



Training Speaker Embedding Extractors Using Multi-Speaker Audio with Unknown Speaker Boundaries

Themis Stafylakis^{1†}, Ladislav Mošner^{2†}, Oldřich Plchoť², Johan Rohdin^{2,1},
Anna Silnova², Lukáš Burget², Jan “Honza” Černocký²

¹Omilia - Conversational Intelligence, Athens, Greece

²Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

tstafylakis@omilia.com, imosner@fit.vutbr.cz

Abstract

In this paper, we demonstrate a method for training speaker embedding extractors using weak annotation. More specifically, we are using the full VoxCeleb recordings and the name of the celebrities appearing on each video without knowledge of the time intervals the celebrities appear in the video. We show that by combining a baseline speaker diarization algorithm that requires no training or parameter tuning, a modified loss with aggregation over segments, and a two-stage training approach, we are able to train a competitive ResNet-based embedding extractor. Finally, we experiment with two different aggregation functions and analyze their behaviour in terms of their gradients.

1. Introduction

Speaker embeddings are low-dimensional vectors that capture the voice characteristics of a person. Their role in speech technology is vital, especially in speaker recognition, diarization, separation, as well as in multi-speaker text-to-speech and voice conversion. Speaker embeddings are extracted with neural network architectures, such as TDNNs and ResNets [1, 2, 3]. In a typical setting, the networks are trained in a supervised way, with a training set consisting of several thousands of speakers and multiple recordings per speaker.

For speaker recognition applications, where the training set should contain high intrinsic and extrinsic within-class variability, the most commonly used publicly available corpora are (a) the NIST-SRE repository (distributed by LDC) [4] and (b) the VoxCeleb dataset [5, 3]. The former has been collected and updated during the last two decades and is primarily composed of telephone conversations between two speakers in a controlled setup. The latter is composed of about 7 thousand celebrities and each such recording is an excerpt of an interview available on YouTube. Its creation was based on an innovative pipeline, combining querying YouTube for videos of each celebrity, a ResNet50-based face recognizer [6] that leverages photos of each celebrity, a SyncNet architecture [7] which verifies that the voice comes from the celebrity appearing at a given frame (active speaker verification), and some minor human verification [5]. The decision thresholds were tuned so that the precision of the search is maximized in favor of recall.

The emergence of self-supervision methods in deep learning has also been applied to training speaker embedding extractors [8, 9, 10, 11, 12]. Several approaches have been examined, some of which employ an audiovisual setting [13, 14].

It should be noted though that the whole pipeline of VoxCeleb is heavily dependent on the existence of audiovisual material, and therefore is not applicable to recordings that do not

contain a video component (e.g. telephone conversations, radio broadcasts, etc.). Furthermore, the pipeline inevitably rejects potentially useful portions of the audio of the celebrity, coming from video excerpts during which the celebrity’s face does not appear in the video.

In this work, we explore a novel way of training speaker embedding extractors on datasets containing conversational data (i.e. recordings with more than one speaker and a single channel). Contrary to self-supervised approaches, we assume that the speaker-of-interest labels are available on a recording level. What we consider as unknown are the speech chunks of recordings the celebrities appear in. We believe that this is a realistic setup since labels can be fairly easily obtained, e.g. using queries with celebrity names to a database, similarly to VoxCeleb.

Our method consists of two stages. In the first one, the task is to estimate which chunks of a VoxCeleb2 full-length (uncut) recording belong to the target celebrity. In the second stage, the extractor is trained using these chunks with standard supervised training. The first stage, which is obviously the challenging one, involves a very basic diarization algorithm in order to estimate clusters. We assume that the clusters are of high speaker purity and low speaker coverage, i.e. speakers are typically clustered in more than one cluster. To find the chunks that correspond to the target celebrity, we train from scratch a speaker embedding extractor with Cross-Entropy, Additive Angular Margin (AAM, [15]) loss and a classification head over training celebrities. We create minibatches by randomly sampling segments from each cluster, and we feed those segments to the embedding network. The network extracts a 2D tensor containing the segment-level logits (i.e. number of training celebrities times clusters), which we aggregate across the cluster dimension using either max pooling or its soft version (the well-known log-sum-exp operator). This yields a vector of aggregated logits of size equal to the number of training celebrities which is fed to the softmax layer, and AAM loss is applied using the target celebrity as label. Once the model is trained, we keep those chunks that are classified by the network as belonging to the target celebrity. Finally, based on these chunks, we train a new extractor from scratch, showing competitive performance compared to one trained on the original VoxCeleb segments.

