DESIGNING AN EFFECTIVE METRIC LEARNING PIPELINE FOR SPEAKER DIARIZATION

Vivek Sivaraman Narayanaswamy[*], *Jayaraman J. Thiagarajan*[†], *Huan Song*[‡], *and Andreas Spanias*[*]

[*]Arizona State University, [†]Lawrence Livermore National Labs, [‡]Bosch Research North America
Email:{vnaray29@asu.edu, jjayaram@llnl.gov, huan.song@us.bosch.com, spanias@asu.edu}

ABSTRACT

State-of-the-art speaker diarization systems utilize knowledge from external data, in the form of a pre-trained distance metric, to effectively determine relative speaker identities to unseen data. However, much of recent focus has been on choosing the appropriate feature extractor, ranging from pre-trained $i-$vectors to representations learned via different sequence modeling architectures (e.g. 1D-CNNs, LSTMs, attention models), while adopting off-the-shelf metric learning solutions. In this paper, we argue that, regardless of the feature extractor, it is crucial to carefully design a metric learning pipeline, namely the loss function, the sampling strategy and the discrimnative margin parameter, for building robust diarization systems. Furthermore, we propose to adopt a fine-grained validation process to obtain a comprehensive evaluation of the generalization power of metric learning pipelines. To this end, we measure diarization performance across different language speakers, and variations in the number of speakers in a recording. Using empirical studies, we provide interesting insights into the effectiveness of different design choices and make recommendations.

Index Terms— Speaker diarization , metric learning, attention models, inverse distance weighted sampling

1. INTRODUCTION

Speaker diarization refers to the problem of attributing relative speaker identities without any prior information about speakers or the nature of speech [1]. This is often used as the first step before invoking downstream inference tasks such as speaker recognition. Diarization systems can be severely challenged by variabilities in acoustic conditions and the need to adapt to speakers with different characteristics. Posed as an unsupervised learning problem, its success relies heavily on the choice of an appropriate distance metric for performing clustering. While classical approaches [2, 3, 4] resorted to careful feature design coupled with a predefined metric, for example cosine similarity between $i-$vectors, more recent solutions have emphasized the importance of integrating a metric learning pipeline into diarization systems [5, 6, 7]. This naturally allows knowledge inferred from an external data source to be utilized while performing diarization on an unseen target data. Powered by recent advances in deep neural networks, there is a surge in interest to construct generalizable latent spaces, that will make the learned metric highly effective for even unseen speakers [7].

In general, metric learning aims to utilize latent features in data to effectively compare observations [8, 9]. This amounts to inferring key factors in data, while encoding higher order interactions, to ensure that examples from the same speaker are within smaller distances, compared to examples from a different speaker [6]. While a variety of formulations exist for supervised metric learning [7, 10], recent approaches have relied on deep networks to construct embeddings that satisfy the supervisory constraints. Popular examples include the *siamese* [11], *triplet* [12, 8], and *quadruplet* [13] networks. By coupling sequence modeling techniques with these deep metric learning formalisms, recent works such as [5, 7] produce state-of-the-art diarization performance, while entirely dispensing the need for explicit feature design.

Though most existing works have focused on choosing the right sequence modeling architecture, it is critical to understand the impact of different components in the metric learning pipeline, from the context of generalization performance. In this paper, we consider three critical components in deep metric learning algorithms, and perform empirical studies to understand their impact on diarization performance: (i) loss function, (ii) strategy for sampling negative examples, and (iii) margin parameter selection. By performing a fine-grained evaluation of generalization to different language speakers and variations in the number of speakers, our study provides interesting insights into choosing the right metric learning architecture for reliable performance.

2. DIARIZATION SYSTEM OVERVIEW

An overview of the diarization system used in our work is illustrated in Figure 1. Though, several existing solutions build

This work was supported in part by the ASU SenSIP center, Arizona State University. Portions of this work were performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE- AC52-07NA27344.

