Noname manuscript No.
(will be inserted by the editor)

Improving speech embedding using crossmodal transfer learning with audio-visual data

Nam Le · Jean-Marc Odobez

Received: date / Accepted: date

Abstract Learning a discriminative voice embedding allows speaker turns to be compared directly and efficiently, which is crucial for tasks such as diarization and verification. This paper investigates several transfer learning approaches to improve a voice embedding using knowledge transferred from a face representation. The main idea of our crossmodal approaches is to constrain the target voice embedding space to share latent attributes with the source face embedding space.The shared latent attributes can be formalized as geometric properties or distribution characterics between these embedding spaces. We propose four transfer learning approaches belonging to two categories: the first category relies on the structure of the source face embedding space to regularize at different granularities the speaker turn embedding space. The second category -a domain adaptation approach- improves the embedding space of speaker turns by applying a maximum mean discrepancy loss to minimize the disparity between the distributions of the embedded features. Experiments are conducted on TV news datasets, REPERE and ETAPE, to demonstrate our methods. Quantitative results in verification and clustering tasks show promising improvement, especially in cases where speaker turns are short or the training data size is limited. The analysis also gives insights the embedding spaces and shows their potential applications.

Keywords Speaker diariazation, multimodal identification, metric learning, transfer learning, deep learning

Nam Le · Jean-Marc Odobez
Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédéral de Lausanne, Switzerland
E-mail: nle@idiap.ch, odobez@idiap.ch

1 Introduction

Learning speaker turn representation is the fundamental problem to enable comparing or clustering speech segments for multimedia indexing or interactive dialogue analysis. State-of-the-art Gaussian-based speaker diarization methods have been shown to be successful in various types of content such as radio, TV broadcast news, telephone conversation and meetings [34, 25, ?]. 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 [42]. As shown in TV series or movies, state-of-the-art approaches do not perform well [9, 5] 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, research has been conducted along two fronts: better methods to learn speaker turn embeddings or utilizing the multimodal nature of video content. For instance, the recent work on speaker turn embedding using triplet loss shows certain improvements [6, 12, 29], where as other multimodal related works focus on late fusion of two streams by propagating labels [3, 7] or high level information such as distances or overlapping duration [15, 41].

1.1 Motivation

In this work, we combine the two fronts of embedding learning and multimodal processing by investigating crossmodal transfer learning approaches to improve directly a speaker turn embedding using a face embedding . An overview of our framework is illustrated in Figure 1. First, on the visual side, we rely on the state-of-the-art advances in deep face embedding [37, 43]. Indeed recently, learning face embeddings has made significant achievements in all tasks, including recognition, verification, and clustering [30, 31, 43, 37, 49]. On the acoustic side, we exploit the deep architecture to learn a speaker turn embedding with triplet loss (*TristouNet*) of [6], which achieved improvement on short utterances. By projecting both acoustic signals and face images into a common hypersphere, one can unify the two embedding spaces, thus enabling the knowledge to be shared across modalities. The discrepancy between the two domains is formulated as an added regularizing term which measures differences between the two embedding spaces.

Our motivation for crossmodal transfer learning and adaptation is twofold. First, we can point to the difference in training data of two modalities. There are hundreds of thousands images from thousands identities in any standard face dataset. However, collecting labeled speech data is more challenging because we cannot use Internet search engines similarly to face images in [37, 48]. Also, manual labeling speech segments is much more costly or the labels

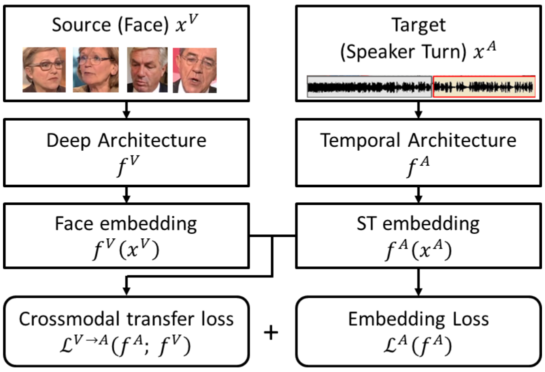

Fig. 1 Overview of our proposed method. Face embedding model is pretrained and used to guide the training of speaker turn embedding model through crossmodal transfer loss $\mathcal{L}^{V\rightarrow A}(f^A; f^V)$. Speaker turn embedding is trained with the combination of the embedding loss $\mathcal{L}^A(f^A)$ and the crossmodal transfer loss.

have to be obtained indirectly from visual modules[36]. Thus, we aim at mitigating the need for massive datasets and take advantage of pretrained face embeddings through transfer learning and domain adaptation.

Second, we can observe that although one cannot find the exact voice of a person given only a face, when given a small set of voice candidates, it is possible to pick a voice which is more likely to come from the given face than others. This means that there are shared commonalities between the two embedding spaces such as age, gender, or ethnicity; or in other words, if a group of people share common facial traits, we expect their voices to also share common acoustic features. Thus, there are latent attributes which are shared between the two modalities.

1.2 Our approach and main contributions

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 them to the target domain, a speaker turn embedding. We hypothesize that these latent attributes can be related to the geometry of the space or to the underlying distributions of features. Therefore, by transferring properties of the source embedding feature space (*i.e.* face embedding) onto the target embedding feature space (*i.e.* speaker turn embedding), we can improve the performance.