2. Related work

In real-world applications, obtaining labels of plentiful data for strongly supervised training is a costly and resource-demanding process. It turns out to be much easier and cheaper to acquire data points with incomplete, inexact, or inaccurate labels [16] leading to *weakly supervised* training. In our study, we aim

†These authors contributed equally to this work.

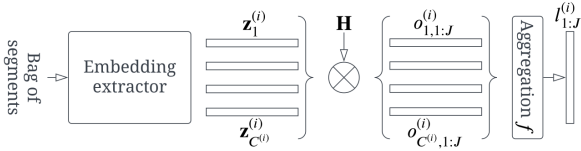


Figure 1: *Proposed method to weakly supervised training of a speaker embedding extractor. The \otimes symbol stands for matrix multiplication.*

at employing *inexact* supervision, which generally deals with data having coarse-grained labels. The task at hand is commonly referred to as *multi-instance learning* [17] where labels (speaker identities) are known for input *bags* (recordings). Bags contain individual *instances* (audio segments), hence the name. According to a multi-instance learning taxonomy [18], the presented approach falls within vocabulary-based methods of an embedded space paradigm. Despite enforcing bag classification in training, instance embeddings are of interest in test time.

As summarized in [18], multi-instance learning has been applied to tasks of multiple fields. In speech processing, weak supervision in the context of automatic speech recognition was explored in [19]. The authors trained a word-level acoustic model with the bag-of-words annotations. Their and our approaches are similar in terms of aggregation via log-sum-exp operation. The closest to our approach is that presented in [20, 21], where the goal is speaker identification and verification. In contrast to our method, [20] requires a trained diarization model which provides audio segments for the i-vector [22] extraction. I-vectors are used as input to a speaker classifier trained with an objective encompassing weak labels. Eventually, the classifier labels data for the final model training.

Finally, there are works and datasets that deal with the problem of multi-speaker enrolment or test recordings [23, 24]. However, the extractor is trained on fully-supervised methods and their focus is on scoring rather than on training.

3. Proposed method

3.1. Assumptions

We assume that we have a set of R recordings for training an embedding extractor. Moreover, we assume that we know all the celebrities appearing in each recording¹. We do not know though where these celebrities appear in the recording, for how long they speak, or how many other speakers participate in the recording. We denote the set of celebrities by \mathcal{J} and $J = |\mathcal{J}|$ is their number. Note that we use the term *celebrities* as a reference to VoxCeleb, to describe those speakers included in the training set and the classification head of the network.

3.2. Baseline Speaker Diarization

There are several speaker diarization algorithms that can yield state-of-the-art performance, ranging from methods that combine neural approaches with probabilistic clustering (e.g. [25, 26]) to fully end-to-end neural approaches (e.g. [27]). A weakness of these methods is their dependence on annotated datasets in order to train. As the purpose of this paper is to minimize the need for annotation, we will employ a very baseline diarization approach, namely (a) speaker change detection, followed by (b) BIC-based agglomerative hierarchical clustering with Gaussians on MFCC features for modeling segments, followed by (c)

¹In this work, we removed recordings with more than one celebrity, although our method can be fairly easily extended to include such cases.

Viterbi-based boundary refinement using (maximum-likelihood trained) Gaussian Mixture Models. No pretrained models (such as Universal Background Model) are used for diarization. The SIDEKIT software is used for diarizing the recordings [28]. We use the default hyperparameters, which in general tend to over-estimate the number of speakers in a recording. This is needed for our algorithm in order to avoid merging chunks of different speakers in the same cluster, and to ensure that during training, at least one sampled segment corresponds to the target celebrity of the recording (see Section 3.3.2). The clustering is fixed during training, i.e. we do not re-cluster the chunks while training the extractor, as some self-supervised algorithms do [9, 10].

3.3. Training the extractor with score-level aggregation functions

3.3.1. Creating minibatches

Assume that i th recording has been split into a set of clusters with index set $\mathcal{C}^{(i)}$, where $|\mathcal{C}^{(i)}| = C^{(i)}$. Each cluster $c \in \mathcal{C}^{(i)}$ comprises chunks $S_c^{(i)}$. A minibatch indexed by b contains a set of segments $\mathcal{B}^{(b)}$ corresponding to the clusters of recordings $\mathcal{I}^{(b)}$, where size $|\mathcal{I}^{(b)}|$ varies with b . The set of segments in a minibatch is defined as $\mathcal{B}^{(b)} = \bigcup_{i \in \mathcal{I}^{(b)}} \mathcal{S}^{(i)}$, where $\mathcal{S}^{(i)} = \{s_c^{(i)}\}_{c \in \mathcal{C}^{(i)}}$. Segments $s_c^{(i)}$ are cut out of chunks $S_c^{(i)}$ that are uniformly sampled from clusters and each cluster is represented by at least one chunk.