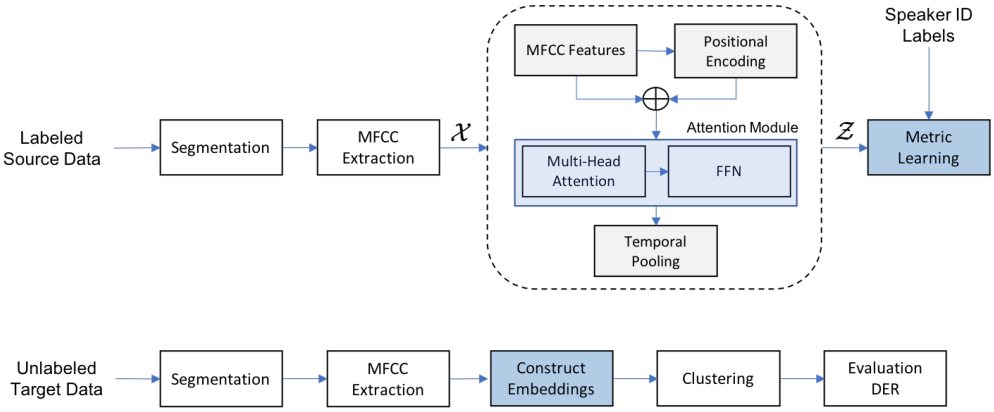

Fig. 1. An overview of the diarization system adopted in this work. Following the state-of-the-art approach in [7], we use raw MFCC features along with deep metric learning to infer embeddings for diarization. The focus of this work is to effectively design different components of metric learning, such that improved generalization is achieved.

upon pre-designed features, such as $i-$vectors, our setup follows the approach in [7] and operates directly on the mel frequency cepstral coefficients (MFCCs) to extract speaker embeddings. In the first stage, an out-of-domain labeled source dataset is utilized to perform metric learning, wherein the speaker ID is used to define positive and negative triplets (or quadruplets). For learning latent features from the sequence data, we adopt the state-of-the-art attention models [14]. Diarization is then performed on a different target dataset by first extracting embeddings with the pre-trained model and then organizing segments using a clustering algorithm.

Preprocessing: In our setup, the speech recordings considered are temporally segmented into non-overlapping segments of equal duration (fixed at 2 seconds). The MFCC features are then extracted using 25ms Hamming windows with 15ms overlap. Consequently, each data sample corresponds to a temporal sequence $\mathbf{x}_i \in \mathbb{R}^{T \times d}$ where T is the number of MFCC frames and d is the number of MFCC dimensions.

Architecture: In this work, attention models are used to learn speaker embeddings from MFCC features, using a metric learning objective. Attention mechanism is a widely-adopted strategy in sequence modeling, wherein a parameterized function is used to determine relevant parts of the input to focus on, in order to make decisions [15, 16]. We use a popular implementation of attention models, *Transformer* [14], which employs the scalar dot-product attention mechanism.

This architecture uses a *self-attention* mechanism to capture dependencies within the same input and employs multiple attention heads to enhance the modeling power. One useful interpretation of self-attention is that it implicitly induces a graph structure for a given sequence, where the nodes are time-steps and the edges indicate temporal dependencies. Furthermore, instead of a single attention graph, we can actually consider multiple graphs corresponding to the different attention heads, each of which can be interpreted to encode different types of edges and hence can provide complementary information about different types of dependencies. This concept is referred to as using *multiple attention heads*. Song *et al.* [7] utilized a variant of this architecture for speaker diarization and our system follows their implementation. As illustrated in Figure 1, the model consists of a multi-head, self-attention mechanism with a feed-forward network (FFN) stacked together L times to learn the deep representations. Besides, positional encoding is included to exploit the ordering information from a sequence.

Clustering: After obtaining the embeddings $\mathcal{Z}$ from the pre-trained model from the out-of-domain data, we perform x-means [17] to estimate the number of speakers, and then use k-means clustering with the estimation. Note, we force x-means to produce at least 2 clusters.

Evaluation Metric: Following standard practice, we use diarization error rate (DER) as the evaluation metric and utilize the `pyannote.metric` [18] package.

3. DESIGN METHODOLOGY

In this section, we describe design choices that we considered pertinent to loss function, the sampling strategy and selection of the margin. The design choices on these components result in a total of 11 realizations of the metric learning pipeline.

3.1. Choice of loss function

We consider two state-of-the-art loss functions to build speaker embeddings from MFCC features with attention models: triplet loss, and quadruplet loss. Denoting the attention model as $\mathcal{A}(.)$, and the Euclidean distance between a pair of embeddings as $D_{ij} = \|\mathcal{A}(\mathbf{x}_i) - \mathcal{A}(\mathbf{x}_j)\|_2$, we describe the definitions of the loss functions in detail.