Because different properties can be used as constrains to be transferred, we investigate 4 different strategies continuing from our previous works [28, 27]. Out of these, 3 strategies aim at transferring spatial constraints of the embedding at different levels of granularity. Meanwhile, the fourth strategy focuses on the distributions of multimodal features. More precisely, they are:

- 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.
- Relative distance transfer: One can argue that exact similar location in the embedding spaces is hard to achieve given the fuzzy relationship between the 2 modalities. It may be sufficient to only enforce relative order between identities. Therefore, this approach constrains that 2 people who look more similar will have more similar voices.
- Clustering structure transfer: This approach focuses 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.
- Maximum mean discrepancy: This approach is different in nature than the previous ones, following the hypothesis that the crossmodal commonalities can be expressed as the discrepancy between the distributions of the two embedded features. In other words, by minimizing the difference between the distribution of speech features and the distribution of the visual features, one can achieve a better speech embedding. We use maximum mean discrepancy, which is the statistical measure of difference between 2 distributions [18], as the regularizing term.

Experiments conducted on 2 public datasets REPERE and ETAPE show significant improvement over the competitive baselines, especially when dealing with short utterances. Our results also show that by transferring knowledge from the visual domain, one can still learn competitive speaker turn embeddings when there are limited data. Our contributions are also supported by crossmodal retrieval experiments and the visualization of our intuition. It is important to note that while target embedding transfer, clustering structure transfer, and MMD approach were originally proposed in [28] and [27], relative distance transfer is a novel contribution of this paper to fully explore the granularity of embedding spaces.

The rest of the paper is organized as follows. Sec. 2 reviews other works related to ours, Sec. 3 introduces embedding learning with triplet loss and the architecture that we use in 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 [3]. Meanwhile in [7], 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. [41] linearly combines the acoustic distance and the face representation distance of speaking tracks to perform graph-based optimization; while [15] 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 has been proposed, including using deep neural networks [24,38]. However, it is only suitable for supervised tasks and has only been tested on limited datasets limited with 6 identities. 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 [37,43] and speaker turn [6] 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 three geometric-based approaches draw inspiration from a different line of research in transfer learning. First, we can point to coupled matching of image-text and heterogeneous recognition [32,23,**?**] or harmonic embedding [43] as related background for our target embedding transfer. These works focus on learning the direct mapping between an item in one domain to one item in the other. 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 a main loss to optimize. Meanwhile, as the learning target is an Euclidean embedding space in both modalities, relative distance transfer is inspired by metric imitation [10] or multi-tasks metric learning [4]. In our work, the triangular relationship is transferred across modalities instead of neighbourhood structure [10] or across tasks of the same modality [4]. Finally, as one identity is enforced to have the same neighbors in both face embedding and speech embedding spaces, our clustering structure transfer is therefore closely related to transfer learning through projection ensemble [11]. Although co-clustering information and cluster correspondence inference have also been used in transfer learning on traditional tasks of text mining [50,33], we are first to expand that concept into exploiting clustering structure of person identities for crossmodal learning.

Unlike the previous 3 transferring methods which emphasize on the geometric properties of the embedding spaces, the domain adaptation approach relies on the underlying distributions of the features within the embedding spaces.

A popular method in visual domain adaptation [14,40,1] is to minimize the maximum mean discrepancy (MMD) loss between the 2 feature sets. Maximum mean discrepancy (MMD) loss, proposed by [18], is a non-parametric approach to compare distributions. Intuitively, the 2 feature sets are projected into the kernel space and the distance between the means of the 2 distributions in this kernel space is used as the measurement of discrepancy. This is an interesting approach since non-parametric representations are well-suited for representing complex multimodal data in high-dimensional spaces. Our work is the first attempt in unifying the audio and visual domains into a single feature space which shares the commonalities by minimizing MMD loss between feature distributions in 2 embedding spaces.

3 Preliminary

This section first briefly recall the concept of learning embedding and the triplet loss. Then we review the TristouNet architecture, which is used as the main architecture for learning speech embedding in our work.

3.1 Triplet loss

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

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. 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 [48], 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. Thus, the triplet loss to train a projection f is defined as:

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

[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.

in which

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

Figure 2 shows an example of an embedding space, in which samples from difference classes are separated. By choosing $h << D$, one can learn a projection to a space that is both distinctive and compact.

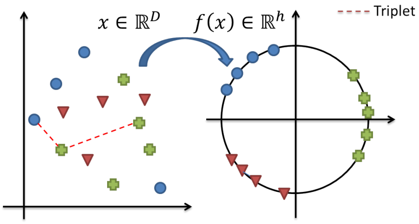

Fig. 2 Illustration of an embedding space. In this example, an embedding function f is learned to project the input samples in R^D into the embedding space R^h. In this embedding space, samples from the same class (with the same color) will have smaller distances than their distances with samples from a different class.

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.

3.2 Learning speaker turn embedding with triplet loss

The fundamental task we are interested in is to learn a good speaker turn embedding so that given 2 speaker segments, one can compare them directly for verification or clustering. To this end, one can employ different architectures with different input representations such as time delay neural networks [44] or convolutional neural networks on spectrograms [36]. In this work, we use the architecture proposed by [6], which learns speech emebedding using triplet loss, for compatibility with our visual models.

