Representation and Metric Learning Advances for Deep Neural Network Face and Speaker Biometric Systems

Victoria Mingote, Antonio Miguel

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{vmingote,amiguel}@unizar.es

Abstract

Nowadays, the use of technological devices and face and speaker biometric recognition systems are becoming increasingly common in people daily lives. This fact has motivated a great deal of research interest in the development of effective and robust systems. However, although face and voice recognition systems are mature technologies, there are still some challenges which need further improvement and continued research when Deep Neural Networks (DNNs) are employed in these systems. In this manuscript, we present an overview of the main findings of Victoria Mingote’s Thesis where different approaches to address these issues are proposed. The advances presented are focused on two streams of research. First, in the representation learning part, we propose several approaches to obtain robust representations of the signals for text-dependent speaker verification systems. While in the metric learning part, we focus on introducing new loss functions to train DNNs directly to optimize the target task for text-dependent speaker, language and face verification and also multimodal diarization.

Index Terms: Representation Learning, Metric Learning, Speaker Verification, Face Verification, Multimodal Diarization

1. Introduction

During the last decades, deep learning techniques have become the dominant approaches in many different fields and tasks, including biometric recognition. As advanced as these techniques are, they still have some problems to be solved such as overfitting issues when the task has limited data that generate problems to generalize to unseen data. Moreover, in some cases, applying techniques with impressive results in a specific task to another task may not produce the expected behaviour. Therefore, in this manuscript, we present a summary of the thesis [1] defended by Victoria Mingote entitled “Representation and Metric Learning Advances for Deep Neural Network Face and Speaker Biometric Systems”[1]. Along this thesis, we have presented different approaches to deal with the previous issues.

First of all, we have focused our efforts on improving the generation of signal representations for the text-dependent speaker verification task, since this task has a strong dependency on phonetic content. Thus, not only the information related to the speaker identity is needed to generate a good enough representation for the verification process. While in the last part, we have analyzed the fact that most of the available verification systems are not always optimized toward the goal task. Hence, we have proposed several approaches using new training loss functions that are based on the final verification metrics. These training loss functions can be applied for each verification task. The following subsections summarise and present the conclusions for each specific part of this dissertation.

2. Representation Learning

At the beginning of this thesis, we have analyzed the effect of introducing the most widespread Deep Neural Network (DNN) architectures to establish the baseline systems for face and text-dependent speaker verification tasks [2, 3]. In this analysis, we have discovered that state-of-the-art DNNs established for many tasks, including face verification, did not perform as efficiently as we expected for text-dependent speaker verification. In the study to find the cause of this poor performance, Fig.1 shows the cosine similarity matrix of the embeddings from two speakers. This illustration depicts that different speakers pronouncing the same phrase have high cosine similarities. So, this effect will lead to detection errors in the text-dependent speaker verification task where the order of the phonetic information is relevant since the speaker identity and the correct phrase are jointly verified to grant access to the systems. Note that this issue could be derived from the use of the global average pooling mechanism since the phonetic order is neglected and the verification performance error metrics achieved are too high.

![Figure 1: Matrix of cosine similarity of embeddings from two speakers pronouncing different phrases.](image)

2.1. Deep Neural Network Supervectors

Motivated by the previous issue, in this thesis, we have presented a new successful approach based on a phonetic phrase alignment mechanism for DNNs to replace the global average pooling [4, 5, 6, 7]. The use of the alignment mechanism allows us to keep the temporal structure and encode the speaker identity and phrase information in a neural network supervector as representation for each utterance. Moreover, different types of approaches such as Hidden Markov Model (HMM) or Gaussian Mixture Model (GMM) combined with Maximum A Posteriori can be employed as alignment mechanism.

The experiments performed with the RSR2015 text-dependent speaker verification database have confirmed that the application of the different alignment mechanisms to obtain the supervector instead of employing global average embeddings provides better results. Fig.2 proves that global average embeddings compared to neural network supervectors are not able
to separate between the same identity with different phrase and the same identity with different phrase which is the base of the text-dependent speaker verification task.