(i) Triplet loss (Trip) [8]: In a triplet network, every input to the attention model is a group of 3 samples $\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n$, where

$\mathbf{x}_a$ denotes an anchor, $\mathbf{x}_p$ denotes a positive sample from the same class as $\mathbf{x}_a$, and $\mathbf{x}_n$ a negative sample from a different class. Every sample in the set is processed independently by the attention model, and we compute the triplet loss as:

$$l_{\text{trip}}(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) = \max(0, D^2_{ap} - D^2_{an} + \alpha), \qquad (1)$$

where the margin parameter α characterizes the separation between D^2_{an} and D^2_{ap}, such that $D^2_{an} \geq D^2_{ap} + \alpha$. Unlike the contrastive loss [11], the triplet loss does not impose a global margin of separation, and allows a certain amount of distortion in the embedding space.

(ii) Quadruplet loss (Quad) [13]: A well-known criterion for achieving high generalization ability to unseen classes is to reduce the intra-class variability while enlarging the inter-class variability. The recent study on quadruplet network shows that, by adding such a modeling term into the triplet loss, one can decrease the generalization error [13]. More specifically, quadruplet loss includes an additional sample $\mathbf{x}_q$ to the input set of $\{\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n\}$, where $\mathbf{x}_q$ is from a class different than both $\mathbf{x}_a$ and $\mathbf{x}_n$. As a result, the modeling of the intra- and inter-class variations can be achieved by targeting that $D^2_{qn} \geq D^2_{ap} + \alpha_2$, in addition to the triplet loss:

$$\begin{aligned} l_{\text{quad}}(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n, \mathbf{x}_q) = \max(0, D^2_{ap} - D^2_{an} + \alpha_1) \\ + \max(0, D^2_{ap} - D^2_{qn} + \alpha_2) \end{aligned} \qquad (2)$$

where α_1 has the similar effect as in Equation 1 while α_2 balances the two criteria in the training process.

3.2. Choice of sampling strategy

The sampling process is critical in training metric learning architectures, and recent studies have demonstrated that a good sampling strategy can be equally important as the loss formulation in achieving state-of-the-art performance [19]. We describe different sampling strategies under the setting of triplet loss, but the design choice is similar under other loss functions. Given an anchor-positive pair, the naïve way to obtain a negative sample is sample at random. However, this can be sub-optimal as a large number of random examples selected can be *easy* negatives, which do not contribute to the loss at all. In this paper, we consider the following strategies: *(i) Semi-Hard Mining (SH) [12]:* In constrast to random sampling, one can select only a *hard* negative which satisfies $D^2_{an} \geq D^2_{ap} + \alpha$. However, as shown in [12] this can typically lead to a collapsed model. In order to construct more useful triplets (or quadruplets), it is prudent to select only those sets of embeddings that satisfy $D^2_{ap} \leq D^2_{an} \leq D^2_{ap} + \alpha$, referred as *semi-hard* negatives.

(ii) Distance-Weighted Sampling (DW) [19]: Although semi-hard negative mining is effective in practice, it is still a heuristic approach and may not be optimal to cover the high-dimensional sampling space. Consequently, Wu *et al.* analyzed existing sampling strategies, and hypothesized that a better approach can be to reduce the sampling bias, and be more exposed to classes which may lie at the edge of a latent space [19]. Specifically, we construct a discrete probability measure for each example based on the inverse distances to the anchor, and draw samples with the assigned probabilities.

Table 1. Overall performance of different configurations of metric learning. Our recommendations are showed in green.