The TristouNet architecture is illustrated in Figure 3. Firstly, a bidirectional Long Short-Term Memory (LSTM) recurrent networks [22] receives the input sequence x^A to produce the hidden forward and backward outputs. Secondly, average pooling is then applied along the temporal axis of each output yielding 2 fixed length vectors. These 2 forward and backward vectors are then concatenated as input for the fully connected layers for projection into higher dimensional space. Finally, the output are L^2-normalized into the Euclidean hypersphere.

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

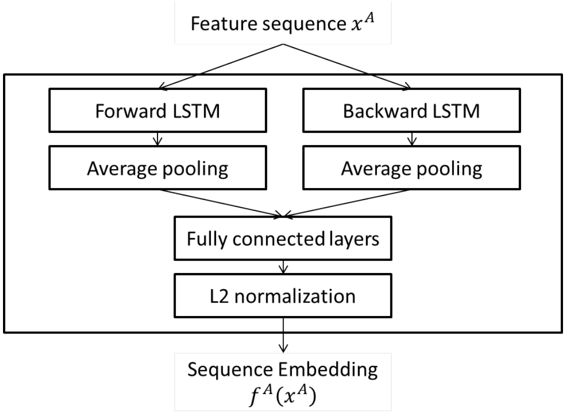

Fig. 3 TristouNet architecture

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 [37]). 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 [43]). 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 [6]. 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 this section, we will expand the embedding learning concept into multimodal data to learn different feature embedding spaces. Then we will in turn describe our transfer learning methods to use one embedding space to improve the other.

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:

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

and

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

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:

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

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:

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

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 [32, **?**]. 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.

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:

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

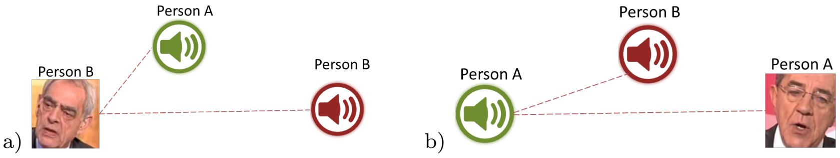

Fig. 4 Examples of multimodal triplets in target embedding transfer. Each triplet consists of mixing samples from both modalities. (a) (V, A, A) triplet where the anchor comes from visual domain. (b) (A, V, A) triplet where the positive comes from visual domain.

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 $T_{tar}$
```

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). In Figure 4, 2 examples of (V, A, A) and (A, V, A) are shown. 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 [21], 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 [43].

4.2 Relative distance transfer

The correspondence between faces and voices is not a definitive one-to-one, *i.e.* it is not trivial to precisely select the face corresponding to a voice one has heard. Therefore target embedding transfer might not generalize well even when achieving low training error. Instead of the exact locations, the relative distance transfer approach works at a lower granularity and aims to mimic the

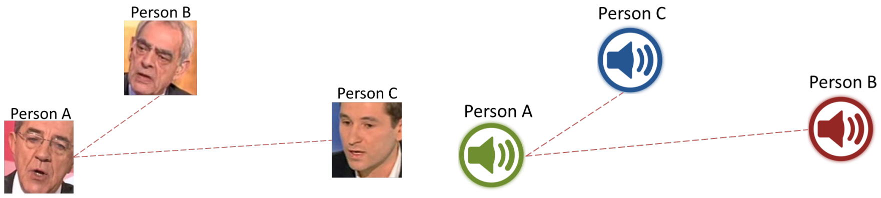

Fig. 5 An example of a transfer triplet in relative distance transfer. In the visual domain, $d(personA, personB) < d(personA, personC)$. However, this inequality is not satisfied in the audio domain. Therefore they form a negative triplet for training.

discriminative power (i.e. the notion of being close or far) of the face embedding space. Thus, it does not directly transfer the embeddings individual instances but the relative distances between their identities.

Before computing relative distances, let us define the mean face representation M_y of a person and the distance between identities within the face embedding space. Concretely, let X_{y_i} be the set of faces of identity y_i, the mean face representation M_{y_i} of person y_i is computed as:

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

where X_{y_i} is the set of visual samples with identity y_i. From $\{M_{y_i}\}$, we can define the distance between identities as:

$$d(y_i, y_j) = d(M_{y_i}, M_{y_j}), \tag{10}$$

The goal is then to collect in the set T_{rel} all audio triplets (a, p, n) *with arbitrary identities* where the sample p has an identity which is closer to the identity of the anchor sample a than the identity of the sample n, as defined in the face embedding. In other words, if within the face embedding space the relative distances among the 3 identities of the triplet (a, p, n) follows:

$$d(M^V_{y_a}, M^V_{y_p}) < d(M^V_{y_a}, M^V_{y_n}), \tag{11}$$

then this relative condition must hold in the speaker turn embedding space as well:

$$d(f^A(x^A_a), f^A(x^A_p)) + \alpha < d(f^A(x^A_a), f^A(x^A_n)) \tag{12}$$

Then, at each epoch, Eq. 11 and 12 can be used to collect the set T_{rel}, as shown in Alg. 2, and the regularizing transfer loss $\mathcal{L}^{V \to A}(f^A)$ can then be defined as the average sum of the standard triplet loss over this set. Figure 5 illustrates how such negative triplets are formed using Eq. 11 and 12. In theory, relative distance transfer can achieve the same training error as with target embedding transfer, but leave more freedom to the relaxation of the exact location of the embedded features.

Algorithm 2 Relative distance transfer triplet set.

