Improving speaker turn embedding by crossmodal transfer learning from face embedding

Nam Le[1,2], Jean-Marc Odobez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédéral de Lausanne, Switzerland
{nle, odobez}@idiap.ch

Abstract

Learning speaker turn embeddings has shown considerable improvement in situations where conventional speaker modeling approaches fail. However, this improvement is relatively limited when compared to the gain observed in face embedding learning, which has proven very successful for face verification and clustering tasks. Assuming that face and voices from the same identities share some latent properties (like age, gender, ethnicity), we propose two transfer learning approaches to leverage the knowledge from the face domain learned from thousands of identities for tasks in the speaker domain. These approaches, namely target embedding transfer and clustering structure transfer, utilize the structure of the source face embedding space at different granularities to regularize the target speaker turn embedding space as optimizing terms. Our methods are evaluated on two public broadcast corpora and yield promising advances over competitive baselines in verification and audio clustering tasks, especially when dealing with short speaker utterances. The analysis gives insight into characteristics of the embedding spaces and shows their potential applications.

1. Introduction

As the daily production of broadcast TV and internet content is growing quickly everyday, it is an essential task to make large multimedia corpora easily accessible through search and indexing. Therefore, research effort has been devoted to unsupervised segmentation of videos into homogeneous segments according to person identity, one of which is speaker diarization, *i.e.* segmenting an audio stream according to the identity of the speaker. It allows search engines to answer the question ”who speaks when?” and to create rich transcription of ”who speaks what?”.

In the literature, state-of-the-art Gaussian-based speaker diarization methods have been shown to be successful in various types of content such as radio or TV broadcast news, telephone conversation and meetings [24, 19, 27]. In these contents, the speech signal is mostly prepared speech and clean audio, the number of speakers is limited, and the duration of speaker turn (i.e. a speech segment of one speaker) is more than 2 seconds on average. When these conditions are not valid, in particular the assumption of speaker turn duration, the quality of speaker diarization deteriorates [30]. As shown in TV series or movies, state-of-the-art approaches do not perform well [7, 3] when there are many speakers (from 28 to 48 speakers), or speaker turns are spontaneous and short (1.6 seconds on average in the Game of Thrones TV series). To alleviate these shortcomings of speaker diarization, researches have been proposed along two fronts: better methods to learn speaker turn embeddings or utilizing the multimodal nature of video content. The recent work on speaker turn embedding using triplet loss shows certain improvements [4]. Other multimodal related works focus on late fusion of two streams by propagating labels [2, 5] or high level information such as distances or overlapping duration [11, 29].

In this work, we unite the two fronts by proposing crossmodal transfer learning from a face embedding to improve a speaker turn embedding. Indeed recently, learning face embeddings has made significant achievements in all tasks, including recognition, verification, and clustering [31, 26]. To transpose these advances to the speaker diarization domain, a neural network for speaker turn embedding trained with triplet loss (*TristouNet*) was proposed in [4]. Nevertheless, the improvement of this network architecture over the Gaussian-based methods was quite incremental compared to the gain obtained when using such methods in learning face embeddings. To explain this disparity between modalities, one can point to the clear difference in amounts of training data, as there are hundreds of thousands images from thousands identities in any standard face dataset. The limited size of speech data is very challenging to overcome because we cannot use Internet search engines to collect speech segments similarly to face images in [26, 34]. Moreover, manual labeling speech segments is much more costly. To mitigate the need of massive dataset, we take advantage of pretrained face embeddings by relying on the multimodal nature of person diarization.

Although transfer learning is widely applied in other topics [35, 23], transferring between acoustic and visual domains has mainly been applied to the task of speech recognition [25], in which the two streams are highly correlated. On the other hand, with respect to identity, because there is not a definite one-one inference from a face to a voice, it is still an open question of how to apply transfer learning between a face embedding and a speaker embedding. To answer this question, we start with an observation. Although one cannot find the exact voice of a person given only a face, however, if given a small set of potential candidates, it is possible to pick a voice which is more likely to come from the given face than other voices. For example, when most candidates are male voices then it is more likely to find the correct one if the voice is female. Thus, there are latent attributes which are shared between the two modalities. Rather than relying on multimodal data with explicit shared labels such as genders, ages, or accent and ethnicity, we want to discover the latent commonalities from the source domain, a face embedding, and transfer to the target domain, a speaker turn embedding. Therefore, our hypothesis is that by enforcing the speaker turn embedding to have the same geometric properties with the face embedding with respect to identity, we can improve the performance of the speaker turn embedding.

Because from one space, there are different properties to be used as constrains to be enforced on the other space, we propose two different strategies aiming at different granularity for transferring:

- Target embedding transfer: We are given the identity correspondences between the 2 modalities. Hence, given the 2 inputs from the same identity, one can force the desired embedded features of the speaker turn to be close to embedded features of the face. Minimizing the disparity between the 2 embedding spaces with respect to identity will act as a regularizing term for optimizing the speaker turn embedding.
- Clustering structure transfer: This approach focus on discovering shared commonalities between the 2 embedding spaces such as age, gender, or ethnicity. If a group of people share common facial traits, we expect their voices to also share common acoustic features. In particular, the shared common traits in our case is expressed as belonging to the same cluster of identities in the face embedding space.

