Triplet loss-based embeddings for forensic speaker identification in Spanish

Emmanuel Maqueda1 · Javier Alvarez-Jimenez2 · Carlos Mena3 · Ivan Meza4

Received: 1 February 2021 / Accepted: 17 August 2021 / Published online: 4 September 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract
With the advent of digital technology, it is more common that committed crimes or legal disputes involve some form of speech recording where the identity of a speaker is questioned [28]. In face of this situation, the field of forensic speaker identification has been looking to shed light on the problem by quantifying how much a speech recording belongs to a particular person in relation to a population. In this work, we explore the use of speech embeddings obtained by training a CNN using the triplet loss. In particular, we focus on the Spanish language, which has not been extensively studied. We propose extracting the embeddings from speech spectrograms samples, then explore several configurations of such spectrograms, and finally, quantify the embeddings quality. We also show some limitations of our data setting which is predominantly composed by male speakers. At the end, we propose two approaches to calculate the likelihood ratio given out speech embeddings and we show that triplet loss is a good alternative to create speech embeddings for forensic speaker identification.

Keywords Forensic speaker identification · Triplet loss · Likelihood ratio · Embedding space

1 Introduction

The use of Triplet loss [35] was popularized with the introduction of the FaceNet architecture which was aimed to face identification tasks. This loss allows to train a neural network, commonly a convolutional neural network (CNN), to produce a vector representation of an image. The goal is that the neural network learns a mapping from face images to an Euclidean space; it is desirable that distances among face image positions directly correspond to a measure of face similarity. Within this setting, a desired outcome is that images of the same face will cluster together and images of different faces will be separated by a margin. Triplet loss has been applied into different types of images: objects [38], person re-identification [6], information retrieval [15].

Recently, Triplet loss has been proposed for speech tasks, for instance: the speaker verification task [41], for speaker turn [4], speech emotion classification [16], among other tasks. However, its applicability to forensic speaker identification has not been explored, particularly for the Spanish language. Forensic Speaker Identification (FSI) focuses on gathering and quantifying the evidence that will be presented in a court. FSI addresses the question if a specific recording registers or not, speech produced by a specific person [26, 31]. The more basic scenario in FSI consists of two sample speech recordings, a reference sample and a questioned sample. For the reference sample, we always know the identity of the speaker. This certainty
is guaranteed by the chain of custody, and this is because we know the conditions in which the recording was taken, including the speaker’s identity. On the other hand, for the questioned recording we are not sure about the identity of the person whose voice is in the recording. In a case that involves FSI, the identity of the speaker in the questioned recording is contested regarding the identity in the reference recording; one of the involved parts affirms that the voice in the reference and the questioned recordings are the same (same-speaker hypothesis), while the other part affirms the contrary (different-speaker hypothesis).

The goal in a FSI case is not only about matching the two recordings by their similarity, same-speaker hypothesis, FSI requires a stronger legal standard which makes also necessary to quantify the chances of the questioned sample to be associated to other speakers of the population, this addresses the different-speaker hypothesis. This measurement is known as typicality. With these two measurements, similarity and typicality, it is common to calculate the likelihood ratio (LR) [13]. LR offers a quantifiable measurement that updates the odds of one of the hypothesis; this information should be taken into consideration in the context of the presence of other evidence regarding the case.

In this work, we propose to extract speech embeddings from speech spectrogram samples for reference and questioned speech recordings in order to quantify the LR. We start by presenting related work in section 2. We continue by presenting the details of our neural model and the implementation of the triplet loss in section 3. We present two sets of results: first, in section 5.1, we measure the quality of the embeddings by proposing inner and outer speaker distance metrics, together with the use of the well-established silhouette clustering metric. Second, in section 5.2, we propose two ways to calculate the LR in terms of the speech embedding distances. Once we show that speech embeddings are an option to be used in FSI, we discuss some ethical aspects to be considered in section 6. Finally, we summarize our main findings in section 7.