```
1: Input $f^A$, $f^V$, $\{M^y\}_{y=1..K}$, $\{(x_i^A, x_i^V, y_i)\}_{i=1..N}$
2: $T_{rel} = \emptyset$
3: for $\forall(a, p, n)/y_a \neq y_p \wedge y_a \neq y_n$ do
4:     if $d(M_{y_a}^V, M_{y_p}^V) < d(M_{y_a}^V, M_{y_n}^V)$ then
5:         $d_{a,p} = d(f^A(x_a^A), f^A(x_p^A))$
6:         $d_{a,n} = d(f^A(x_a^A), f^A(x_n^A))$
7:         if $d_{a,p} + \alpha > d_{a,n}$ then
8:             $T_{rel} = T_{rel} \cup (a, p, n)$
9: Output $T_{rel}$
```

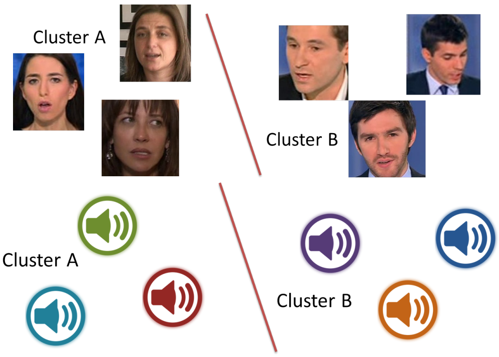

Fig. 6 In the visual domain, the identities form 2 clusters (ie. male vs female). We expect the samples in the audio domain to also from the same clustering structure. The audio embedding model is trained to not only discriminate between identities but also to form the same structure.

4.3 Clustering structure transfer

The common idea of the target transfer and relative distance transfer methods 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:

- The set of mean faces of each idendity $\{M^V_{y_i}\}$ is calculated following Eq. 9 in relative distance transfer.
- K-Means is performed on the set of $\{M^V_{y_i}\}$. 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^A_i, y_i) as the second label, and define accordingly a triplet loss relying on this second label (i.e by considering the instances (x^A_i, c_{y_i})). This step is illustrated in Figure 6, where the audio samples are assigned the cluster labels from the face domain. 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.3.

Algorithm 3 Clustering structure transfer triplet set.

```
Input $f^A$, $f^V$, $g_m$, $\{(x^A_i, x^V_i, y_i)\}_{i=1..N}$
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
    $d_{a,p} = d(f^A(x^A_a), f^A(x^A_p))$
    $d_{a,n} = d(f^A(x^A_a), f^A(x^A_n))$
    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.

4.4 Domain adaptation with maximal mean discrepancy

In the previous 3 methods, the emphasis was put on the geometric properties of the embedding spaces with respect to the labels. Therefore, the constrains between spaces are established only if we are given the multimodal correspondence between identities. Hence, these methods may not make full use of the face embedding, which was trained with more identities, and most of these are not present in the audio dataset. To overcome the dependence on labels, we focus on minimizing the difference between the two embedding feature distributions directly.

MMD is a statistical test to quantify the similarity between two distributions p and q on a domain $\mathcal{X}$ by mapping the data to a high dimensional feature space. The observations $X = x_1, ..., x_m$ and $Y = y_1, ..., y_n$ are drawn independently and identically distributed (i.i.d.) from p and q respectively.

To test whether $p = q$, we first introduce a class of function $\mathcal{F}$, which contains $f : \mathcal{X} \rightarrow \mathbb{R}$, each f can be simply viewed as a linear mapping function. Given $\mathcal{F}$, the measure of discrepancy between p and q can be estimated as:

$$\mathrm{MMD}[X, Y] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{j=1}^{n} f(y_j) \right) \tag{13}$$

By defining $\mathcal{F}$ as the set of functions in the unit ball in a universal Reproducing Kernel Hilbert Space (RKHS), it was shown that $\mathrm{MMD}[\mathcal{F}, X, Y] = 0$ if and only $ifp = q$ [18].

Let ϕ be the the mapping to the RKHS and $k(\cdot, \cdot) =< \phi(\cdot), \phi(\cdot) >$ be the universal kernel associated with this mapping. MMD can be computed as the distance between the mean of the two sets after mapping each sample to the RKHS:

$$\begin{aligned} \mathrm{MMD}^2(X, Y) &= \left|\left| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(y_j) \right|\right|^2 \qquad (14) \\ &= \sum_{i,j=1}^{m} \frac{k(x_i, x_j)}{m^2} - 2 \sum_{i,j=1}^{m,n} \frac{k(x_i, y_j)}{mn} + \sum_{i,j=1}^{n} \frac{k(y_i, y_j)}{n^2} \end{aligned}$$

The MMD between the distributions of two sets of observations is equivalent to the distance between the sample means in a high-dimensional feature space. In practice, Gaussian or Laplace kernels are often chosen as they are shown to be universal [45]. We choose kernel k associated to ϕ to be Radial Basis Function kernel, *i.e.* $k(u, v) = \exp(-d(u, v)^2)/\sigma$ in Eq. 14.

