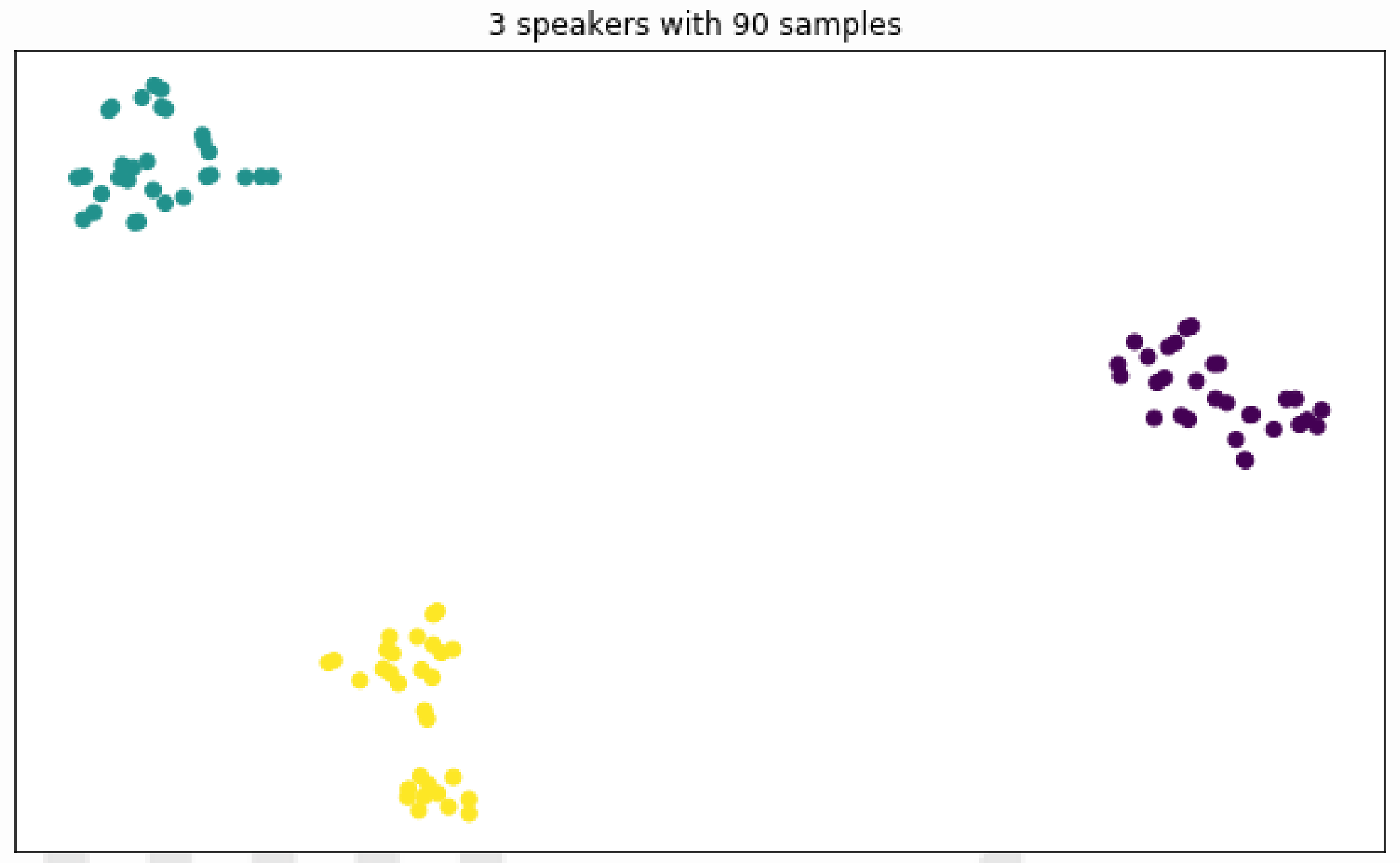
Towards Forensic Speaker Identification in Spanish using Triplet Loss

Emmanuel Maqueda, Javier Álvarez Jiménez and Ivan Vladimir Meza Ruiz*.
(emmaqueda@comunidad.unam.mx, javieralvarezim@nube.unadmexico.mx, ivanvladimir@turing.iimas.unam.mx)
Applied Mathematics and Systems Research Institute, National Autonomous University of Mexico

Abstract

This work uses of a triplet loss deep network setting for the forensic identification of speakers in Spanish. We train a convolutional network to produce vector representations of speech spectrogram slices. Then we test how similar these vectors are for a given speaker and how dissimilar are compared with other speakers.