Figure 2: Visualizing Average embeddings vs Neural Network Supervectors for 30 phrases from female using i-SNE.

2.2. Knowledge Distillation with Teacher-Student Architectures

During the development of the alignment mechanism, the effect of varying the size of the training data has also been studied to check that an improvement could be achieved if the available training data is larger. So, we have also noted that the limited amount of training data in the RSR2015 database is another relevant issue to the application of powerful DNNs. This lack of data could produce overconfident predictions and systems are not able to generate enough generalised representations for new data. To address this problem, we have introduced a new architecture philosophy based on the Knowledge Distillation (KD) approach [8]. This architecture consists of two simultaneously trained networks, known as teacher-student architecture. Using this architecture, the student network learns to mimic the predictions of the teacher network capturing the uncertainty. Thus, this approach (BDK) provides robustness to the systems during the training process. In addition, we have included two alternatives to augment the variability in the input of the networks with the Random Erasing (RE) data augmentation which helps to handle a potential overfitting issue due to the lack of data.

The results achieved with the RSR2015-Part I and Part II database have confirmed that the teacher-student architectures improved the generalization capability and better model the variability introduced by the input signals. To illustrate this, the performance results with RSR2015-Part I are shown in Table 1.

Table 1: Experimental results on RSR2015-Part I eval set. These results were obtained to compare the different neural networks with both alignment techniques.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>EER%</th>
<th>min/med/medDCT10</th>
<th>min/med/medCLLR/CLLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN / HMM</td>
<td>0.73</td>
<td>0.142/0.172</td>
<td>0.030/0.032</td>
</tr>
<tr>
<td>CNN (BDK)</td>
<td>0.80</td>
<td>0.149/0.153</td>
<td>0.033/0.036</td>
</tr>
<tr>
<td>CNN / GMM</td>
<td>0.92</td>
<td>0.201/0.271</td>
<td>0.037/0.050</td>
</tr>
<tr>
<td>CNN (BDK)</td>
<td>0.96</td>
<td>0.158/0.156</td>
<td>0.027/0.033</td>
</tr>
</tbody>
</table>

2.3. Class and Distillation Tokens with Multi-head Self-Attention Mechanisms and Memory Layers

The alignment mechanism introduced previously proved to be effective for the text-dependent speaker verification task, but this technique has a main problem, since the temporal alignment is obtained using an external method such as HMM or GMM. Therefore, another alternative approach has been proposed for processing the temporal information, which consists of Multi-head Self-Attention (MSA). MSA layers allow the model to focus on the most relevant frames of the sequence to keep the phonetic information and better discriminate between utterances and speakers. The architecture proposed to use the MSA layers also introduces phonetic embeddings and memory layers to enhance the discrimination [9]. The use of phonetic embeddings compared to standard positional embeddings improves the performance of attention mechanism, as these embeddings help to guide the attention mask to focus on certain phonetic information. Furthermore, the model capacity has been improved by the use of the memory layers. Apart from these techniques, another approach has been incorporated in the above architecture where two learnable vectors have been introduced called class and distillation tokens [10] as Fig.3 shows. These tokens are concatenated to the input before the first MSA layer. Using these tokens during training, temporal information is kept and encoded into the tokens, so that at the end, a global utterance descriptor similar to the supervector is obtained.

Figure 3: Teacher-student architecture used to create the system, where the dashed line indicates the process of backpropagation of the gradients.

The combination of these techniques has been evaluated on the RSR2015-Part II and DeepMine-Part I text-dependent speaker verification databases [11, 12]. Results achieved shown in both databases the power of this kind of approach, even for the RSR2015 database which has suffered from overfitting problems when larger DNN architectures were used in the initial attempts of this thesis due to the small size of the database. Moreover, the use of this approach allows us to enhance the interpretability of what the DNN learns as Fig.4 depicts.