3.3.2. Aggregating over the cluster dimension

In the forward pass, each segment is processed individually (i.e. independently of the recording it belongs to) including the statistics pooling layer, the embedding, and the dot-products of the embeddings with the linear layer of the classification head $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_J]$, where \mathbf{h}_j denotes the set of weights for speaker j . This operation yields a 2D tensor (ignoring the batch dimension) of similarities between segments and celebrities. Tying between segments of the same recordings is implemented via an aggregation operation, which transforms the tensor into a vector of size J , indicating the similarity between the recording and celebrities.

The segment-level similarities are defined as $\mathbf{o}_j^{(i)} = \{o_{c,j}^{(i)}\}_{c \in \mathcal{C}^{(i)}}$, where $o_{c,j}^{(i)} = \mathbf{h}_j^\top \mathbf{z}_c^{(i)}$ and $\mathbf{z}_c^{(i)}$ is the embedding of the example belonging to the c th cluster. The recording-level similarities $\{l_j^{(i)}\}_{j \in \mathcal{J}}$ are derived by an aggregation function:

$$l_j^{(i)} = f(\mathbf{o}_j^{(i)}; \tau) \quad (1)$$

as shown in Fig. 1. We refer to these recording-level similarities as logits, since they are the ones passed to the SoftMax function

$$p(j|\mathcal{S}^{(i)}) = \frac{\exp(sl_j^{(i)})}{\sum_{j' \in \mathcal{J}} \exp(sl_{j'}^{(i)})}, \quad (2)$$

where s is the scale of the AAM loss, while the target speaker logit is penalized by the margin of the AAM loss [15].

3.3.3. Aggregation functions

The most commonly used aggregation functions are the average and max pooling. However, both functions have certain undesired properties for our task. Average pooling would encourage the embeddings of all segments of the i th recording to be close

Table 1: Comparison of VoxCeleb2 derivatives.

Dev. set	Celebrities	Recordings	Hours
Original	5,994	145,569	2,369.0
Restricted	5,987	110,940	1,884.3
Uncut	5,987	110,940	10,211.6
Self-labeled	5,935	88,084	4,483.8

to the representation \mathbf{h}_j of the target celebrity of the recording. This property makes it inadequate for the task since only a fraction of the segments belongs to the target celebrity.

Max pooling seems to be more adequate since it focuses only on the segment having the highest similarity with the target celebrity. However, max pooling propagates gradients only through this highest-similarity segment, ignoring all other segments of the recordings. Furthermore, max pooling might make the network hard to train in the warm-up phase, when the network is randomly initialized.

Given the above, we also experiment with a soft version of max pooling, the log-sum-exp:

$$l_j^{(i)} = f_{\text{LSE}}(\mathbf{o}_j^{(i)}; \tau) = \tau \log \frac{1}{C^{(i)}} \sum_{c \in C^{(i)}} \exp\left(\frac{1}{\tau} o_{c,j}^{(i)}\right), \quad (3)$$

where $\tau > 0$ is a temperature. Note that for $\tau \rightarrow 0^+$ we obtain max pooling, i.e. $f_{\text{LSE}}(\mathbf{o}_j^{(i)}; \tau) \rightarrow \max_c \left(\{o_{c,j}^{(i)}\}_{c \in C^{(i)}}\right)$.