2 Related work

Forensic speech science has advanced in last three decades in which it has established different methodologies and techniques to face the speech identification problem [26]. In the case of the methodology, there has been a paradigm shift towards empirically grounded methods [23, 34]. This shift has been motivated by the requirements of admissibility of science evidence that had become a standard in some courts around the world [5, 11]. The main result of this shift has been the adoption of the likelihood ratio (LR) as a mean to introduce the evidence in court. LR is formulated in the following manner:

\[
LR = \frac{p(E|H_s)}{p(E|H_d)}
\]

where \(E\) represents the evidence, in FSI this is the quantification of speech properties in the questioned recorded sample. \(H_s\) correspond to the same-speaker hypothesis and \(H_d\) to the different-speaker hypothesis. The numerator can be considered a similarity score while the denominator a typicality one. LR is not to be considered independent of the other facts of the case; on the contrary its meaning only depends on the strength or lack of the rest of the evidence and their compatibility with either of the hypotheses.

On the other hand, from the point of the techniques there has been several proposals which allow the quantification of \(p(E|H_s)\) and \(p(E|H_d)\). One approach that was extensively explored was the statistical analysis supposing a Gaussian distribution of speech features [31]. A common approach is to measure a specific phonetic and phonological speech properties (e.g., formants) in a specific context (e.g., a word [29, 30]). Motivated by the application of multi-variable statistics in the forensic field [1], new approaches were suggested together with kernel approaches to improve the statistical analysis of the speech evidence [24, 32].

Another proposed method is the use of Gaussian-Mixture-Model/Universal Background Model (GMM – UBM) [3]. This methodology reflects more a generative machine learning approach which depends on data intensive algorithms; for this reason in these methodologies there is no necessity of measuring a specific voice property. The GMM – UBM model is parametric depends on a dataset of recordings which is used to define the model parameters, in this case the parameters of the Gaussian mixtures. A GMM speaker-specific model is created to quantify the similarity term, and a UBM general model, based on the population of possible speakers, is generated to quantify the typicality term. With the advent of different machine learning techniques, some other machine learning-based approaches had been proposed, such as using support vector machines, boosting algorithms and random forest [25, 37]. A fundamental piece to adopt ML techniques is its appropriateness to calculate the LR [12]. Recently, there has been proposal that exploits the discriminative power of Neuronal Networks and their capability of producing representations; some examples of these approaches are: DNN senone i-vectors [10], bottleneck features [39] and x-vectors [36].

On the other hand, with the progress of self-supervised methods for training deep learning networks there has been advances in proposals for learning representations for embedding spaces. Contrastive methods have been proposed to compare a questioned sample to a set of samples.
of known speakers; this setting corresponds to a setting of speaker verification were there is access to samples of the speaker to identify, it is a matter of verifying if the two sets of samples match [14, 36]. In recent years there has been progress into using siamese networks [7] for creating good embeddings and to be applied in speaker identification [18, 27]. Of particular interest are the advances with the Triplet Loss, since it maps raw speech representation into an Euclidean space [35]. The use of these embeddings for the speaker identification task is trivial since the distance among embeddings can be used to determine which speaker is close to a questioned recording. Here it is important to notice that speaker identification does not quantify typicality, but just similarity. This approach had been used in different scenarios; [19] uses a Residual CNN and GRU that transforms a spectrogram into a 512 dimension vector, it proposes to use cosine similarity to guide the triplet loss and it evaluates accuracy and error to recognize the speaker [40] presented a CNN Inception Resnet to generate a 128 dimension embedding, it focused on $L2$ norm, and it proposed validation and false accept rate to evaluate its system [22] modifies the triplet loss to make it more efficient; it also uses accuracy (top 1 and top 5) to evaluate its system. In all these three cases, the resulting performances were superior to previous approaches.

Of particular interest for our experimentation is the difference between female and male voices, since as it will be presented further down that our dataset has an imbalance among this type of speakers. According to the medical notion, the voice originates in the throat of the speaker specifically in the larynx. There is an understanding that the size of the larynx correlates with sexual characteristics which at the same time determine the sex (or pitch) of the voice as an acoustic event. In average, the male larynx is larger than the female larynx and is naturally inclined to produce a lower pitched speech. However this is not a rule since for men and women speech can overlap. In this regard, the sexual characteristics of the vocal tract, biologically determined, determine sex in speech; however, the identification phase is of one of the genders, since it is constituted by a subjective interpretation [2, 8, 20]. With this in mind our experimentation will be on perception which means on gender.