Experiments conducted on 2 public datasets REPERE and ETAPE show significant improvement over the competitive baselines, especially when dealing with short utterances. Our contributions are also supported by crossmodal retrieval experiments and the visualization of our intuition.

The rest of the paper is organized as follows. Sec. 2 reviews other works related to ours, Sec. 3 introduces triplet loss and the motivation of our work, Sec. 4 describes our transfer methods in details. Sec. 5 presents and discusses the experimental results, while Sec. 6 concludes the paper.

2. Related Work

Below we discuss prior works on audio-visual person recognition and transfer learning which share similarities with our proposed methods.

As person analysis tasks in multimedia content such as diarization or recognition are multimodal by nature, significant effort has been devoted to using one modality to improve another. Several works exploit labels from the modality that has superior performance to correct the other modality. In TV news, as detecting speaker changes produces less false alarm rate and less noise than detecting and clustering faces, speaker diarization hypothesis is used to constrain face clustering, *i.e.* talking faces with different voice labels should not have the same name [2]. Meanwhile in [5], because face clustering outperforms speaker diarization in TV series, labels of face clusters are propagated to the corresponding speaker turns. Another approach is to perform clustering jointly in the audio-visual domain. [29] linearly combines the acoustic distance and the face representation distance of speaking tracks to perform graph-based optimization; while [11] formulates the joint clustering problem in a Conditional Random Field framework with the acoustic distance and the face representation distance as pair-wise potential functions. Beside late fusion of labels, early fusion of features proposed in [18, 28] is only suitable for supervised tasks; and because their datasets are limited with 6 identities, the case is not conclusive. Note that the aforementioned works focus on aggregating two streams of information whereas we emphasize on the transfer of knowledge from one embedding space to another. By applying recent advances in embedding learning, with deep networks for face [26, 31] and speaker turn [4] our goal is not only to improve the target task (as speaker turn embedding in our case) but also provide a unified way for multimodal combination.

Each of our learning approaches draw inspiration from a different line of research. First, we can point to coupled matching of image-text or heterogeneous recognition [21, 17, 22] or harmonic embedding [31] as related background for our target embedding transfer. Since it is arguable that audio-visual identities contain less correlated information, our method uses the one-one correspondence as a regularization term rather than as an optimal goal. Second, co-clustering information and cluster correspondence inference have been used in transfer learning on traditional tasks of text mining [35, 23]. As one identity is enforced to have the same neighbors in both face embedding and speech embedding spaces, our work is therefore closely related to metric imitation [8] or transfer learning through projection ensemble [9]. In this work, we expand that concept into exploiting clustering structure of person identities for crossmodal learning.

3. Triplet loss and motivation

Given a labeled training set of $\{(x_i, y_i)\}$, in which $x_i \in \mathbb{R}^D, y_i \in \{1, 2, .., K\}$, we define an embedding as $f(x) \in \mathbb{R}^d$, which maps an instance x into a d-dimensional Euclidean space. Additionally, this embedding is constrained to live on the d-dimensional hypersphere, *i.e.* $||f(x)||_2 = 1$. Within the hypersphere, the distance between 2 projected instances is simply the Euclidean distance: $d(f(x_i), f(x_j)) = ||f(x_i) - f(x_j)||_2$

In this embedding space, we want the intra-class distances $d(f(x_i), f(x_j)), \forall x_i, x_j / y_i = y_j$ to be minimized and the inter-class distances $d(f(x_i), f(x_j)), \forall x_i, x_j / y_i \neq y_j$ to be maximized. A major advantage of embedding learning is that the projection f is class independent. At test time, we can expect examples from a different class, or identity, to still satisfy the embedding goals. This makes embedding learning suitable for verification and clustering tasks.

To achieve such embedding, one method is to learn the projection that optimizes the triplet loss in the embedding space. Unlike other losses such as verification loss [34], triplet loss encourages a relative distance constraint. A triplet consists of 3 data points: (x_a, x_p, x_n) such that $y_a = y_p$ and $y_a \neq y_n$ and thus, we would like the 2 points (x_a, x_p) to be close together and the 2 points (x_a, x_n) to be further away by a margin α in the embedding space [1]. Formally, a triplet must satisfy:

$$d(f(x_a), f(x_p)) + \alpha < d(f(x_a), f(x_n)), \forall (x_a, x_p, x_n) \in T \tag{1}$$

where T is the set of all possible triplets of the training set, and α is the margin enforced between the positive and negative pairs. By choosing $d << D$, one can learn a projection to a space that is both distinctive and compact. Subsequently, we define the loss to be minimized as:

$$\mathcal{L}(f) = \frac{1}{|T|} \sum_{(x_a, x_p, x_n) \in T} l(x_a, x_p, x_n, f) \tag{2}$$

in which

$$l(x_a, x_p, x_n; f) = \max\{0, d(f(x_a), f(x_p)) - d(f(x_a), f(x_n)) + \alpha\} \tag{3}$$