3.3.4. Gradients and loss function

It is interesting to derive the gradients when log-sum-exp is applied. Let j^* be the (ground-truth) celebrity of the i th recording and let $L^{(i)}$ be the CE loss. By applying the chain rule, we get

$$\frac{\partial L^{(i)}}{\partial o_{c,j}^{(i)}} = p(c|j, \mathcal{S}^{(i)}) \left[p(j|\mathcal{S}^{(i)}) - \delta_{j,j^*} \right], \quad (4)$$

where δ_{j,j^*} is the Kronecker delta and

$$p(c|j, \mathcal{S}^{(i)}) = \frac{\exp\left(\frac{1}{\tau} o_{c,j}^{(i)}\right)}{\sum_{c' \in C^{(i)}} \exp\left(\frac{1}{\tau} o_{c',j}^{(i)}\right)}. \quad (5)$$

To derive the above, we used the fact that the gradient of $l_j^{(i)}$ w.r.t. $o_{c,j}^{(i)}$ is $p(c|j, \mathcal{S}^{(i)})$. As we observe, the error signal $\epsilon_{j,j^*}^{(i)} = p(j|\mathcal{S}^{(i)}) - \delta_{j,j^*}$ is distributed between the clusters proportionally to $p(c|j, \mathcal{S}^{(i)})$. Moreover, for utterances not containing the j th speaker, the error is proportional to $p(j|\mathcal{S}^{(i)})$, meaning that the norm of the gradient is small for recordings not containing speakers similar to the j th celebrity.

Note also that if we consider length-normalized speaker representations and embeddings, i.e. $\|\mathbf{h}_j\| = \|\mathbf{z}\| = 1$, then $-1 \leq f(\mathbf{o}_j^{(i)}; \tau) \leq +1$ for both aggregation functions we examine, making AAM loss applicable [15].

3.4. Selecting chunks from the celebrities

Once training of the extractor is completed, we need to select the chunks corresponding to the target speaker. There are several methods that can be employed for this task. We implement a simple selection policy, where each chunk (as defined by the segmentation and boundary refinement algorithms of speaker

diarization without the hierarchical clustering) is considered as coming from the celebrity of the recording if and only if the network classifies it as such. That is,

$$S^{(i)} \in \hat{\mathcal{S}}_{j^*} \text{ iff } \operatorname{argmax}_j \left(\{l_j^{S^{(i)}}\}_{j \in \mathcal{J}} \right) = j^*, \quad (6)$$

where $S^{(i)}$ is a chunk from the i th recording, $\hat{\mathcal{S}}_j$ is the set of estimated chunks for the j th celebrity, and j^* is the celebrity of the i th recording. The chunk selection stage is performed after training, so neither an aggregation function nor the AAM margin are applied. We simply classify all segments based on the logits denoted by $\{l_j^{S^{(i)}}\}_{j \in \mathcal{J}}$.

4. Experimental setup

4.1. Speaker embedding extractor and classifier

All presented models trained in a weakly supervised fashion share the same structure: per-segment embedding extractor, an aggregation function with similarity computation, and AAM-based logit computation [15].

The backbone of embedding extractors is based on the ResNet34 architecture [6], widely used in the field of speaker verification [14]. 400 frames of 80-dimensional log Mel-filter bank energy (fbank) features constitute the input to the model. Individual stages of the network comprise 2D convolutional layers with 64, 128, 256, and 256 filters each. The building blocks of stages follow the pre-activation structure [29]. Instead of a standard batch normalization, we opt for instance normalization [30]. The rationale behind this choice is as follows: During training, segments do not need to contain only speech due to a simple diarization. Statistics collected by batch normalization layers are in turn also affected by non-speech content, which is in contrast with the test phase. The resulting embeddings are 256-dimensional. The embedding extraction is followed by aggregation described in Section 3.3.2.

Thanks to the favorable property of the aggregation functions – values bounded in the interval $[-1, 1]$ – the output of the preceding stage can be considered as similarity scores (analogous to the cosine similarity computed in fully supervised training). Therefore, logits for the CE loss function can be computed according to margin-based methods. In our case, the margin is introduced following the AAM formulation. In all of our experiments, we use a scale $s = 30$, see Eq. (2). It is a common practice to increase the margin during training. Since we experimentally found it challenging to schedule it in weakly supervised settings, we keep it fixed at 0.1 (unless explicitly stated).

The models are trained using SGD with a momentum of 0.9. To prevent from diverging in the initial stage of such challenging training, we employ a learning rate warm-up. After reaching a value of 0.2, the learning rate is scheduled according to the loss on the cross-validation set comprising 256 speakers whose recordings were extracted from the training set.

4.2. Strongly and weakly labeled datasets

For the sake of comparison, we choose an exemplary training corpus termed *VoxCeleb2 Uncut dev*. It has been acquired by downloading a complete audio material of recordings present in VoxCeleb2 dev. We stress that we use only full-length audio recordings without a video content of the corresponding clips. It might seem wasteful not to use another available modality, but we target a more general goal – utilization of weakly labeled audio-only data, which is much easier to obtain compared to strongly labeled segments.