Originally proposed as a statistic measure between 2 distributions, MMD is widely used as the loss for domain adaptation[14,40,1]. Let x_s be the samples from the source domain, x_t be the samples from the target domain, and f^s, f^t be their respective feature mapping functions. By minimizing $\mathrm{MMD}(\{f^s(x^s)\}, \{f^t(x^t)\})$, one can minimize the discrepancy between the feature spaces learned from the two domains, thus enhancing the performance on the target domain using the knowledge from the source domain. In our work, we adopted the same strategy after unifying the two embedding spaces of faces and speaker turns respectively.

Given that the two embedding spaces can be constrained to lie within the same hypersphere, one can measure the discrepancy between the distributions

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

	# shows	# people	# tracks
training	98	208	3360
test	35	98	629

of face embedded features $f^V(x_i^V)$ and auditory embedded features $f^A(x_j^A)$ using Eq. 14 as:

$$\mathcal{L}^{V \to A}(f^A) = \mathrm{MMD}(\{f^V(x_i^V)\}, \{f^A(x_j^A)\}) \tag{15}$$

Based on Eq. 15, our objective is to find an embedding which is capable of inferring cross-domain statistical relationships when one exists. Instead of trying to bind faces and voices of the same individual identity geometrically, minimizing Eq. 15 only regulates the statistical properties of the whole population in an unsupervised fashion. Intuitively, minimizing the MMD forces the auditory features to have the same distribution with facial features, which includes having similar density around common attributes or identites. This can be interpreted as a regularizing term in $\mathcal{L}(f^A)$ to effectively use the embedded faces to guide the speaker turn embedding.

5 Experiments

We first describe the datasets and evaluation protocols before discussing the implementation details and the experimental results.

5.1 Datasets

REPERE [16]. 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 [26] 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 Table 1.

ETAPE [17]. 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 [6]. 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.

5.2 Experimental protocols and metrics

The experiments are designed to benchmarking the quality of the embedding space improved by transfer learning. The same/different experiments are designed following the verification protocol, which is based on assessing distances between pairs of samples. Meanwhile the clustering experiments are designed to quantify if the embedding space is discriminative enough to group segments of each identity among other candidates.

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. At each step, 2 clusters with the minimum distance are merged and a new mean is computed. For every step, we compute 3 metrics on the clustering set:

- Weighted cluster purity (WCP) [46]: 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 a weight of 1 for each segment. 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. We can define WCP as:

$$WCP = \frac{1}{N} \sum_{c \in C} n_c \cdot purity_c$$

- 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) [19]: 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 the 1 click needed to annotate segments of the dominant class, to correct each erroneous track of a different class, 1 more click will be added. For a cluster c of n_c speaker segments, the cluster cost is formally defined as:

$$\text{OCI-k}(c) = 1 + (n_c - max(\{n_i^c\})),$$

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 for correctly annotating all speaker segments given an initial clustering.

5.3 Implementation details

Face embedding. Our face model is based on the ResNet-34 architecture [20] trained on the CASIA-WebFaces dataset [48]. This is a collection of 494,414 images from 10,575 identities. We follow the procedure of [37] as follows:

- A DPM face detector [13] 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 $h = 128$ 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 [47] 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) [8], Gaussian divergence (Div.) [2], and the original *TristouNet* [6].

Additional details. Our codes use PyTorch library and are publicly available at:
`gitlab.idiap.ch/software/CTL-AV-Identification/`

5.4 Experimental results

5.4.1 REPERE - Clustering experiment

We applied the audio (or video) hierarchical clustering to the 629 audio-visual test tracks of REPERE. In Figure 7-a, we can compare other methods with the

Table 2 Result of OCI-k metric on the REPERE test set. 'Min' reports minimum value of OCI-k and in parenthesis is the number of clusters where this is achieved. 'At ideal clusters' reports OCI-k obtained when clustering reaches clusters the ideal number of clusters corresponding to 98 identities.

Methods		Min (# clusters)	At 98 clusters
Vision	Rn34-Emb (V)	113 (113)	136
Audio	BIC [8]	451 (390)	525
	Div. [2]	330 (289)	521
	TristouNet [6]	216 (119)	226
Audio-Visual (Ours)	Target	214 (112)	228
	Relative	204 (99)	211
	Structure	207 (107)	216
	MMD	202 (94)	209

two reference systems: Tristounet for speech embedding and ResNet-34 (Rn34-Emb)for face embedding. Face clustering with Rn34-Emb clearly outperforms all speaker turn based methods. This visual system is used as a reference to show the significant difference between the two domains, thus motivating transfer learning to improve the speech embedding. 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 robust with respect to the inter-cluster distances. Overall, all transfer learning methods are consistent and show steady improvements. On average, MMD has a slightly better result than the others. For the geometry-based methods, the lower the granularity level, the higher the gain in performance is. This shows that due to low the correlation between the 2 domains, it is better to use the latent attributes in the data with some relaxation rather than directly enforcing the embedding targets to be the same for both audio and video domain.

Figure 7-b, we compare our best method, MMD, with other state of the art methods. 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. 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 why traditional approaches BIC and Div methods appear much worse with the clustering metrics. More experiments on full diarization are needed in future works.

Table 2 reports the number of clicks to label and correct the clustering results. Our MMD approach reduces the OCI-k by 17 from the closest competitor in both the best cases at the minimum OCI-k and at the ideal number of clusters. This in practice can decrease the effort of human annotation by around 7.5%. While target embedding transfer does not yield any improvement, clustering structure and relative distance transferring methods also show decreases of $4.5 - 6.5\%$.