Sampling	Loss	Margin	DER %
Random	Triplet	Fixed	14.11
Random	Triplet	Adaptive	13.57
Random	Quadruplet	Fixed	13.54
Random	Quadruplet	Adaptive	13.08
Semi-hard	Triplet	Fixed	12.77
Semi-hard	Triplet	Adaptive	14.25
Semi-hard	Quadruplet	Adaptive	13.18
DWS	Triplet	Fixed	12.44
DWS	Triplet	Adaptive	12.98
DWS	Quadruplet	Fixed	12.47
DWS	Quadruplet	Adaptive	12.76

3.3. Choice of margin parameter

Finally, we study the impact of how the margin parameter is chosen. In the *fixed margin (FM)* case, pre-defined values were used throughout the training, $\alpha = 0.8$ ($\alpha_1 = 0.8$ and $\alpha_2 = 0.4$ for quadruplet loss). For the *adaptive margin (AM)* case, as suggested in [13], we compute margin as the difference between the mean of the anchor-negative distance distribution, μ_{an}, and the mean of the anchor-positive distance distribution, μ_{ap}, within every mini-batch processed by the network. During the initial phase of training, the margin is assigned to be a fixed value and as the training progresses, the margin increases and consequently only allows those samples producing a non-zero loss to be evaluated in the gradients computation. Our adaptive margin was calculated as follows:

$$\alpha(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) = \max(0.8, \mu_{an} - \mu_{ap}). \qquad (3)$$

4. EMPIRICAL EVALUATION AND INFERENCES

All variants of the metric learning pipeline were trained on the TEDLIUM corpus which consists of 1495 audio recordings. A set of 1211 speakers with an average recording length of 10.2 minutes were considered after ignoring speakers with less than 45 transcribed segments from the dataset for training. The recordings were down-sampled to 8kHz to match the target CALLHOME corpus. The CALLHOME corpus consists of 780 transcribed, conversation speech recordings from six different languages namely Arabic, Chinese, English, German, Spanish and Japanese, containing 2 to 7 speakers. We

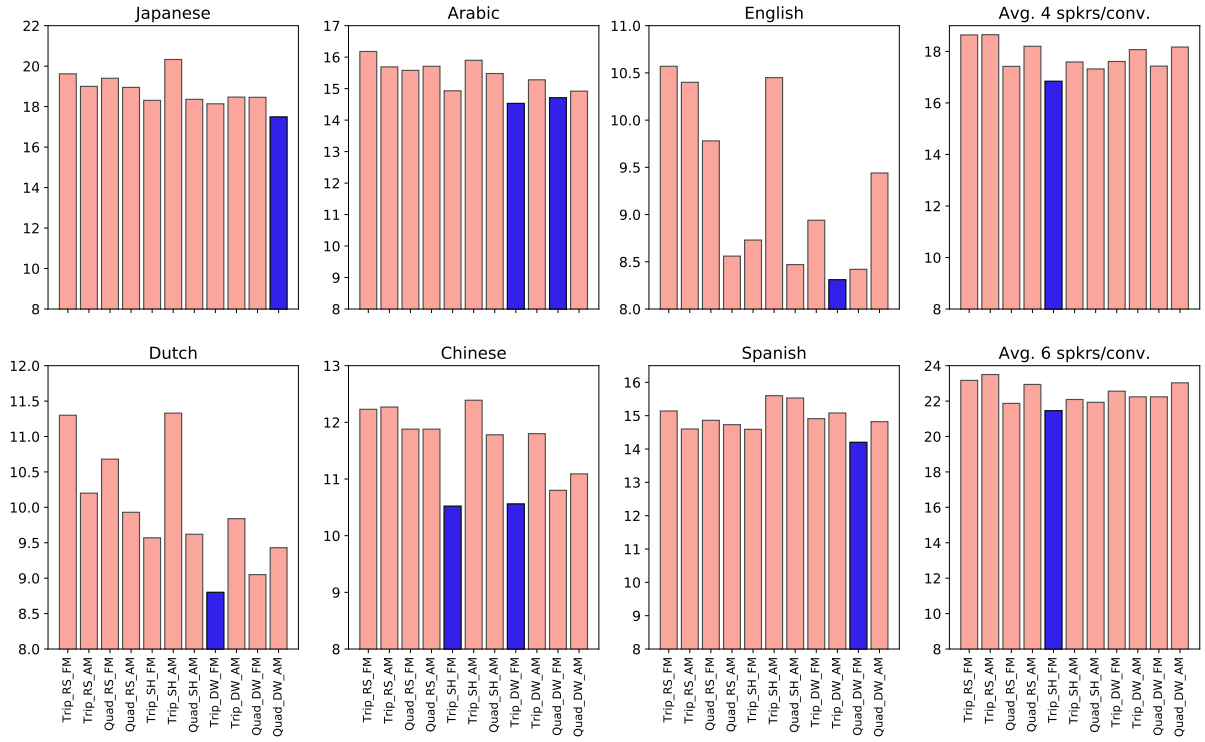

Fig. 2. Fine-grained evaluation of diarization performance. We measure the generalization power of a learned metric by performing diarization with speakers from a specific language, and also with increase in the average number of speakers.

evaluate the overall diarization performance using DER to understand impact of the different design choices.

