overlap. Features are normalized using cepstral mean and variance normalization over a sliding-window of 300 frames.

To compute VAD labels, we use the class-posteriors of an ASR TDNN trained on Fisher English. This TDNN has a total of 114 output classes (senones) modelling either silence (5 classes), noise (2 classes) or speech. The posterior probabilities obtained for each frame are used to map the frame to the speech, silence or noise categories. Frames classified as either silence or noise are dropped.

Having the ASR system to explicitly consider noise classes allows us to filter-out noise, which can not be done with energy-based techniques. In previous work, we have found that this can be very beneficial in noisy conditions.

2.2. Embedding extraction

The system uses a Time Delay Neural Network (TDNN) architecture that is well-known in the Speaker Recognition literature [1, 2]. This architecture is shown in Table 1. Note that it differs in some implementation details, in line with [3], with respect to the original implementation in [2].

Implementation was done using TensorFlow [4]. Training was performed using Stochastic Gradient Descent with an exponentially decaying learning rate with initial value of 0.005. Training stopped when validation loss did not improved for 10 epochs.

2.3. Back-end

The back-end follows a classical LDA-PLDA scoring scheme:

- Embeddings are projected to unit length, centered and whitened.

Table 1. TDNN architecture.

Layer Type	Filter/Stride	Output	Params
Conv1D-batchnorm-ReLU	$5 \times 1/1 \times 1$	$T \times 512$	61.4K
Dense-batchnorm-ReLU	-	$T \times 512$	264.7K
Conv1D-batchnorm-ReLU	$5 \times 1/1 \times 1$	$T \times 512$	1.313M
Dense-batchnorm-ReLU	-	$T \times 512$	264.7K
Conv1D-batchnorm-ReLU	$7 \times 1/1 \times 1$	$T \times 512$	1.837M
Dense-batchnorm-ReLU	-	$T \times 512$	264.7K
Conv1D-batchnorm-ReLU	$7 \times 1/1 \times 1$	$T \times 512$	2.361M
Dense-batchnorm-ReLU	-	$T \times 512$	264.7K
Dense-batchnorm-ReLU	-	$T \times 512$	264.7K
Dense-batchnorm-ReLU	-	$T \times 1500$	775.5K
Stats Pooling (mean+stddev)	-	3000	-
Dense-batchnorm-ReLU	-	512	1.538M
Dense-batchnorm-ReLU	-	512	264.7K
Additive Margin Softmax	-	14413	7.379M
Total			16.85M

- LDA is used to project the embeddings to a lower dimension. (From 512 to 150 in our case.)

- The trial is scored using PLDA.

Both LDA and PLDA are trained on NIST SRE 04-10 + MIXER6 + NIST SRE 18 Eval + NIST SRE 19 CTS + all augmented data (a total of 455.358 utterances from 4.474 speakers).

Finally, we used symmetric normalization (S-Norm) using NIST SRE 04-08 and NIST SRE 18 Evaluation set as the cohort. Only the top 1.000 scores were used to compute mean and standard deviation.

3. RESULTS

Our system obtains $EER = 3.40\%$, $C_{min} = 0.140$, $C_{act} = 0.207$ on the SRE 20 CTS Challenge set.

4. COMPUTATIONAL RESOURCES

The scoring of a single trial involves:

- Feature extraction and VAD for both enrollment and test segments.

- Embedding extraction and post-processing for both segments.

- PLDA scoring.

- Score calibration.

This whole process takes an average of 4.31s with a peak consumption of 250MB RAM. These results were obtained by running 10 repetitions of a single trial scoring using a single thread on a c5xlarge AWS instance.

5. REFERENCES

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, “Speaker recognition for multi-speaker conversations using x-vectors,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5796–5800.

[3] Yi Liu, Liang He, and Jia Liu, “Large Margin Softmax Loss for Speaker Verification,” in *Proc. Interspeech 2019*, 2019, pp. 2873–2877.

[4] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, Software available from tensorflow.org.