3 Triplet loss and neural model

Triplet loss compares an anchor input with two other inputs, a positive input which shares a property with the anchor, in our case it is the identity of the speaker, and with a negative input which does not share such property. The comparison is guided by the following formulation:

$$L(A, P, N) = \max(D(A, P) - D(A, N) + m, 0)$$

where $A$, $P$ and $N$ are vectors representing the anchor, the positive and negative inputs, respectively. $D$ is distance metric and $m$ is a margin. In an ideal setting it is expected that the distance between $A$ and $P$ to be less than distance between $A$ and $N$ at least by a margin $m$, if that is the case the loss is zero when the network calculating the vectors is doing a good job. However, if this is not the case, the loss will be positive and by using back propagation the weights of the model that produces the vectors from raw information will be adjusted. Figure 1 shows the relation between the CNN model and the triplet loss, it is important to notice that the CNN blocks are the same neuronal network that transforms all inputs since the weights are shared. The figure also illustrates the three cases in relation to distances among the embeddings and the margin. $L2$ norm is used as a distance metric.

A common arrangement in Triplet loss is to use the same neuronal model to produce the vector inputs from raw inputs. In this work we propose the use a CNN that will receive segments of the spectrogram of speech recordings. Figure 2 shows this arrangement. We propose a simple CNN composed by 5 layers when possible otherwise 4, this is due to the size of the input patch, since the convolution and max-pooling layers reduce the input’s dimensionality, this means that sometimes the convolution of the last CNN layer cannot be performed, however as it is shown in our experimentation, this does not affect the performance of the architecture. Figure 2 shows a diagram with the specific details of our model.

3.1 Preprocessing of speech signal

The input of our CNN is a segment of a spectrogram that represents a time slice ($t$ in milliseconds) and frequencies information up to 8.5 kHz (enough to characterize the human voice), using always 256 bins for the frequencies. All our recordings are down sampled to 16K, and the spectrogram is normalized and pre-amph. These specific choices are followed from typical pre-processing of speech signals. In order to obtain the spectrogram we use a Hann window, with variable window ($w$ in milliseconds) and hop ($h$ in milliseconds) size. The parameters $t$, $w$ and $h$ allow us to generate an image patch (spectrogram segment) with a variable width but a constant height $W \times 256$. This patch is the CNN’s raw input which will be transformed it into a 1024 dimension embedding for all our experiments.
4 Datasets

In this work we use the Spanish Voxforge dataset, which is entirely based on the recordings from the Voxforge Project\(^1\). The Voxforge Project is a non-profit initiative that aims to collect transcribed speech for use with Free and Open Source Speech Recognition Engines. We chose the Voxforge because the speakers read always the same prompt, a paragraph of *El Quijote*; this eliminates overfitting by the content of what it is said. We expect that our models focus on properties of how things are said. The speakers donated a sample of their voice by registering to the project website, filling a form with relevant information about the speaker and reading some prompts directly through the speaker’s computer microphone. Thanks to this mechanism, we can know the following about every speaker, which is relevant for forensic purposes and our experimentation:

- Username: It could be left in blank or it could be an alias
- Gender: Male/Female
- Age: \(13 - 17 = \text{Youth} / 18 - 64 = \text{Adult} / 65 - \text{Senior}\)
- Native Speaker?: Yes/No
- Dialect: Country or Region

Since the beginning of the project in 2006, several languages have been added by the community, making necessary to clarify that our corpus only contains Spanish recordings until the year 2016. The original recordings were manually segmented into utterances. Table 1 shows the total number of male and female speakers and how they are classified into the corpus according to their nationality.

The Spanish Voxforge dataset is composed by 21,692 recordings which in average last 8.25 seconds. They come in a 16 kHz, 16 bit, mono format. The total duration of the whole recordings is approximately 50 hours. For the experiments, we split the speakers in 80% training, 10% validation and 10% testing.

In addition to Spanish Voxforge dataset, we use LibriVox Spanish Corpus [21] which consists of approximately 73 hours of reading speeches and transcripts in Spanish. The audio data was taken from Spanish-language audiobooks developed by LibriVox, a non-profit project that creates audiobooks from public domain works. The audio is made up of phrases from 300 books with a duration of between three and ten seconds. The audio files were manually segmented. Audio is presented as 16 kHz, 16 bit single-channel FLAC files. When unzipped, they produce *PCM WAV* files.