In spite of its advantages, the triplet loss training is empirical and depends on the training data, the initialization, and triplet sampling methods. For a certain set of training samples, there can be an exponential number of possible solutions that yield the same training loss. One approach to guarantee good performance is to make sure that the training data come from the same distribution of the test data (as in [26]). Another solution for the projection to work in more general unseen cases may be to gather a massive training dataset with more data (as FaceNet was trained with 100-200 millions images of 8 millions of identities [31]). Although it is possible to gather such a large scale dataset for visual information, it is less the case for acoustic data. This explains why speaker turn embedding *TristouNet* only gains slight improvement over Gaussian-based methods [4]. To alleviate the data concern, we tackle the problem of embedding learning from the multimodal point of view. By using a superior face embedding network that was trained on a face dataset with the same identities as in the acoustic dataset, we can regularize the speaker embedding space and thus guide the training process to a better minima.

4. Crossmodal transfer learning

In audio-visual (or multimodal data in general) settings, data contain 2 corresponding streams $\{(x_i^A, x_i^V, y_i)\}$. If the learning process is applied independently to each modality, we can learn 2 projections f_A and f_V into 2 embedding spaces $\mathbb{R}^{d_A}$ and $\mathbb{R}^{d_V}$ following their own respective losses:

$$\mathcal{L}^A(f^A) = \frac{1}{|T^A|} \sum_{(x_a^A, x_p^A, x_n^A) \in T^A} l(x_a^A, x_p^A, x_n^A; f^A) \tag{4}$$

[1] The value of α varies depending on the particular loss function to optimize We use one value of $\alpha = 0.2$ in all cases.

and

$$\mathcal{L}^V(f^V) = \frac{1}{|T^V|} \sum_{(x_a^V, x_p^V, x_n^V) \in T^V} l(x_a^V, x_p^V, x_n^V; f^V) \quad (5)$$

in which $\mathcal{L}^A$ and $\mathcal{L}^V$ are defined from the general embedding loss Eq. 2 to speaker turn embedding and face embedding.

As shown in the experiments, f^V can already achieve significantly lower than the counterpart in acoustic domain, therefore our goal is to transfer the knowledge from face embedding to the speaker turn embedding. Hence, we assume that f^V is already trained with Eq. 5 using the corresponding face dataset (as well as optional external data). Using f_V, an auxiliary term $\mathcal{L}^{V\to A}(f^A)$ is defined to regularize the relationship between voices and faces from the same identity in addition to the loss function used to train speaker turn embedding in Eq. 2. Formally, the final loss function can be written as:

$$\mathcal{L}(f^A) = \mathcal{L}^A(f^A) + \lambda \mathcal{L}^{V\to A}(f^A) \quad (6)$$

The transfer loss $\mathcal{L}^{V\to A}(f^A)$ depends on what type of knowledge is transferred across modalities. λ is a constant hyper-parameter chosen through experiments specifically for each transfer type. In the following sections, different types of $\mathcal{L}^{V\to A}(f^A)$ will be described in details.

4.1. Target embedding transfer

Assuming that f^A projects x_i^A into the same hypersphere as $f^V(x_j^V)$, one can observe that by enforcing $f^A(x_i^A)$ to be in close proximity of $f^V(x_j^V)$ when $y_i = y_j$, f^A could achieve a similar training loss as f^V. In that case, the regularizing term in Eq. 6 can be defined as the disparity between crossmodal instances of the same identity:

$$\mathcal{L}^{V\to A}(f^A) = \sum_{(x_i^A, x_j^V)/y_i = y_j} d(f^A(x_i^A), f^V(x_j^V)) \quad (7)$$

The goal of Eq. 7 is to minimize intra-class distances by binding embedded speaker turns and embedded faces within the same class similarly to coupled multimodal projection methods [21, 22]. In this work, we extend this goal further by adopting the multimodal triplet paradigm to jointly minimize intra-class distances and maximize inter-class distances.

Algorithm 1 Target embedding transfer triplet set.

```
Input $f^A$, $f^V$, $Q_{A,V}$, $\{(x_i^A, x_i^V, y_i)\}_{i=1..N}$
$T_{tar} = \emptyset$
for $\forall(a,p,n)/y_a = y_p \wedge y_a \neq y_n$ do
    for $m_a, m_p, m_n \in \{Q_{A,V}\}$ do
        $d_{a,p} = d(f^{m_a}(x_a^{m_a}), f^{m_p}(x_p^{m_p}))$
        $d_{a,n} = d(f^{m_a}(x_a^{m_a}), f^{m_n}(x_n^{m_n}))$
        if $d_{a,p} + \alpha > d_{a,n}$ then
            $T_{tar} = T_{tar} \cup (a,p,n)$
Output $T_{tar}$
```