Table 2: Speaker verification results with models trained in strongly and weakly supervised fashion. m stands for the AAM margin, \uparrow means scheduling up to the value to the right of the arrow. LSE means log-sum-exp. $P_{tar} = 0.05$ for MinDCF. EER is in %.

	Supervision	Data	m	Aggregation / τ	Speaker repres.	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
						EER	MinDCF	EER	MinDCF	EER	MinDCF
s1	strong	restricted	0.1	–	1	1.56	0.111	1.56	0.100	2.81	0.164
s2	strong	restricted	\uparrow 0.3	–	1	1.24	0.074	1.34	0.087	2.46	0.142
s3	weak	uncut	0.1	max pooling / –	1	4.75	0.283	5.05	0.300	7.44	0.395
s4	weak	uncut	0.1	LSE / 0.5	1	6.03	0.381	6.39	0.391	9.99	0.529
s5	weak	uncut	0.1	LSE / 0.5 \rightarrow 0.1	1	4.53	0.286	4.86	0.292	7.31	0.401
s6	pseudo	self-labeled	\uparrow 0.3	–	1	1.73	0.114	1.80	0.114	3.15	0.184
s7	pseudo	self-labeled	\uparrow 0.3	–	2	1.58	0.103	1.82	0.118	3.23	0.184

Since some video clips are no longer available, *VoxCeleb2 Uncut* is restricted compared to the original *VoxCeleb2* in terms of speakers and recordings. The comparison is presented in Table 1. To provide a fair comparison with baselines trained in a standard supervised way, we limit the speakers and recordings of *VoxCeleb2* to match the *uncut* version – *VoxCeleb2 Restricted dev*. In both weakly- and strongly-supervised training, the data is sampled at 16 kHz. Online reverberation and Musan noise [31] augmentations are applied to boost robustness.

5. Experiments

The evaluation is performed on the standard sets of verification trials – *VoxCeleb1-O*, *E*, and *H*. We present results in terms of equal error rate (EER [%]) and minimum detection cost (MinDCF), with prior probability of target trials $P_{tar} = 0.05$.

5.1. Strong vs. weak supervision

In the first section of Table 2, we present the results of a standard supervised training with both scheduled ($0.1 \rightarrow 0.3$) and constant (0.1) margin m of AAM. We show how enforced between-speaker separability can improve performance to stress the potential of proposed weakly supervised training, which utilizes a constant margin of 0.1. As per our experimentation, however, more sophisticated treatment of margin within our framework is required. We note that the models’ performance is affected by a limited size of *VoxCeleb2 Restricted dev* used for training.

The next set of rows of Table 2 clearly shows that proposed weakly supervised training is viable and already leads to reasonably well-performing models. Obviously, multi-instance learning is much harder than training with strong supervision, which is reflected by a deterioration of the results compared to the baseline. We compare two possible aggregation functions: max pooling and log-sum-exp. Despite its limitations described in Section 3.3.2, the max pooling is a plausible option and even outperforms log-sum-exp with a constant temperature of 0.5. We assume the worse performance yielded by $\tau = 0.5$ is due to over-smoothness. On the other hand, scheduled τ changing from 0.5 to 0.1 eases training, and the aggregation gets closer to the maximum in later stages of training. We observe that a soft version of the maximum aggregation can lead to improved performance. As τ has a significant impact, it calls for further exploration. We suspect that better scheduling has the potential to boost performance.

5.2. Training with self-labeled data

Having trained models in a weakly supervised fashion, we use the best one (s5) to select chunks of the *uncut* set according to (6). The procedure from Section 3.4 provides the *self-labeled*

training set with parameters presented in Table 1. Even though some speakers and recordings have been lost, it is 2.4 times larger than the *restricted* set. We proceed with supervised training with obtained pseudo labels where aggregation function is no longer required.

Table 2 shows that the resulting system s6 greatly outperforms the model used for labeling. At the same time, its performance gets close to that of strongly supervised models.

As a result of inaccurate diarization and the chunk selection approach, the *self-labeled* set contains non-speech signals or chunks with speakers without labels (e.g. interviewers). To deal with label noise, we introduce two representations per speaker in the classification head \mathbf{H} [32, 10]. This allows it to learn true speaker and interference representations for each celebrity. Although a considerable improvement can be observed on *VoxCeleb1-O* (s6 vs s7), the results on the other two sets do not show a similar trend.