We use a total of 146 speakers and as with Spanish Voxforge dataset this is because eight of the non-native speaker files were faulty; we split the speakers in 80% training, 10% validation and 10% testing. Table 2 shows the distribution of speakers segmenting them into native and non-native.

5 Experiments and results

We performed two levels of experimentation. First we explored the speech embeddings by measuring quality of the speech embeddings and the cluster they generate on samples from the same speaker. Second we evaluated the capability of the speech embeddings to be used to calculate an LR; we propose two approaches: one is to use directly distance and questioned as a proxy for the LR; second is to use a ratio between the previous distances divided by the shortest distance to other speaker in the population.

---

\(^1\) http://www.voxforge.org.
Table 1 Number of speakers in the Voxforge corpus and their nationalities

| Country      | Males | Females |
|--------------|-------|---------|
| Argentina    | 143   | 31      |
| Chile        | 69    | 3       |
| Latin America| 148   | 10      |
| Mexico       | 68    | 8       |
| Spain        | 1240  | 411     |
| Unknown      | 45    | 4       |
| Total        | 1713  | 467     |

Table 2 Number of native and non-native speakers in the LibriVox Spanish Corpus

|                  | Natives | Non-natives |
|------------------|---------|-------------|
|                  | 140     | 32          |

5.1 Quality of the embeddings

To quantify the quality of the embeddings, we calculate three metrics: inner average distance for the same speaker samples (IAD), outer average distance between speaker centroids of other speakers (OAD). We expect an (IAD) small, which will signal that samples from the same speaker will land in the same region. For OAD, we expect a large number that will signal that the samples from a different speaker will be far away from other speakers. With these two metrics, we proposed to calculate a distance ratio (DR) that will tell us the relation between a speaker and the rest of speakers, we will expect a small amount for DR signaling good speech embeddings. We also calculated the mean silhouette coefficient (MSC) which ranges from −1 (worst result) to 1 (best result). In this case, a number closer to 1 will indicate that there is less confusion among the clusters of embeddings from the same speaker samples.

The first question we address is how long we have to train the network using triplet loss. For these experiments, we set the parameters to \( t = 2000 \text{ms} \), \( w = 100 \) and \( h = 50 \), we also set a margin of 2 and we train the model during 1, 2 and 3 days (exploratory experiments with shorter time showed that the minimum training time was a day). Table 3 reports the results. As can be seen 1 day of training is enough to reach a good results.

The second question we address was how large has to be the CNN’s input speech segment. For this we have to explore different parameters of the spectrogram: \( h, t \) while we fix \( w \) to 100 ms which is a common window size for speech signal processing. Varying \( h \) allows us to control the amount of information that passes through in a segment, we explored 25 ms, 33 ms and 50 ms. On the other hand \( t \) allows to control the amount of signal that the CNN will ‘see’, we explore 1 s, 1.5 s and 2 s values; we did not try a larger time since some all recordings were at least 2 s, but not necessary longer. We set the margin to 2 since previous experiments had allowed to identify a good compromise between DR and MSC. Table 4 summarizes the main findings for different combinations of these parameters; each model was trained during a day. As can be seen the more information in the patch the smaller the DR is (which is a good result). This points that a more informative patch is obtained by a larger segment with a small hope size.

Figure 3 shows the projection of embeddings. As it can be seen same speaker embeddings clusters, in this case these clusters represent one speaker’s recording from which we extracted 20 samples that became 20 embeddings. The figure illustrates our best \(( t = 2000 \text{ms} \) and \( h = 25 \)) and worst model \(( t = 1000 \text{ms} \) and \( h = 50 \)). As it can be seen in both projections there is ordering of the speakers (clusters of speakers) but also shows some speakers which are close among themselves. It can been seen that our best model produces more organized positions while for the worst model there is some confusion among the clusters in the middle. For this case Table 5 shows the results of our best model with our test data; as can be seen, our metrics remain consistent in this evaluation.