Multimodal triplet loss. In addition to minimizing the audio triplet loss of Eq. 4, we also want two embedded instances to be close if they come from the same identity, regardless of the modality they comes from, and to be far from embedded instances of all other identities in both modalities as well. Concretely, the regularizing term is thus defined as the triplet loss over multimodal triplets:

$$\mathcal{L}^{V\to A}(f^A) = \frac{1}{|T_{tar}|} \sum_{(x_a^{m_a}, x_p^{m_p}, x_n^{m_n}) \in T_{tar}} l(x_a^{m_a}, x_p^{m_p}, x_n^{m_n}; f^A, f^V) \quad (8)$$

where $m_\bullet$ is the modality associated with the sample $x_\bullet^{m_\bullet}$, and the loss l is adapted from Eq. 3 by using the embedding appropriate to each sample modality. The set T_{tar} denotes all useful and valid cross-modal triplets, i.e. with the positive sample to be of the same identity of the anchor ($y_a = y_p$), and the negative sample to be from another identity ($y_a \neq y_n$); and with $(m_a, m_p, m_n) \in Q_{A,V}$, the set of valid modalities (all combinations except (V,V,V), (V,V,A), and (A,A,A) already considered in the primary loss of Eq. 4). For instance, if $(m_a, m_p, m_n) = (A,V,V)$, the loss will foster the decrease of the intra-class distance between $f^A(x_a^A)$ and $f^V(x_p^V)$ while increasing the inter-class distance between x_a^A and x_n^V. The strategy to collect the set T_{tar} at each epoch of the training is described in Alg. 1.

Using Eq. 8 as regularizing term in $\mathcal{L}(f^A)$, one can effectively use the embedded faces as targets to learn a speaker turn embedding. Note that this is similar in spirit to the neural network distillation [16], using one embedding as a teacher for the other. Moreover, the two modalities can be combined straightforwardly as their embedding spaces can be viewed as one harmonic space [31].

4.2. Clustering structure transfer

The common idea of the target transfer method is that people with similar faces should have similar voices. Thus it aims at putting constrains based on the distances among individual instances in the face embedding space. In clustering structure transfer, the central idea does not focus on pair of identities. but rather, we hypothesize that commonalities between 2 modalities can be discovered amongst groups of identities. For example, people within a similar age group are more likely to be close together in the face embedding space, and we also expect them to have more similar voices in comparison to other groups.

Based on this hypothesis, we propose to regularize the target speaker turn embedding space to have the same clustering structure with the source face embedding space, *i.e.* an identities should have the same neighbors in the speaker embedding space as in the face embedding space. To achieve that, we first discover groups in the face embedding space by performing a K-Means clustering on the set of mean identity representations $\{M_{y_i}^V\}$ by following 2 steps:

- Let X_{y_i} be the set of faces of identity y_i, we define the mean face representation M_{y_i} of person y_i as:

$$M_{y_i} = \frac{1}{|X_{y_i}|} \sum_{x_i \in X_{y_i}} f^V(x_i) \tag{9}$$

- K-Means is performed on the set of $\{M_{y_i}^V\}$. We denote by C the number of clusters, the resulting cluster mapping function is defined as:

$$\begin{aligned} g_m : \{1..K\} &\rightarrow \{1..C\} \\ y &\rightarrow c_y \end{aligned}$$

To define the regularizing term $\mathcal{L}^{V \rightarrow A}(f^A)$, we simply consider the set of cluster labels c_{y_i} attached to each audio sample (x_i^A, y_i) as the second label, and define accordingly a triplet loss relying on this second label (i.e by considering the instances (x_i^A, c_{y_i})). In this way, one can guide the acoustic instances of identities from the same cluster to be close together, thus preserving the source clustering structure. How to collect the set of triplet T_{str} to be used for the regularizing term at each epoch is detailed in Alg.2.

Algorithm 2 Clustering struct. transfer triplet set.

```
Input $f^A$, $f^V$, $g_m$, $\{(x_i^A, x_i^V, y_i)\}_{i=1..N}$
Cluster mapping $g_m$: $y \rightarrow c_y, \forall y \in 1 \dots K$
$T_{str} = \emptyset$
for $\forall(a, p, n) / c_{y_a} \neq c_{y_p} \wedge c_{y_a} \neq c_{y_n}$ do
    $d_{a,p} = d(f^A(x_a^A), f^A(x_p^A))$
    $d_{a,n} = d(f^A(x_a^A), f^A(x_n^A))$
    if $d_{a,p} + \alpha > d_{a,n}$ then
        $T_{str} = T_{str} \cup (a, p, n)$
Output $T_{str}$
```

This group structure can be expected to generalize for new identities because even though a person is unknown, he/she belongs to a certain group which share similarities in the face and voice domains. In our work, we only apply K-Means once on the mean facial representations. However, as people usually belong to multiple non-exclusive common groups, each with a different attribute, it would be interesting in further works to aggregate multiple clustering partitions with different initial seeds or with different number of clusters. As the space can be hierarchically structured, one other possibility could be to apply hierarchical clustering to obtain these multiple partitions.