Figure 4: Visualizing two examples of the same phrase of DeepMine-Part I, Ok Google, which are pronounced by different speakers. In both cases, three representations are presented: the spectrogram of each phrase, the attention weights learnt by the class token for each of the 16 heads in the last MSA layer, and the sum of the rows of the previous weight attention matrix.
3. Metric Learning

Throughout the previous part, robust representations were obtained using different approaches based on DNNs. Nevertheless, all the presented approaches were trained with the same loss function which was the traditional Cross-Entropy (CE) loss combined with Ring loss as a complementary loss. Although this training strategy provided reasonably good results for many tasks, its design is not oriented to optimize the verification task itself. Moreover, many systems trained with this strategy usually apply a back-end to perform the verification process. For this reason, we have proposed several new loss functions to train DNN architectures which are based on the verification metrics.

3.1. Optimization of Area Under the ROC Curve

The first approach developed for this part is the use of a triplet neural network as back-end with a training loss function inspired by the Area Under the ROC Curve (AUC) metric. Thus, we have presented a differentiable approximation of this metric called aAUC loss as objective training loss oriented to the goal task [7, 13] as Fig.5 shows. In order to correctly train this kind of approach, the triplet data selection employed is very important. For this reason, we have implemented a smart algorithm to carry out the selection of this triplet training data.

Figure 5: Triplet neural network, the examples are grouped in triplets by the triplet selection to train the network and evaluated the two pairs of embeddings to optimize the loss function.

The use of this type of training strategy have achieved a significant influence on the performance of the system for each of the verification tasks in which it has been applied. Since this approach has been employed to develop text-dependent speaker, language and face verification systems using the RSR2015, LRE and MOBIO databases respectively. Moreover, aAUC loss proposed in this thesis has also been employed for audio segmentation with limited training data in [14, 15].

In Fig.6, we can observe the effect of applying aAUC loss to train the triplet neural network as back-end and use the embeddings extracted from it to make the verification process.

Figure 6: Visualizing Neural Network Supervectors vs Embeddings from aAUC architecture for 30 phrases using t-SNE.

3.2. Approximated Detection Cost Function

The previous aAUC loss has shown relevant success in improving the generalization capability of verification systems using a loss function oriented to the goal task. Nevertheless, this kind of approach with triplets needs smart strategies to create the triplets which involve a high computational cost. Therefore, we have developed the differentiable approximation of the Detection Cost Function (aDCF) [16, 17] to take advantage of the efficiency and speed of multi-class training while a loss function focused on the relevance of decision errors in the verification process is optimized. By optimizing DNNs to minimize this loss function, the system learns to reduce decision errors of verification systems in terms of misses and false alarms. An example of the learning process can be seen in Fig.7 where both terms of error are minimized by gradient backpropagation.

Figure 7: aDCF learning process using the sigmoid functions trained with target and non-target examples.

With this loss function, the verification system trained using the RSR2015-Part I and Part II database have achieved a great performance in both parts. Apart from evaluating the effectiveness of this loss function for training multi-class DNN architectures, we have checked the effects of employing a cosine layer as the last layer in the DNN and compared the performance with some of the state-of-the-art loss functions as Fig.8 depicts.

Figure 8: DET curves for female+male results on RSR2015-Part I using different loss functions.

3.3. Training Enrollment Models by Network Optimization

Following the idea of improving the discrimination ability with back-ends based on metric learning techniques, we have also proposed a new straightforward back-end to address the problems of the high computational cost of our previous back-end with aAUC loss. This approach employs the information learned by the matrix of the last layer of DNN architecture during training with aDCF loss [18, 19, 20]. Using the matrix of this last layer, an enrollment model with a learnable vector is
trained for each enrollment identity to perform the verification process as Fig.9 depicts. This novel back-end employed the matrix from the last layer of the DNN architecture as negative samples combined with the enrollment data as positive samples to train enrollment models for each identity as a binary task. Thus, this training strategy mimics the final verification process.