Table 3 Validation results for different training duration, one day gives the best results and more time does not affect the behavior of the network

| Training days | IAD | IOD | DR   | MSC   |
|---------------|-----|-----|------|-------|
| 1             | 4.77| 27.61| **0.1730** | **0.2475** |
| 2             | **4.61** | 26.62 | 0.1732 | 0.2326 |
| 3             | 4.85| **27.88** | 0.1742 | 0.2358 |

The best performance is highlighted in bold

Table 4 Validation results for different parameters of speech segments

| \( t \) (ms) | \( h \) (ms) | Patch size | IAD | IOD | DR | MSC |
|------------|-------------|------------|-----|-----|----|-----|
| 1000       | 25          | 40 × 256   | 3.21| 22.17| **0.1445** | **0.1924** |
| 1000       | 33          | 30 × 256   | 3.60| 19.36| 0.1860 | 0.1754 |
| 1000       | 50          | 20 × 256   | 4.68| 23.21| 0.2018 | 0.1506 |
| 1500       | 25          | 60 × 256   | 3.57| 25.74| **0.1386** | **0.2209** |
| 1500       | 33          | 45 × 256   | 3.42| 21.31| 0.1605 | 0.2276 |
| 1500       | 50          | 30 × 256   | 3.98| 23.46| 0.1695 | 0.1721 |
| 2000       | 25          | 80 × 256   | 4.69| 30.24| 0.1541 | **0.2766** |
| 2000       | 33          | 60 × 256   | 4.44| 29.70| **0.1495** | 0.2487 |
| 2000       | 50          | 40 × 256   | 4.77| 27.60| 0.2019 | 0.2475 |

The best performance is highlighted in bold

*signals the models with only 4 convolutional layers rather than 5 which were possible for the other patches sizes
As presented in Sect. 4, there is an imbalance between female and male speakers, and between the nationalities of the speakers. To quantify the effect of these imbalances, we performed more evaluations. In the case of gender we created three models: with only female recordings (F, 373 training and 47 validation speakers), with only male recordings (M, 373/1370/171) and training with a comparable amount of male recordings with female recordings (M*, 373/47). For these experiments, we set the parameters to $t = 2000\text{ms}, w = 100\text{ms}$ and $h = 50\text{ms}$, which gives a fast system (small patch) while having a good performance (long window). Table 6 shows the main results on the corresponding validation dataset. As it can be seen there is an effect of gender mismatch; however this is not severe for the case of training with female speakers and using the model with male speakers. On the other direction, we can notice a severe drop in performance. As expected, the best setting is to train with the biggest amount of recordings of a gender and evaluating on that gender.

Table 7 shows our findings for the nationality. Given the amount of speakers, we decided to compare two types of speakers from two regions: Latin America (L) and Spain (S). We set two models, L (373/48) and S (373/48). As we can see there is not a notable difference given the nationality, considering the region Latin America packs more nationalities in the dataset we cannot see an effect of the regional accents on the capabilities of triplet loss in producing good embeddings. Similarly to gender, the mismatch between training and evaluation datasets speakers produces drops in the performance, but not as severe as one might expect.

Finally we show our results for native and non-native speakers. As in the previous section there is variation in the number of native and non-native speakers in the corpus to quantify the effect of these imbalances, we generated three evaluations: with only native recordings (N, 92 training and 4 evaluation of speakers), with 11% of the evaluation set (N, 92/11) and training with comparable amount of speaker recordings for evaluating non-native speakers (NN, 32/4). For these experiments, we set the parameters of our best model $t = 2000\text{ms}, w = 100\text{ms}$ and $h = 25\text{ms}$.
Table 8 shows the results of the proposed metrics for our experiments comparing natives vs non-native speakers. As can be seen, the results are consistent compared to previous experiments, except for the models trained with non-native speakers, where a slight drop in performance is shown, which may be due to the amount of speakers used in training compared to native speakers. However, the performance of the model trained with native speakers and evaluated with non-native speakers of Spanish is remarkable good, which has a similar performance to its counterpart with evaluation of native speakers, this reinforces the theory that the accent acquired when learning a language does not generate variations in the performance of our models, and however this may also be dependent on the level of language management of a non-native speaker of the Spanish language. It will be necessary to carry out more experimentation considering the level of knowledge of the language of the non-native speaker.