Table 1. Statistics of tracks extracted from REPERE. The training and test sets have disjoint identities.

	# shows	# people	# tracks
training	98	208	1876
test	35	98	629

5. Experiments

We first describe the datasets and evaluation protocols before discussing the implementation details and the experimental results. Our codes and pretrained models are publicly available. [2]

5.1. Datasets

REPERE [12]. We use this standard dataset to collect people tracks with corresponding voice-face information. It features programs including news, debates, and talk shows from two French TV channels, LCP and BFMTV, along with annotations available through the REPERE challenge. The annotations consist of the timestamps when a person appears and talks. By intersecting the talking and appearing information, we can obtain all segments with face and voice from the same identity. As REPERE only contains sparse reference bounding box annotation, automatic face tracks [20] are aligned with reference bounding boxes to get the full face tracks. This collection process is followed by manual examination for correctness and consistency and to remove short tracks (less than 18 frames ≈ 0.72s). The resulting data is split into training and test sets. Statistics are shown in Tab. 1.

ETAPE [13]. This standard dataset contains 29 hours of TV broadcast. In this paper, we only consider the development set to compare with state-of-the-art methods. Specifically, we use similar settings for the "same/different" audio experiments than in [4]. From this development set, 5130 1-second segments of 58 identities are extracted. Because 15 identities appear in the REPERE training set, we remove them and retain 3746 segments of 43 identities.

[2] gitlab.idiap.ch/software/CTL-AV-Identification/

5.2. Experimental protocols and metrics

Same/different experiments. Given a set of segments, distances between all pairs are computed. One can then decide if a pair of instances has the same identity if their (embedded) distance is below a threshold. We can then report the equal error rate (EER), *i.e.* the value when the false negative rate and the false positive rate become equal as we vary the threshold.

Clustering experiments. From a set of all audio (or video) segments, a standard hierarchical clustering is applied using the distance between cluster means in the embedded space as merging criteria. Then, each time 2 clusters are merged, we compute 3 metrics on the clustering set:

- Weighted cluster purity (WCP) [32]: For a given set of clusters $C = \{c\}$, each cluster c has a weight of n_c, which is the number of segments within that cluster. At initialization, we start from N segments with weight 1 each. The purity $purity_c$ of a cluster c is the fraction of the largest number of segments from the same identities to the total number of segments in the cluster n_c, *i.e.* $WCP = \frac{1}{N}\sum_{c\in C} n_c \cdot purity_c$
- Weighted cluster entropy (WCE): A drawback from WCP is that it does not distinguish the errors. For instance, a cluster with 80% purity, 20% error due to 5 different identities is more severe than if it is only due to 2 identities. To characterize this point, we thus compute the entropy of a cluster, from which WCE is calculated as: $WCE = \frac{1}{N}\sum_{c\in C} n_c \cdot entropy_c$
- Operator clicks index (OCI-k) [14]: This is the total number of clicks required to label all clusters. If a cluster is 100% pure, only 1 click is required. Otherwise, besides 1 click to annotate segments of the dominant class, then 1 extra click is needed to correct each erroneous track of a different class. For a cluster c of n_c speaker segments, the cluster cost is formally defined as: OCI-k$(c) = 1 + (n_c - max(\{n_i^c\}))$, where n_i^c denotes the number of segments from identity i in the cluster. The cluster clicks are then added to produce the overall OCI-k measure. This metric simultaneously combines the number of clusters and cluster quality in one number to represent the manual effort in practical applications.

5.3. Implementation details

Face embedding. Our face model is based on ResNet-34 [15] trained on CASIA-WebFaces [34]. We follow the procedure of [26] as follows:

- A DPM face detector [10] is run to extract a tight bounding box around each face. No further preprocessing is performed except for randomly flipping training images.
- ResNet-34 is first trained to predict 10,575 identities by minimizing cross entropy criteria. Then the last layer is removed and the weights are frozen.
- The last embedding layer with a dimension of $d = 128$ is learned using Eq. 5 and the face tracks of the REPERE training set.

Speaker turn embedding. Our implementation of *TristouNet* consists of a bidirectional LSTM with the hidden size of 32. It is followed by an average pooling of the hidden state over the different time steps of the audio sequence, followed by 2 fully connected layers of size 64 and 128 respectively. As input acoustic features to the LSTM, 13 Mel-Frequency Cepstral Coefficients (MFCC) are extracted with energy and their first and second derivatives.

Optimization. All embedding networks are trained using a fixed $\alpha = 0.2$ and the RMSProp optimizer [33] with a 10^{-3} learning rate. From each mini-batch, both hard and soft negative triplets are used for learning.

Baselines. We compare our speaker turn embedding with 3 approaches: Bayesian Information Criterion (BIC) [6], Gaussian divergence (Div.) [1], and the original *TristouNet* [4].