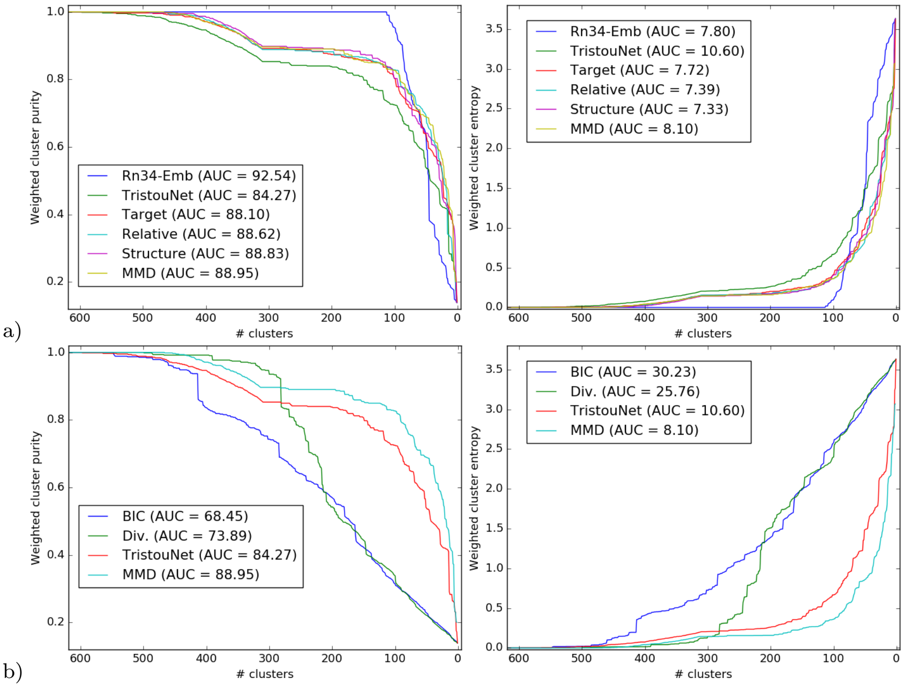

Fig. 7 Weighted cluster purity (WCP) and weighted cluster entropy (WCE) evaluation of hierarchical clustering on REPERE. (a) Comparison of our transferring approaches against the baseline TristouNet and the face embedding using ResNet-34 (b) Comparison of our MMD approach against state-of-art audio systems.

5.4.2 ETAPE - biometrics 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 Table 3. Most 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 using MMD, we can improve *TristouNet* with a relative improvement of 3.7% of EER. We should remark that in [6], the original *TristouNet* achieved 17.3% and 14.4% when being trained and tested on 1s sequences and 2s sequences respectively. It is important to note that our models are trained on a smaller dataset (8h vs. 13.8h of ETAPE data in [6]) 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 datasets, *i.e.* combining REPERE and ETAPE training sets.

Table 3 EER reported on all pairs of 3746 sequences in ETAPE dev set.

	BIC[8]	Div.[2]	*TristouNet*[6]	Target	Relative	Structure	MMD
EER	32.4	28.9	16.1	16.30	15.59	15.62	15.5

Table 4 Performance when training data are limited. EER is reported on ETAPE dev set. OCI-k is reported on REPERE.

	60%					30%				
	[6]	Tar.	Rel.	Str.	MMD	[6]	Tar.	Rel.	Str.	MMD
Min OCI-k	274	241	256	255	229	249	232	218	225	213
OCI-k@98	285	255	268	271	231	263	250	242	229	221
EER	19.1	18.0	18.2	18.3	18.4	16.9	16.7	16.4	15.9	16.5

5.4.3 Results with limited data.

To benchmark the generalization of our approaches, the same verification and clustering protocols fromp revious subsections are applied when the amount of training audio data is reduced and the results are reported in Table 4. Transferring methods perform particularly better in this scenario. In most cases, networks trained with MMD loss achieves better figures. As the amount of training data decreases, the performance of the audio-only system quickly deteriorates, especially in the clustering protocol. On the other hand, our visual guided system is less affected. When using only 60% of data, MMD outperforms audio-only *TristouNet* in OCI-k by 45 points , *i.e.* reducing the manually effort by 16%. Interestingly, both systems perform better with 30% of data than with 60%. One explanation is that although there are fewer samples, they are more balanced among identities. Considering both metrics, this balance in identities helps MMD the most because it is a density-based method. Therefore, the imbalance in the dataset can leads to a skew in the distribution and reduce the effectiveness of MMD. Meanwhile, as target transfer is a sample-based method, the balance in the 30% set does not improve as significantly.

5.4.4 Combining with a different embedding regularizer

In this experiment, we explore how our crossmodal losses can be combined with the recent intra-class loss [29]. Intra-class loss is a soft constraint on the averaged pair-wise distance between samples from the same class. It is also a regularizer preventing the scattering of these samples within the embedding space to increase the intra-class compactness.

In Table 5, we first present the results of [29] using the same benchmarks with 30% of the training dataset. Comparing to our methods in Table 4, it achieves comparable results with MMD. Subsequently, intra-class loss is linearly combined with all of our crossmodal losses to yield the rest of Table 5.