### 5.2 Speech embeddings for FSI

In these experiments, we aim to establish a way to calculate the LR based on the distances among embeddings. The more straightforward proposal is to use a normalized distance between the centroids of the reference and questioned sample embeddings, we call this approach distance-based \( D \). The second proposal is to use a ratio between the same distance \( D \) and the distance to the closest speaker in the population in order to account for typicality, we call this approach distance ratio \( DR \). The first proposal can be formalized as:

\[
LR_D = \frac{D(q, r)}{N} \tag{3}
\]

while the second proposal as:

\[
LR_{DR} = \frac{\min (D(q, p))}{D(r, q)} \quad \forall p \in \text{Population} \tag{4}
\]

Where \( q \) is the centroid of the questioned embedding samples, \( r \) is the centroid of the reference and \( p \) is a centroid of a speaker from the population and \( N \) is a normalizing factor.

For the experimentation we used two Forensic Speech Identification scenarios: genuine (were the same-speaker hypothesis is true) and impostor (where the different-speaker hypothesis is true). Per speaker we randomly selected three recordings from where we sampled 15 segments per recording as a reference. In the case of the genuine scenario we selected a fourth recording as the questioned source of samples while for the impostor scenario we randomly selected an extra recording from a different speaker. Additionally, as our population we randomly selected 100 different speakers from which we extracted the same amount of samples that our reference (i.e., 45). It is important to remark that the recordings for these settings came from the validation split of the data. With this considerations, we have 218 genuine cases and 218 impostor cases. Figure 4 shows for both approaches the LR scores: distance based \( (D) \) and distance ratio \( (DR) \).

For the case of the distance as a proxy for \( LR \) \((D)\), as
expected genuine cases are concentrated with lower values with a mean in 0.31 while impostor are located with a mean of 0.59. The Equal Error Rate (EER) for the metric is of 0.1467 with a sensitivity of 1.80. For the case of distance ratio (DR) we see that it corresponds better with the common interpretation of LR where the score indicates an update in the belief on the same speaker hypothesis. In this case, lower scores are associated to impostor cases with a mean of 0.64 and larger score corresponds to genuine cases with a mean of 1.15. This approach has an EER 0.13 and sensitivity index of 1.79.

Figure 5 shows the DET and ROC curves in a log scale for both scores. DET curves are common in the forensic field; in particular it compares the False Match Rate (FMR) with False Non-Match Rates (FNMR) of the system. For both DET curves we see that as the number of false positives grows, it is harder to mismatch a case. A better system produces a curve located to the left bottom corner. Our experiments show that both approaches D and DR share a lot of predictive power. It is important to take into account that DET curves are related to ROC curves, which are more common in the ML field. Within the ROC curves it is common to measure Area Under the Curve (AUC) to compare two systems; in this case we can observe that the DR approach has AUC of 0.9475 while D a AUC score of 0.9402, this points to a slightly more robust discriminative power for the DR metric.

6 Ethical discussion

In recent years, the topic of sharing personal information has become extremely relevant due to the intensive interaction between modern societies and technology, in particular technology that is able to take autonomous decisions. Nowadays, simple things like clicking a “like” button or typing some words to search in a search engine could trigger undesired advertisements or the risk to be scammed in creative ways [17]. As computer scientists, these types of observations raise the question of how datasets are share through the internet in order to contribute to the advances of science but at the same time, not to harm anyone in anyway, as it can be understood in the Hippocratic Oath for artificial intelligence practitioners [9].

In this work, the use of the Voxforge Spanish dataset raises some privacy concerns, since the donors accepted that their voices are part of a database destined to create language technologies however while maintaining anonymity. On the other hand, due to the type of information provided by each donor (mainly a personalized username), it could be easy to identify some of them and performing actions against them like identity fraud in certain speech systems [33]. Fortunately, some of this concerns were contemplated by the developers of the Voxforge datasets which allowed and option to opt-out; however it could hardly be enforced for all distributed copies [42]. Following this concern our experiments do not rely on the username field and it only releases the models, not the recordings. We also will honor any opt-out petition to the Voxforge dataset.

Other aspect of concern regarding this work is if the proposed system is ready to be deployed and used in a real court case. From our perspective, even though the experiments show an interesting and promising performance, it is clear that our models are not ready for deployment in court. In particular, the selection of the population has to be properly. In a real case, it must be transparent who will be included in the population to calculate typicality. Additionally, the results of our models have to be used in the context of other evidence, as previously mentioned the LR has the goal of updating the odds of a legal hypothesis, but it is not a definitive test of identity. So in this context, any attempt to use our models as a recognition system will be ill-advised.
7 Conclusions

In this work, we explored the use of the triplet loss function and a convolution neural network (CNN) for the forensic identification of speakers. We have shown that this setting offers an alternative to well-established approaches. Within this setting, the goal is to train a convolutional network (CNN) to produce an embedding of a speech sample. Although previous research has shown that such embeddings can cluster speakers, in our experimentation we have shown that the distance among clusters can be used to approximate a likelihood ratio (LR) which is a common measurement to convey the plausibility of a legal forensic hypothesis.