Figure 9: (a) Left: Example of Embedding Extraction and Training Enrollment Model ID1. (b) Right: Example of Embedding Extraction and Training Enrollment Model ID2.

The experiments using this strategy have improved the system performance and also, the calibration of the resulting systems for the text-dependent speaker verification task with the RSR2015-Part II database as Table 2 presents. Moreover, this new back-end approach has been applied to a multimodal diarization task where face enrollment models have been trained for each identity. Table 3 shows that the use of these models for identity assignment allows us to achieve a relevant improvement in the IberSPEECH-RTVE 2020 Multimodal Diarization Challenge [21] over average embeddings and cosine similarity.

Table 2: Experimental results on RSR2015-Part II eval set. These results were obtained to compare the approach proposed and the cosine baseline.

<table>
<thead>
<tr>
<th>Back-end</th>
<th>Init EER%</th>
<th>Females/Male min/wtDCF</th>
<th>min/CLLR/CLLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>–</td>
<td>5.10</td>
<td>1.90/0.201</td>
</tr>
<tr>
<td>Enroll Model rand</td>
<td>4.72</td>
<td>0.85/0.83</td>
<td>0.190/0.185</td>
</tr>
<tr>
<td>Enroll Model avg</td>
<td>4.46</td>
<td>0.79/0.82</td>
<td>0.170/0.174</td>
</tr>
<tr>
<td>Improvement (%)</td>
<td>12.55</td>
<td>7.06/11.83</td>
<td>11.92/13.43</td>
</tr>
</tbody>
</table>

Table 3: Experimental results on RTVE 2020 Multimodal Diarization test set. These results for face diarization part were obtained to compare the approach proposed and the baseline.

<table>
<thead>
<tr>
<th>Back-end</th>
<th>Non-target Examples DER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Embedding</td>
<td>–</td>
</tr>
<tr>
<td>Enrollment Models</td>
<td>3302</td>
</tr>
<tr>
<td></td>
<td>57</td>
</tr>
</tbody>
</table>

3.4. Log-Likelihood Ratio Cost

Previous dDCF loss has been proven to be an effective approach for training DNN architectures. However, this loss function has a major drawback since it is an application-dependent metric and needed some prior or cost parameter assumptions. In addition, an approximation of the real DCF metric has to be made to use it as objective loss function. Hence, as an alternative approach, Log-Likelihood Ratio Cost (CLLR) has been implemented [22]. CLLR is an application-independent metric which measures the expected log costs of the scores generated for each target and non-target example. Therefore, using CLLR as objective loss, the end-to-end system is trained as in Fig.10 to learn to minimize these costs by obtaining good scores.

The performance of systems trained with CLLR loss has been evaluated on the RSR2015-Part II database. The results in Fig.11 confirm the improvement achieved in the generalization of the learned representations when one of the specific verification metrics is applied to develop the systems.

Figure 10: Graphical example of the main idea behind how DNNs are optimized by minibatch using CLLR loss.

Figure 11: DET curves for female+male results on RSR2015-Part II using the different loss functions evaluated.

4. Conclusions

In this thesis, the main contributions are the following. First, we have presented new approaches to improve the generalization capacity to generate signal representations when facing new situations for text-dependent speaker, language and face verification than the existing ones. The proposed approaches attempt to mimic the final verification process or exploit the efficiency to train multi-class architectures using objective loss oriented to the goal task. These approaches improve the generalization of learned representations for verification. In terms of articles, 3 journal articles and 5 conference papers have been published as result of this part.

5. Acknowledgements

This thesis has received funding from the Spanish Ministry of Economy and Competitiveness and the European Social Fund (TIN2017-85854-C4-1-R), by the Government of Aragon (Reference Group T36_20R), and by Nuance Communications, Inc.
6. References