Although DER collectively considers false alarms, missed detections and confusion errors, most existing systems evaluated on CALLHOME [5] account for only the confusion rate and ignore overlapping segments. Following this convention, we use the oracle speech activity regions and use only the non-overlapping sections. Additionally, there is a collar tolerance of 250ms at both beginning and end of each segment. Table 1 shows the overall diarization performance obtained using different realizations of the metric learning pipeline. By studying the impact of different design choices, we make the following observations: Both the triplet and quadruplet losses are highly effective in constructing generalizable latent spaces, however their performances are suboptimal when random sampling was used. On the other hand, distance weighted negative sampling boosts the performance significantly in all cases. However, semi-hard negative mining was useful only with the triplet loss. Finally, in contrast to state-of-the-art results in vision applications, using an adaptive margin did not seem to provide any improvements to the performance, except in the case of random sampling. Overall, we find that triplet loss with distance weighted sampling produced the lowest DER (12.44%) on the CALLHOME dataset, which is the best reported performance in the literature so far.

Figure 2 shows the language specific diarization performance of the architectures. It can be clearly observed that the DERs for English conversations are relatively lower than other languages. We attribute this to the fact the metric was trained using recordings in English. Another possible reason for higher DER in other languages can be related to the fact that speaker identity is the only semantic information used for training the network and as a result the model is not able to generalize the learned embeddings independent of language.

We also studied the effect of number of speakers in a conversation on the expected diarization performance - for this experiment, we randomly concatenated speech segments from multiple conversations, to produce two variants of the CALLHOME dataset where the average number of speakers per conversation was increased to 4 and 6 respectively. It can be observed from Figure 2, that the DER increases substantially for all the architectures considered, which clearly evidences the gaps in the current art of generalizing a distance metric to unseen scenarios.

In summary, we find that the choice of metric learning pipeline has a crucial role in diarization performance with unseen datasets, and we identify configurations that produce state-of-the-art results. However, we notice significant performance variability across speakers from different languages, and that the performance of diarization systems quickly degrades with increase in number of speakers per conversation.

5. REFERENCES

[1] Sue E Tranter and Douglas A Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] Jan Prazak and Jan Silovsky, "Speaker diarization using plda-based speaker clustering," in *Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2011 IEEE 6th International Conference on*. IEEE, 2011, vol. 1, pp. 347–350.

[3] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 413–417.

[4] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[5] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, "Speaker diarization using deep neural network embeddings," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4930–4934.

[6] Gaël Le Lan, Delphine Charlet, Anthony Larcher, and Sylvain Meignier, "A triplet ranking-based neural network for speaker diarization and linking," *Proc. Interspeech 2017*, pp. 3572–3576, 2017.

[7] Huan Song, Megan Willi, Jayaraman J Thiagarajan, Visar Berisha, and Andreas Spanias, "Triplet network with attention for speaker diarization," *Proc. Interspeech 2018*, pp. 3608–3612, 2018.

[8] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[9] Samuel Berlemont, Grégoire Lefebvre, Stefan Duffner, and Christophe Garcia, "Siamese neural network based similarity metric for inertial gesture classification and rejection," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, 2015, vol. 1, pp. 1–6.

[10] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopez Moreno, "Speaker diarization with lstm," *arXiv preprint arXiv:1710.10468*, 2017.

[11] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, 2015, vol. 2.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[13] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[16] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems*, 2015, pp. 1693–1701.

[17] Dan Pelleg, Andrew W Moore, et al., "X-means: Extending k-means with efficient estimation of the number of clusters.," in *Icml*, 2000, vol. 1, pp. 727–734.

[18] Hervé Bredin, "pyannote. metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, 2017.

[19] R Manmatha, Chao-Yuan Wu, Alexander J Smola, and Philipp Krähenbühl, "Sampling matters in deep embedding learning," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2859–2867.