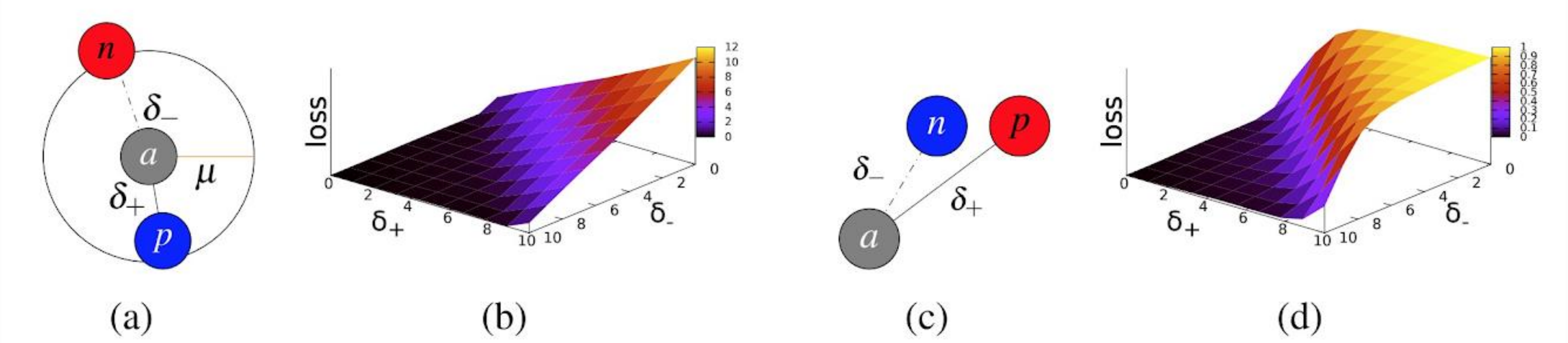
3 speakers with 90 samples

Background - Forensic Linguistics

Forensic speaker identification focuses on gathering and quantifying the evidence for the identification of a person through their voice. However, it is not only a case of matching two recordings by their similarity but in the case of forensic analysis it also necessary to quantify the chances of the recordings to be confused within recordings of speakers of the population (tipicality).

Triplet Loss

Triplet loss evaluates three vector representations of two objects (originally a picture, in our case a slide of audio). The first and second representations correspond to the same object identity, while the third representation corresponds to a second object identity. The goal of triplet loss is to enforce that the two first representations are close in relation with the third one.



(a) (b) (c) (d)

Image taken from: https://atcold.github.io/pytorch-Deep-Learning/images/week11/11-2/tml.png

Dataset, Arquitecture & Methodology

The input of our CNN is a slice of a spectrogram composed by 60ms and frequencies information up to 8.5kHz. This setting creates a patch of 200 × 256 pixels.

This is feed into a five convolutional layer with 32 kernels (2 first layers) and 64 kernels (3 last layers); each convolutional layer is followed by a batch normalization layer, max polling (size 2) and a ReLU activation function. The output of the CNN network is a 1D − 1024 dimension vector which represents the speech audio slice.