5.4. Experimental results

5.4.1 Face embedding

We conducted this experiment to choose the best (more accurate) face embedding to transfer to the audio domain amongst the following candidates:

- VGG-Face: the model from [26], which was pretrained using 2.6M faces of 2622 identities.
- Rn34-FC: ResNet-34 trained with CASIA-WebFaces and using the activation of the last layer before the softmax classification as features.
- Rn34-Emb: the embedding layer learned using the face tracks of the REPERE dataset.

From the REPERE test set, 6000 pairs of tracks (3000 negative, 3000 positive) are selected for benchmarking the embeddings using the same/different experimental setting. We compare using the EER and the AUC of the ROC curve. From Tab. 2, we can see that the Rn34-FC slightly outperforms VGG-Face, and that further using a triplet loss learned using the face tracks of the REPERE data helps improving the results. Thus in the following experiments, Rn34-Emb is chosen as embedding to transfer to the audio domain.

Table 2. Results of face representations on 6000 pairs of REPERE test tracks.

	VGG-Face	Rn34-FC	Rn34-Emb
AUC - ROC	99.02	99.15	99.43
EER	4.35	3.6	3.15

Table 3. Result of OCI-k metric on the REPERE test set. 'Min' reports minimum value of OCI-k and its number of clusters. 'At ideal clusters' reports OCI-k at 98 clusters corresponding to 98 identities.

	Min (# clusters)	At 98 clusters
Rn34-Emb (V)	113 (113)	136
BIC [6]	451 (390)	525
Div. [1]	330 (289)	521
TristouNet [4]	275 (124)	285
Target	241 (123)	255
Structure	255 (132)	271

5.4.2 REPERE - Clustering experiment

We applied the audio (or video) hierarchical clustering to the 629 audio-visual test tracks of REPERE. Results are presented in Fig. 1. Face clustering with Rn34-Emb clearly outperforms all speaker turn based methods. This visual system is used as reference to show the significant difference between the two domains. At the beginning, Div. first merges longer audio segments with enough data so it achieves higher purity. However, as small segments get progressively merged, the performance of BIC and Div. quickly deteriorate due to the lack of good voice statistics.

Our transferring methods surpass *TristouNet* in both metrics, especially in the middle stages, when the distances between clusters becomes more confusing. This shows that the knowledge from the face embedding helps distinguishing confusing pairs of clusters. The gap in WCE also means that our embedding is also more consistent with respect to the inter-cluster distances. We should note that in WCP and WCE, segments count as one unit and are not weighted according to their duration as done in traditional diarization metrics. This is one reason while traditional approaches BIC and Div methods appear much worse with the clustering metrics. More experiments on full diarization are needed in future works.

Tab. 3 reports the number of clicks to label and correct the clustering results. Our target embedding transfer reduces the OCI-k by 30 from the closest competitor in both the best case and with the ideal number of clusters. This in practice can decrease the effort of human annotation by $10-12\%$. Clustering structure transfer method also shows improvement of 7-10%.

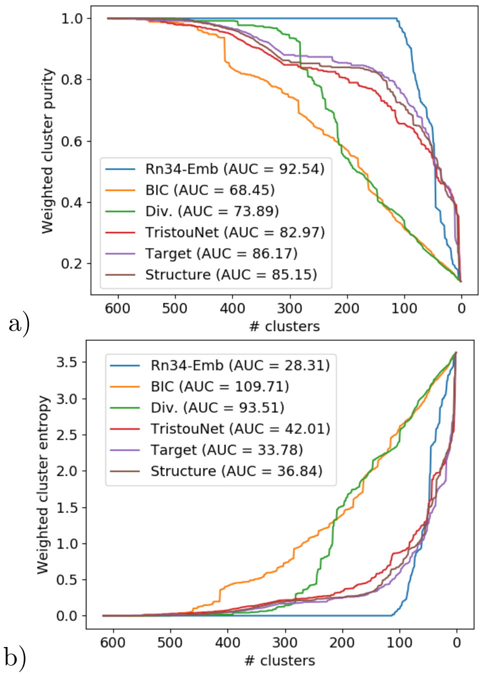

Figure 1. Evaluation of hierarchical clustering on REPERE. (a) weighted cluster purity. (b) weighted cluster entropy.

5.4.3 ETAPE - Same/different experiment

From the ETAPE development set, 3746 segments of 43 identities are extracted. From these segments, all possible pairs are used for testing and the EER is reported in Tab.4. All of our networks with transferred knowledge outperform the baselines. With short segments of 1 second, BIC and Div. do not have enough data to fit the Gaussian models well, therefore they perform poorly. By transferring from visual embedding, we can improve *TristouNet* with a relative improvement of 6% of EER. We should remark that in [4], the original *TristouNet* achieved 17.3% and 14.4% when being trained and tested on 1s sequences and 2s sequences respectively. However, it is important to note that our models are trained on a smaller dataset (4.5h vs. 13.8h of ETAPE data in [4]) and from an independent training set (REPERE vs. ETAPE). Using our transfer learning methods, the speaker turn embedding model could be easily trained by combining different dataset, *i.e.* combining REPERE and ETAPE training sets.