Our experimentation points out that the CNN is able to produce a good embedding representation for long speech samples and finer resolution of the spectrogram. We hypothesize this is because it contains more information of a speaker voice. In particular we have focused on Spanish from Latin America and Spain, we have shown that our approach is not affected by variant of Spanish; similarly, when using native and non-native speakers of Spanish, our experiments indicate that there may be a dependency of the model when being trained with non-native speakers; however a model trained with native speakers seems to have no problems in the performance of the model. However, gender has a strong effect on the performance. So far our results point out that models have to be gender dependent, but more researches have to be done on this point (see [8]).

Finally, our results also showed that triplet loss conveys a good discriminative power when the distances are used to approximate LR. We proposed two approaches for the calculation of LR; the first one is solely based on the distances among a questioned and reference samples (D). This approach can be considered independent of the population. The second approach uses the ratio between the first approximation D and the distance of the questioned sample to the closest speaker from the population (DR). We conclude that this second option provides a better alternative, since it conveys information in two aspects: first numerically it corresponds to other LR scores, and second it has a better predictive power (AUC=0.9458); this makes it suitable for the forensic identification of speakers.

Acknowledgements The authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory’s Deep Learning Platform for Language Technologies with the project Experimentos con voz, traduccion y clasificacion de textos (id. PAPTL, 01-008). We also acknowledge Fernanda Hernandez and Sandra Vázquez who were involved in early stages of the development of the code and experimentation.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aitken CG, Lucy D (2004) Evaluation of trace evidence in the form of multivariate data. J Royal Statistical Soc: Series C (Appl Statistics) 53(1):109–122
2. Azul D (2013) How do voices become gendered? a critical examination of everyday and medical constructions of the relationship between voice, sex, and gender identity. In: Challenging popular myths of sex, gender and biology, pp. 77–88. Springer
3. Bimbot F, Bonastre JF, Friedouille C, Gravier G, Magrin-Chagnolleau I, Meignier S, Merlin T, Ortega-García J, Petrovska-Delacrétaz D, Reynolds DA (2004) A tutorial on text-independent speaker verification. EURASIP J Adv Signal Process 2004(4):101962
4. Bredin H (2017) Tristomet: triplet loss for speaker turn embedding. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5430–5434. IEEE
5. Champod C, Vuille J (2011) Scientific evidence in europe-aidmissibility, evaluation and equality of arms. Int Commentary on Evidence 9(1)
6. 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 (2016)
7. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 1, pp. 539–546. IEEE (2005)
8. Ertam F (2019) An effective gender recognition approach using voice data via deeper lstm networks. Appl Acoustics 156:351–358
9. Etzioni O (2018) A hippocratic oath for artificial intelligence practitioners. Tech rep https://techcrunch.com/2018/03/14/a-hippocratic-oath-for-artificial-intelligence-practitioners. Accessed 01/21/2021
10. Garcia-Romero D, McCree A (2014) Supervised domain adaptation for i-vector based speaker recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4047–4051. IEEE (2014)
11. Giannelli PC (1980) The admissibility of novel scientific evidence: Frye v. united states, a half-century later. Colum. L. Rev. 80:1197
12. Gonzalez-Rodriguez J, Fierrez-Aguilar J, Ortega-Garcia J (2003) Forensic identification reporting using automatic speaker recognition systems. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing. 2003. Proceedings. (ICASSP ’03.), vol. 2, pp. II-93. https://doi.org/10.1109/ICASSP.2003.1202302
13. Good IJ (1991) Weight of evidence and the bayesian likelihood ratio. The use of statistics in forensic science pp. 85–106
14. Heigold G, Moreno I, Bengio S, Shazeer N (2016) End-to-end text-dependent speaker verification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5115–5119. IEEE (2016)