Speakers	2,180
Recordings	21,692
Avg. duration	8.25

Results

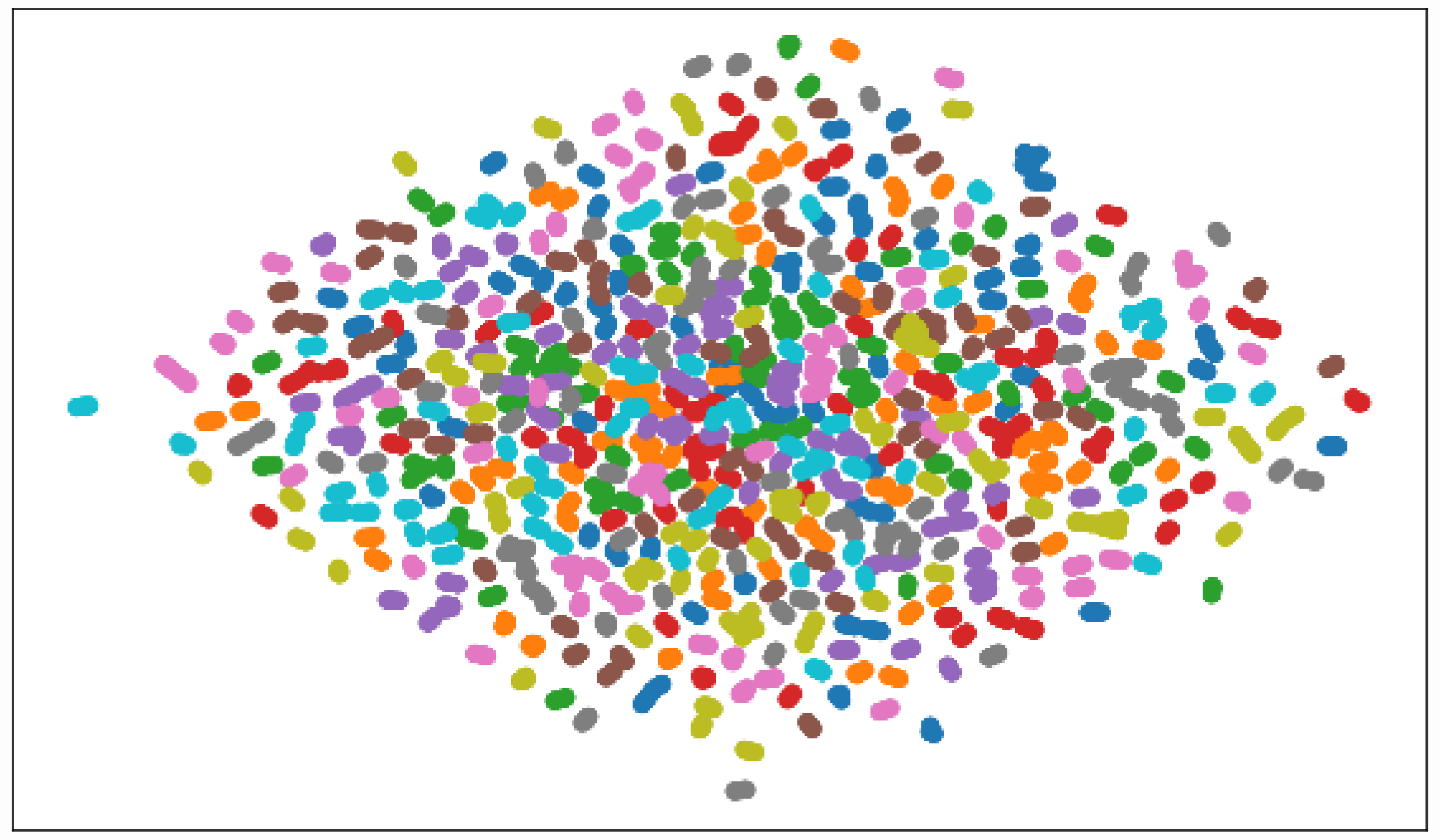
In the case of the loss function we enforced three different margins of 0.2, 0.5 and 0.8, this means that a sample from two different speakers should be at least separated by the set margin.

We calculate three metrics: inner average distance for speaker samples (IAD), outer average distance between speaker and sample centroids of other speakers (OAD). With these two metrics we calculate a simile to Likelihood Ratio (LR). We also calculate the mean Silhouette Coefficient (MSC) which ranges from −1 (worst result) to 1 (best result).

Margin	IAD	OAD	LR	MSC
0.2	**0.449**	3.16	0.142	0.248
0.5	0.891	6.221	0.141	0.225
0.8	2.04	**12.32**	**0.166**	**0.256**

Conclusions & Future Work

We have established that a larger margin for the the loss gives the better results in terms of how close the samples from a speaker are, and how far are these to other speakers. However, our experiments show there is room for improvement since it overlap among speakers can be noticed. Future work will focus on introducing new methods to train the triplet loss.



218 speakers with 60.0 samples (silhuoete score 0.259)

References: Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N. (2016). Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In Proceedings of the iEEE conference on computer vision and pattern recognition (pp. 1335-1344), Hernández-Mena, C. VoxForge Spanish Corpus. In Personal collection, Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.