Comparison of transfer methods. Though the difference is small, target embedding shows an advantage in both the REPERE clustering experiments and in the ETAPE experiment. It seems that as the level of granularity decreases, the performance decreases. It could be interesting in future work to combine these different transfer method to see whether any further gain could be obtained.

Table 4. EER reported on ETAPE dev set. Note that our V → A transfer methods are trained on 1s. sequences (* denotes reported results from [4])

BIC[6]		Div.[1]		[4]	V → A transfer	
1s.	2s.*	1s.	2s.*	1s.	Target	Structure
32.4	20.5	28.9	22.5	19.1	18.0	18.3

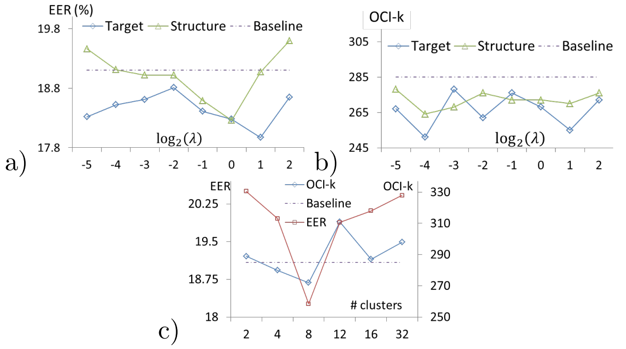

Figure 2. Result of different values of hyperparameters. (a)EER on ETAPE as λ changes, (b) OCI-k on REPERE as λ changes, (c) EER on ETAPE and OCI-k on REPERE as the number of clusters for structure transfer changes.

5.4.4 Parameter sensitivity

In all our transfer learning settings, we need to choose one hyper parameter λ, and the number of clusters for structure transfer setting. Hence, we perform benchmarking with different values of λ and report results in Fig. 2. In Fig. 2-(a) and (b), we can observe our methods are quite insensitive to this hyper parameter λ. Each of them has a different optimal value, which is due to the difference in the nature of each method. Fig.2-(c) shows how structure transfer performs under different granularity. Further analysis in the characteristics of clusters is presented in next subsection.

5.4.5 Further multimodal analysis

Cross modal retrieval. One interesting potential of target embedding transfer is the ability to connect a voice to a face of the same identity. To explore this aspect, we formulate a retrieval experiment: given 1 instance of the source embedding domain (voice or face), its distances to the embedding of 1 correct identities and 9 distractors in the enrolled domain are computed and ranked accordingly. There are 4 different settings depending on the within or cross domain retrieval: audio-audio, visual-visual, audio-visual, and visual-audio. Fig. 3-(a) shows the average precision of 980 different runs when choosing from the top 1 to 10 ranked results (Prec@K). Although the cross modal retrieval settings cannot compete with their single modality counterparts, they perform better than random chance and show consistency between the face embedding and speaker turn embedding. This proves that the two modalities cannot be coupled as in coupled matching learning but can be used as a regularizer of one another.

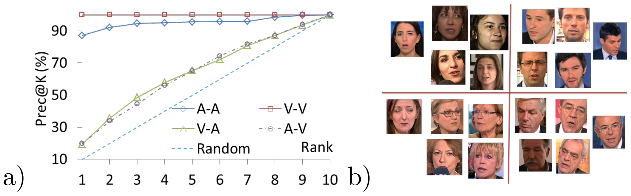

Figure 3. Analysis of different transferring type. (a) Prec@K of cross modal id retrieval using target transfer, (b) visualization of shared identities in 4 clusters across both modalities.

Shared clusters across modalities. Fig. 3-(b) visualizes 4 clusters which share the most common identities across the 2 modalities, when using the face embedding and the speaker embedding with structure transfer. One can observe 2 distinct characteristics among the clusters which are automatically captured: gender and age. It is noteworthy that these characteristics are discovered without any supervision.

6. Conclusion

We have proposed two different approaches to transfer knowledge from a source face embedding to a target speaker turn embedding. Each of our approaches explore different properties of the embedding spaces at different granularity. The results show that our methods improved speaker turn embedding in the tasks of verification and clustering. This is particularly significant in cases of short utterances, an important situation that can be found in many dialog cases, *e.g.* TV series, debates, or in multi-party human-robot interactions where backchannels and short answers/utterances are very frequent. The embedding spaces can also provide potential discovery of latent characteristics and a unified crossmodal combination. Another advantage of the transfer learning approaches is that each modality can be trained independently with their respective data, thus allowing future extension using advance learning techniques or more available data. In the future, experiments with more complicated tasks such as person diarization or large scale indexing can be performed to explore the possibilities of each proposal. Also, working with other corpora in different languages is an interesting direction.