One can first see that combining target transfer and intra-class loss does not give any improvements. This can be explained as the face embedding has a different intra-class structure. Hence, forcing the audio embedding to

Table 5 Performance when combining crossmodal regularizers and intra-class loss. EER is reported on ETAPE dev set. OCI-k is reported on REPERE.

	[29]	[29] + Tar.	[29] + Rel.	[29] + Str.	[29] + MMD
Min OCI-k	214	231	226	216	209
OCI-k@98	221	244	240	228	223
EER	16.3	16.7	15.8	16.0	16.3

match both this structure and another explicit intra-class constrain can be conflicting as both losses work on the sample-level granularity. Meanwhile, combining intra-class loss with relative distance transfer and structure transfer can slightly improve EER because these geometric properties are at inter-class level, ie. relationship across identities, and are complementary with intra-class compactness. Finally, MMD concerns with sample density in an unsupervised manner, it is only slightly beneficial in clustering metrics to use intra-class loss with MMD.

5.4.5 Parameter sensitivity analysis

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 varying as power of 2. This experiment is performed on the training set with 60% of data for computational reason and results are reported in Figure 8. All of our methods are insensitive to λ except for relative distance transfer. Target embedding transfer is the most stable one, as its constrain is more specific than that of the rest. Each of them has a different optimal value, which is due to the difference in the nature of each method. One possible explanation for the case of relative distance transfer (Figure 8-b) when $\lambda \geq 2$ is that there is no proximity constrains on the location of the embedded features, thus instability is not bounded and can increase at test time. Figure8-(e) shows how structure transfer performs under different granularity. Further analysis in the characteristics of clusters is presented in next subsection.

5.4.6 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 instance with the same identity and to 9 distractors in the enrolled domain are computed and ranked accordingly. This experiment is similar to that of crossmodal biometrics [39,35]. There are 4 different settings depending on the within or cross domain retrieval: audio-audio, visual-visual, audio-visual, and visual-audio. Figure 9-(a) shows the average precision of 980 different

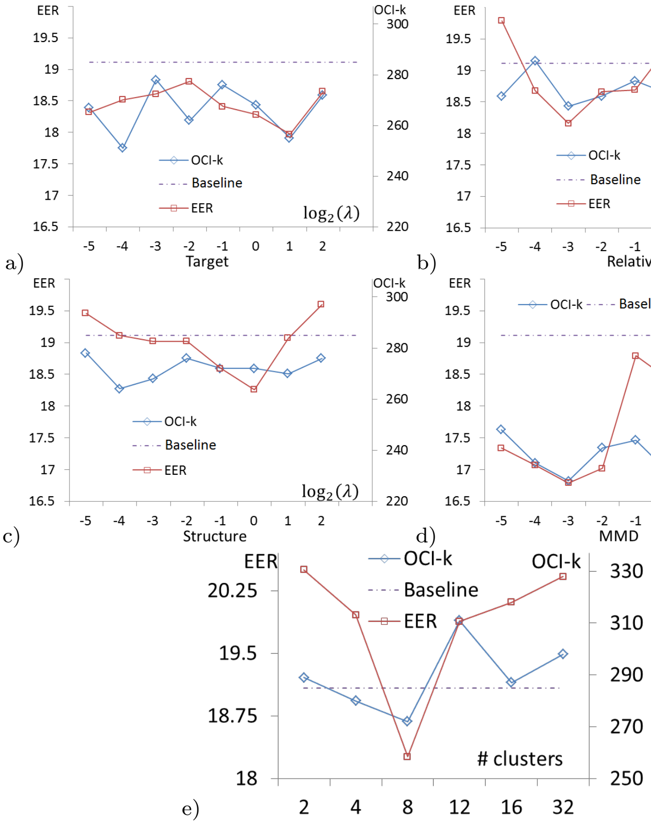

Fig. 8 Result of different values of hyperparameters. The baseline is EER and OCI-k of the standard Tristounet. (a-d) EER on ETAPE and OCI-k on REPERE of target, relative, structure, and MMD respectively as λ changes. (e) EER on ETAPE and OCI-k on REPERE as the number of clusters for structure transfer changes.

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 shows that though the two modalities cannot be used as in coupled matching learning, they can be used as a regularizer of one another.

Shared clusters across modalities. Figure 9-(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.

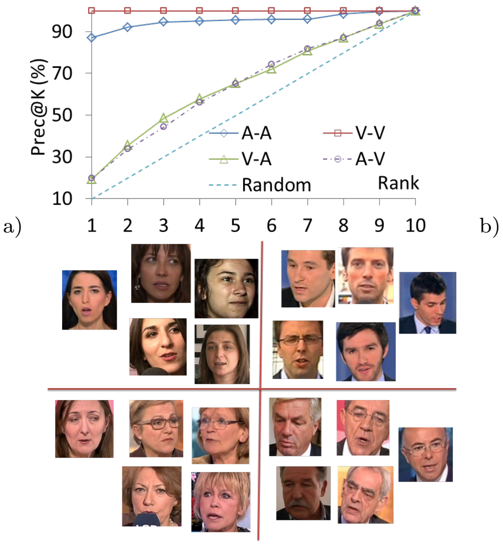

Fig. 9 Analysis of different transferring type. (a) Prec@K of cross modal id retrieval using target transfer, (b) visualization of shared identities in 4 clusters across both modalities.