6. Conclusions

In this paper, we presented an approach to weakly supervised training of speaker embedding extractors with mere recording-level labels. We demonstrated our method using the full-length audio recordings of *VoxCeleb2* clips. We initially trained an extractor equipped with an aggregation function to estimate chunks belonging to the target celebrity. Using these chunks, we trained a model from scratch, attaining performance close to the model trained on the original *VoxCeleb* segments.

Encouraged by the results, we plan to extend the approach by introducing alternative ways of aggregation (such as attention) and leverage recordings containing more than one celebrity. We will also explore combinations of weak and self-supervision for training the model. Finally, we will experiment with applying a second, embedding-based diarization stage (using embeddings from the first extractor) and make use of other (non-celebrity) speakers in training.

7. Acknowledgements

The work was supported by Czech Ministry of Interior project No. VJ01010108 ”ROZKAZ”, Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, European Union’s Horizon 2020 project No. 833635 ROXANNE and Horizon 2020 Marie Skłodowska-Curie grant ESPERANTO, No. 101007666. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project ”e-Infrastructure CZ – LM2018140”.

8. References

- [1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [2] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.
- [3] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.
- [4] S. O. Sadjadi, "NIST SRE CTS superset: A large-scale dataset for telephony speaker recognition," *arXiv preprint arXiv:2108.07118*, 2021.
- [5] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [7] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.
- [8] T. Stafylakis, J. Rohdin, O. Plchot, P. Mizera, and L. Burget, "Self-supervised speaker embeddings," in *Proc. Interspeech 2019*, 2019, pp. 2863–2867.
- [9] D. Cai, W. Wang, and M. Li, "An iterative framework for self-supervised deep speaker representation learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6728–6732.
- [10] J. Thienpondt, B. Desplanques, and K. Demuynck, "The IDLAB VoxCeleb speaker recognition challenge 2020 system description," *arXiv preprint arXiv:2010.12468*, 2020.
- [11] J. Slavíček, A. Swart, M. Klčo, and N. Brümmer, "The Phonexia VoxCeleb Speaker Recognition Challenge 2021 System Description," *arXiv preprint arXiv:2109.02052*, 2021.
- [12] S. H. Mun, M. H. Han, D. Lee, J. Kim, and N. S. Kim, "Bootstrap equilibrium and probabilistic speaker representation learning for self-supervised speaker verification," *IEEE Access*, vol. 9, pp. 167 615–167 627, 2021.
- [13] D. Cai, W. Wang, and M. Li, "Incorporating visual information in audio based self-supervised speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.
- [14] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, "VoxSRC 2021: The Third VoxCeleb Speaker Recognition Challenge," *arXiv preprint arXiv:2201.04583*, 2022.
- [15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [16] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 08 2017.
- [17] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.
- [18] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.
- [19] V. Pratap, Q. Xu, T. Likhomanenko, G. Synnaeve, and R. Collobert, "Word Order Does Not Matter For Speech Recognition," *arXiv preprint arXiv:2110.05994*, 2021.
- [20] T. Alu m e, "Training Speaker Recognition Models with Recording-Level Labels," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1066–1072.
- [21] M. Karu and T. Alu m e, "Weakly Supervised Training of Speaker Identification Models," *arXiv preprint arXiv:1806.08621*, 2018.
- [22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [23] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.
- [24] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Interspeech*, 2016, pp. 818–822.
- [25] M. Diez, L. Burget, S. Wang, J. Rohdin, and H.  ernock y, "Bayesian HMM based x-vector clustering for Speaker Diarization," in *Proc. Interspeech 2019*, 2019, pp. 346–350.
- [26] A. Silnova, N. Br ummer, J. Rohdin, T. Stafylakis, and L. Burget, "Probabilistic embeddings for speaker diarization," in *Odyssey 2020: The Speaker and Language Recognition Workshop*, 2020.
- [27] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," in *Proc. Interspeech 2020*, 2020.
- [28] P.-A. Broux, F. Desnous, A. Larcher, S. Petitrenaud, J. Carrive, and S. Meignier, "S4D: Speaker diarization toolkit in python," in *Proc. Interspeech 2018*, 2018.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630–645.
- [30] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *arXiv preprint arXiv:1607.08022*, 2016.
- [31] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [32] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, "Sub-center ArcFace: Boosting Face Recognition by Large-Scale Noisy Web Faces," in *Computer Vision – ECCV 2020*, 2020, pp. 741–757.