Acknowledgement. We thank Hervé Bredin for useful discussion. This work was supported by the EU projects EUMSSI (FP7-611057) and MuMMER (H2020-688147).

References

[1] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006. 6, 7, 8

[2] M. Bendris, B. Favre, D. Charlet, G. Damnati, and R. Auguste. Multiple-view constrained clustering for unsupervised face identification in TV-broadcast. In *ICASSP)*. IEEE, 2014. 1, 2

[3] X. Bost and G. Linares. Constrained speaker diarization of TV series based on visual patterns. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014. 1

[4] H. Bredin. TristouNet: Triplet Loss for Speaker Turn Embedding. In *ICASSP*, New Orleans, USA, 2017. IEEE. 1, 2, 3, 6, 7, 8

[5] H. Bredin and G. Gelly. Improving speaker diarization of TV series using talking-face detection and clustering. In *Multimedia*. ACM, 2016. 1, 2

[6] S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA broadcast news transcription and understanding workshop*, 1998. 6, 7, 8

[7] P. Clément, T. Bazillon, and C. Fredouille. Speaker diarization of heterogeneous web video files: A preliminary study. In *ICASSP*. IEEE, 2011. 1

[8] D. Dai, T. Kroeger, R. Timofte, and L. Van Gool. Metric imitation by manifold transfer for efficient vision applications. In *CVPR*. IEEE, 2015. 3

[9] D. Dai and L. Van Gool. Unsupervised high-level feature learning by ensemble projection for semi-supervised image classification and image clustering. *arXiv preprint arXiv:1602.00955*, 2016. 3

[10] C. Dubout and F. Fleuret. Deformable part models with individual part scaling. In *BMVC*, 2013. 6

[11] P. Gay, E. Khoury, S. Meignier, J.-M. Odobez, and P. Deleglise. A Conditional Random Field approach for Audio-Visual people diarization. In *ICASSP*. IEEE, 2014. 1, 2

[12] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard. The REPERE corpus: a multimodal corpus for person recognition. In *LREC*, 2012. 5

[13] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert. The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC*, 2012. 5

[14] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*. IEEE, 2009. 6

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*. IEEE, 2016. 6

[16] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[17] D. Hu, X. Lu, and X. Li. Multimodal learning via exploring deep semantic similarity. In *ACM Multimedia*, 2016. 3

[18] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang. Deep multimodal speaker naming. In *ACM Multimedia*, 2015. 2

[19] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Esteve, and C. Jacquin. Automatic named identification of speakers using diarization and {ASR} systems. In *ICASSP*, 2009. 1

[20] N. Le, A. Heili, D. Wu, and J.-M. Odobez. Temporally subsampled detection for accurate and efficient face tracking and diarization. In *ICPR*. IEEE, 2016. 5

[21] A. Li, S. Shan, X. Chen, and W. Gao. Cross-pose face recognition based on partial least squares. *Pattern Recognition Letters*, 2011. 3, 4

[22] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia*, 2016. 3, 4

[23] M. Long, W. Cheng, X. Jin, J. Wang, and D. Shen. Transfer learning via cluster correspondence inference. In *ICDM*. IEEE, 2010. 2, 3

[24] C. Ma, P. Nguyen, and M. Mahajan. Finding speaker identities with a conditional maximum entropy model. In *ICASSP*, 2007. 1

[25] S. Moon, S. Kim, and H. Wang. Multimodal transfer deep learning with applications in audio-visual recognition. In *Multimodal Machine Learning Workshop at NIPS*, 2015. 2

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 1, 2, 3, 6

[27] J. Poignant, L. Besacier, and G. Quénot. Unsupervised Speaker Identification in {TV} Broadcast Based on Written Names. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2014. 1

[28] J. S. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan. Look, Listen and Learn - A Multimodal LSTM for Speaker Identification. In *AAAI*, 2016. 2

[29] G. Sargent, G. B. de Fonseca, I. L. Freire, R. Sicre, Z. Do Patrocínio Jr, S. Guimarães, and G. Gravier. PUC Minas and IRISA at multimodal person discovery. In *MediaEval*, 2016. 1, 2

[30] A. K. Sarkar, D. Matrouf, P.-M. Bousquet, and J.-F. Bonastre. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In *Interspeech*, 2012. 1

[31] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: a Unified Embedding for Face Recognition and Clustering. In *CVPR*, 2015. 1, 2, 3, 4

[32] M. Tapaswi, O. M. Parkhi, E. Rahtu, E. Sommerlade, R. Stiefelhagen, and A. Zisserman. Total cluster: A person agnostic clustering method for broadcast videos. In *Indian Conference on Computer Vision Graphics and Image Processing*. ACM, 2014. 6

[33] T. Tieleman and G. Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012. 6

[34] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2, 3, 6

[35] F. Zhuang, P. Luo, H. Xiong, Q. He, Y. Xiong, and Z. Shi. Exploiting associations between word clusters and document classes for cross-domain text categorization. *Statistical Analysis and Data Mining*, 2011. 2, 3