6 Conclusion

In this paper, we have investigated 4 different methods to transfer knowledge from a source face embedding to a target speaker turn embedding. Inspired by state-of-the-art machine learning literature, each of our approaches explore different properties of the embedding spaces. Depending on the properties exploited, our methods can be categorized into two groups. The first group uses the geometric features at different levels of granularity; *i.e.* through direct target, relative distances, or neighborhood regularization. Meanwhile the second group uses the regularization of the underlying common feature distribution; *i.e.* through regularizing maximum mean discrepancy.

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. It is also interesting to use the face embedding guidance to expand the speech identification data.

Acknowledgement. This work was supported by the EU projects EUMSSI -

Event Understanding through Multimodal Social Stream Interpretation (FP7-611057) and MuMMER - MultiModal Mall Entertainment Robot (H2020-688147).

References

1. M. Baktashmotlagh, M. Harandi, and M. Salzmann. Distribution-matching embedding for visual domain adaptation. *The Journal of Machine Learning Research*, 17(1):3760–3789, 2016.
2. C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1505–1512, 2006.
3. M. Bendris, B. Favre, D. Charlet, G. Damnati, and R. Auguste. Multiple-view constrained clustering for unsupervised face identification in TV-broadcast. In *ICASSP)*, pages 494–498. IEEE, 2014.
4. B. Bhattarai, G. Sharma, and F. Jurie. CP-mtML: Coupled projection multi-task metric learning for large scale face retrieval. In *CVPR*. IEEE, 2016.
5. 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.
6. H. Bredin. TristouNet: Triplet Loss for Speaker Turn Embedding. In *ICASSP*, New Orleans, USA, 2017. IEEE.
7. H. Bredin and G. Gelly. Improving speaker diarization of TV series using talking-face detection and clustering. In *ACM Multimedia*, pages 157–161. ACM, 2016.
8. 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.
9. P. Clément, T. Bazillon, and C. Fredouille. Speaker diarization of heterogeneous web video files: A preliminary study. In *ICASSP*. IEEE, 2011.
10. D. Dai, T. Kroeger, R. Timofte, and L. Van Gool. Metric imitation by manifold transfer for efficient vision applications. In *CVPR*. IEEE, 2015.
11. 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.
12. S. Dey, S. Madikeri, and P. Motlicek. End-to-end text-dependent speaker verification using novel distance measures. *Proc. Interspeech 2018*, pages 3598–3602, 2018.
13. C. Dubout and F. Fleuret. Deformable part models with individual part scaling. In *BMVC*, 2013.
14. Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.
15. 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.
16. 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.
17. 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.
18. A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. *NIPS*, 2007.
19. M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*. IEEE, 2009.
20. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*. IEEE, 2016.
21. G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
22. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

23. D. Hu, X. Lu, and X. Li. Multimodal learning via exploring deep semantic similarity. In *ACM Multimedia*, 2016.
24. Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang. Deep multimodal speaker naming. In *ACM Multimedia*, 2015.
25. 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.
26. 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.
27. N. Le and J.-M. Odobez. A domain adaptation approach to improve speaker turn embedding using face representation. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 411–415. ACM, 2017.
28. N. Le and J.-M. Odobez. Improving speaker turn embedding by crossmodal transfer learning from face embedding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 428–437, 2017.
29. N. Le and J.-M. Odobez. Robust and discriminative speaker embedding via intra-class distance variance regularization. *Proc. Interspeech 2018*, pages 2257–2261, 2018.
30. L. Leng, J. Zhang, G. Chen, M. K. Khan, and K. Alghathbar. Two-directional two-dimensional random projection and its variations for face and palmprint recognition. In *International Conference on Computational Science and Its Applications*, pages 458–470. Springer, 2011.
31. L. Leng, J. Zhang, J. Xu, M. K. Khan, and K. Alghathbar. Dynamic weighted discrimination power analysis in dct domain for face and palmprint recognition. In *Information and Communication Technology Convergence (ICTC), 2010 International Conference on*, pages 467–471. IEEE, 2010.
32. A. Li, S. Shan, X. Chen, and W. Gao. Cross-pose face recognition based on partial least squares. *Pattern Recognition Letters*, 32(15):1948–1955, 2011.
33. M. Long, W. Cheng, X. Jin, J. Wang, and D. Shen. Transfer learning via cluster correspondence inference. In *ICDM*. IEEE, 2010.
34. C. Ma, P. Nguyen, and M. Mahajan. Finding speaker identities with a conditional maximum entropy model. In *ICASSP*, 2007.
35. A. Nagrani, S. Albanie, and A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
36. A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.
37. O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.
38. 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.
39. A. Roy and S. Marcel. Introducing crossmodal biometrics: Person identification from distinct audio & visual streams. In *BTAS*. IEEE, 2010.
40. A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. *arXiv preprint arXiv:1603.06432*, 2016.
41. 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 Workshop*, 2016.
42. 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.
43. F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: a Unified Embedding for Face Recognition and Clustering. In *CVPR*, 2015.
44. D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. Deep neural network embeddings for text-independent speaker verification. *INTERSPEECH*, 2017.
45. I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2001.
46. 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.

47. 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):26–31, 2012.
48. D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
49. Y. Zheng, D. K. Pal, and M. Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5097, 2018.
50. 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: The ASA Data Science Journal*, 4(1):100–114, 2011.