Towards forensic speaker identification in Spanish using triplet loss

Emmanuel Maqueda
Facultad de Estudios Superiores Cuatitlan
Universidad Nacional Autónoma de México
emmaqueda@comunidad.unam.mx

Javier Alvarez Jiménez
Universidad Abierta y a Distancia de México
javier.alvarezjim@nube.unadmexico.mx

Ivan Vladimir Meza Ruiz
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
ivanvladimir@turing.iimas.unam.mx

Abstract

This work explores the use of a triplet loss deep network setting for the forensic identification of speakers in Spanish. Within the framework, 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. Based on these metrics we propose the calculation of the Likelihood Radio which is a cornerstone for forensic identification.

1 Introduction

Triplet loss [1] was introduced as a method to train a CNN to produce good vector representation for the face identification task. 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 identity. The goal of triplet loss is to enforce that the two first representations are close in relation with the third one.

On the other hand, 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*). In this work we propose to measure the inter-speaker and outer-speaker distances as a proxy of quantifying similarity and tipicality, and in this way to get a Likelihood Radio.

2 Methodology

The input of our CNN is a slice of a spectogram composed by $2s$ and frequencies information up to $8.5kHz$. This setting creates a patch of 200×256 pixeles, 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. 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.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

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**

Table 1: Variations of metrics by augmenting the margin.

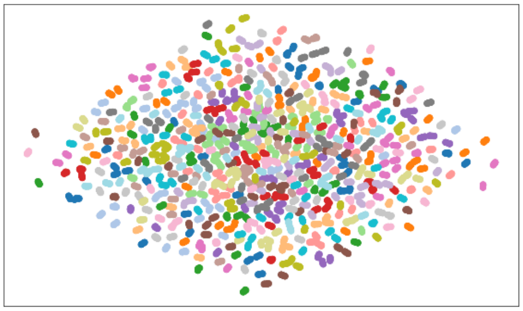

Figure 1: $2D$ projection of vector representations of speech samples (same color same speaker, margin 0.8, validation set, 218 speakers)

3 Dataset and Results

In this work we use the Spanish Voxforge dataset [2]. This corpus is composed by $2,180$ Spanish speakers, $21,692$ recordings which in average last $8.25 secs$. We split the speakers in 80% tranning, 10% validation and 10% testing.

In order to measure the performance of the network, we calculate three metrics: inner average distance for speaker samples (IAD), outer average distance between speaker and centroids of other speakers (OAD). With these two metrics we propose to 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). Table 1 shows the main scores for the different margins.

As it can be seen, the lower the margin, the larger the difference between the inner and outer distance. Even though for a small margin the samples are the closest (IAD), for larger margin these are further away (OA). In particular, our proposal for LR benefits from larger margins. On the other hand, we can observe that this behaviour is confirmed by MSC which is larger for a higher margin. Figure 1 shows a $2D$ projection of the validation samples, we can observe that samples from the same speaker are clustered together, but there is space where the clusters touch or overlap.

4 Conclusions

In this work we explored the use of a triplet loss network setting for the forensic identification of speakers. We have established that a larger margin for 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.

5 Bibliography

[1] 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).

[2] Hernández-Mena, C. VoxForge Spanish Corpus. In *Personal collection.*

[3] 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.