15. Hoffer E, Ailon N (2015) Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition, pp. 84–92. Springer
16. Huang J, Li Y, Tao J, Lian Z et al (2018) Speech emotion recognition from variable-length inputs with triplet loss function. In: Interspeech, pp. 3673–3677
17. Irani D, Webb S, Li K, Pu C (2011) Modeling unintended personal-information leakage from multiple online social networks. IEEE Internet Comput 15(3):13–19
18. Khan U, Hernando Pericás FJ (2020) Unsupervised training of siamese networks for speaker verification. In: Interspeech 2020: the 20th Annual Conference of the International Speech Communication Association: 25–29 October 2020: Shanghai, China, pp. 3002–3006. International Speech Communication Association (ISCA) (2020)
19. Li C, Ma X, Jiang B, Li X, Zhang X, Liu X, Cao Y, Kannan A, Zhu Z (2017) Deep speaker: an end-to-end neural speaker embedding system. arXiv preprint arXiv:1705.02304
20. McDermott R, Hatemi PK (2011) Distinguishing sex and gender. PS. Political Sci Politics 44(1):89–92
21. Mena CDH (2020) Librivox spanish ldc2020s01. Linguistic Data Consortium, Web Download. Philadelphia
22. Mo J, Xu L (2020) Self-attention networks for speaker identification with negative-focused triplet loss. In: J Phys: Conference Series, vol. 1601, p. 052004. IOP Publishing
23. Morrison GS (2009) Forensic voice comparison and the paradigm shift. Sci Justice 49(4):298–308
24. Morrison GS (2011) A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: multivariate kernel density (mrvkd) versus gaussian mixture model-universal background model (gmm-ubm). Speech Commun 53(2):242–256
25. Morrison GS, Enzinger E, Ramos D, González-Rodríguez J, Lozano-Díez A (2020) Statistical models in forensic voice comparison, CRC Press LLC Boca Raton, Florida
26. Morrison GS, Zhang C, Enzinger E (2019) Forensic voice science. The Bloomsbury Companion to Phonet pp. 183–197
27. Ouyang K, Liang Y, Liu Y, Rosenblum DS, Yang W (2020) Unsupervised learning of disentangled location embeddings. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
28. Rose P (1997) Identifying criminals by their voice—the emerging applied discipline of forensic phonetics. Aus Language Matt 5(2):6–7
29. Rose P (1998) A forensic phonetic investigation into non-contemporaneous variation in the 1-pattern of similar-sounding speakers. In: Fifth International Conference on Spoken Language Processing (1998)
30. Rose P (1999) Long-and short-term within-speaker differences in the formants of australian“ hello”. J Int Phonet Assoc pp. 1–31
31. Rose P (2002) Forensic speaker identification. cRc Press
32. Rose P, Lucy D, Osanai T et al (2004) Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical random effects model—a non-idiot’s bayes’ approach. Proceedings of the 10th Australian International Conference on Speech Science and Technology (2004)
33. Safavi S, Gan H, Mporas I, Sotudeh R (2016). Fraud detection in voice-based identity authentication applications and services. In: 2016 IEEE 16th international conference on data mining workshops (ICDMW), pp. 1074–1081. IEEE (2016)
34. Saks MJ, Koehler JJ (2005) The coming paradigm shift in forensic identification science. Science 309(5736):892–895
35. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823 (2015)
36. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5329–5333. IEEE (2018)
37. Univaso P, Ale JM, Gurlekian JA (2015) Data mining applied to forensic speaker identification. IEEE Latin Am Transa 13(4):1098–1111
38. Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, Chen B, Wu Y (2014) Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386–1393 (2014)
39. Yaman S, Pelecanos J, Sarikaya R (2012) Bottleneck features for speaker recognition. In: Odyssey 2012-The Speaker and Language Recognition Workshop
40. Zhang C, Koishiida K (2017) End-to-end text-independent speaker verification with triplet loss on short utterances. In: Interspeech, pp. 1487–1491
41. Zhang C, Koishiida K, Hansen JH (2018) Text-independent speaker verification based on triplet convolutional neural network embeddings. IEEE/ACM Trans on Audio, Speech, and Language Process 26(9):1633–1644
42. Zimmer M (2010) “but the data is already public”: on the ethics of research in facebook. Ethics Inform Technol 12(4):313–325

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.