A Domain Adaptation Approach to Improve Speaker Turn Embedding Using Face Representation

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

ABSTRACT

This paper proposes a novel approach to improve speaker modeling using knowledge transferred from face representation. In particular, we are interested in learning a discriminative metric which allows speaker turns to be compared directly, which is beneficial for tasks such as diarization and dialogue analysis. Our method improves the embedding space of speaker turns by applying maximum mean discrepancy loss to minimize the disparity between the distributions of facial and acoustic embedded features. This approach aims to discover the shared underlying structure of the two embedded spaces, thus enabling the transfer of knowledge from the richer face representation to the counterpart in speech. Experiments are conducted on broadcast TV news datasets, REPERE and ETAPE, to demonstrate the validity of our method. 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.

CCS CONCEPTS

• Information systems → Video search; **• Computing methodologies →** *Transfer learning*;

KEYWORDS

Multimodal person diarization, domain adaptation, metric learning

ACM Reference format:
Nam Le, Jean-Marc Odobez Idiap Research Institute, Martigny, Switzerland École Polytechnique Fédéral de Lausanne, Switzerland . 2017. A Domain Adaptation Approach to Improve Speaker Turn Embedding Using Face Representation. In *Proceedings of 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, November 13–17, 2017 (ICMI'17),* 5 pages.
https://doi.org/10.1145/3136755.3136800

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 methods such as Bayesian Information Criterion (BIC) [8] or Gaussian divergence [2] have been shown to be successful in contents where the speech signal is mostly prepared and clean, the number of speakers is limited, and the duration of speaker turn is more than 2 seconds on average [19, 21, 25]. When these conditions are not valid, in particular the assumption of speaker turn duration, the quality of speaker diarization deteriorates [29], which was the case in TV series or movies [4, 9] or in human-robot interactions where backchannels and short answers/utterances are very frequent. In this paper, we propose an approach to improve speaker turn representation by combining recent deep metric learning advances [5] with the adaptation of the face embedding domain.

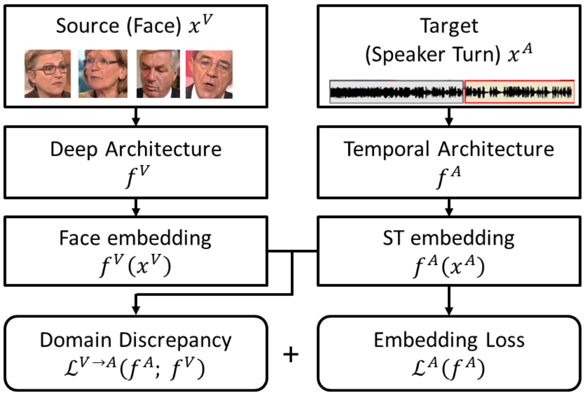
Source (Face) x^V
Target (Speaker Turn) x^A
Deep Architecture f^V
Temporal Architecture f^A
Face embedding $f^V(x^V)$
ST embedding $f^A(x^A)$
Domain Discrepancy $\mathcal{L}^{V\to A}(f^A;\, f^V)$
+
Embedding Loss $\mathcal{L}^A(f^A)$

Figure 1: Overview of our proposed method. Face embedding model is pretrained and used to guide the training of speaker turn embedding model through domain adaptation.

As person analysis in multimedia content is multimodal by nature, significant effort has been devoted to using one modality to improve another. In person diarization, one can either use labels from the modality with superior performance to correct the other [3, 6] or perform audio-visual clustering jointly [12, 28] For speaker identification, early fusion of features is proposed in [18, 26] but it is only suitable for supervised tasks with small dataset. Other works in audio-visual combination include co-training in active speaker detection [7] or transfer learning in speech recognition [22].

In our work, we emphasize on improving the embedding space directly using the knowledge of the other modality through domain adaptation, in contrast to the aforementioned works focusing on aggregating two streams of information. Although domain adaptation is applied on a wide range of computer vision tasks [11, 27], to our best knowledge we are the first to apply it across acoustic and visual domains. An overview of our framework is illustrated in Fig. 1. First, we rely on the state-of-the-art advances in deep face embedding [24, 30]. Indeed recently, learning face embeddings has made significant achievements in all tasks, including recognition, verification, and clustering [24, 30]. On the acoustic domain,

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136800

we employ the speaker turn embedding trained with triplet loss (*TristouNet*) in [5], 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 the differences between two distributions of embedding features. Following works in visual domain adaptation [1, 11, 27], we opt for maximum mean discrepancy (MMD) loss [15], which is a non-parametric approach to compare distributions, as non-parametric representations are well-suited for complex multimodal data in high-dimensional spaces.

Our motivation for crossmodal 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 very challenging because we cannot use Internet search engines similarly to face images in [24, 34] and manual labeling speech segments is much more costly. Thus, we aim at mitigating the need for massive datasets and take advantage of pretrained face embeddings through domain adaptation.

Second, we can observe that although one cannot find the exact voice of a person given only a face, however, if given a small set of candidates, it is possible to pick a voice which is more likely to come from the given face than other voices. This means that there are shared commonalities between the two embedding spaces such as age, gender, or ethnicity; *i.e.* if a group of people share common facial traits, we expect their voices to also share common acoustic features. We hypothesize that these commonalities can be enforced through the distributions of the two embedded features. Thus by creating a regularizing term so that the distribution of speaker turn embedded features is as close as that of face embedded features, we can improve the performance of the speaker turn embedding.

We demonstrate the method through experiments conducted on 2 datasets REPERE and ETAPE. The results show significant improvement over the competitive baselines in the tasks of verification and clustering, especially when dealing with short utterances.

2 PRELIMINARIES

2.1 Embedding learning with triplet loss

Given a labeled training set $\{(x_i, y_i)\}$, in which $x_i \in \mathbb{R}^D, y_i \in \{1, 2, .., K\}$, embedding learning is a class of algorithms which learn a function $f(x) \in \mathbb{R}^h$ which maps an instance x into a h-dimensional space. In this new embedding space, we want the intra-class distances $d(f(x_i), f(x_j))/y_i = y_j$ to be minimized and the inter-class distances $d(f(x_i), f(x_j))/y_i \neq y_j$ to be maximized. By choosing $h << D$, one can learn a projection to a space that is both distinctive and compact. A major advantage of embedding learning is that the projection f is class independent. Hence, embedding learning is suitable for both supervised and unsupervised tasks.

To achieve such embedding, one method is to learn the projection that optimizes the triplet loss in the embedding space. A triplet consists of 3 data points: anchor point x_a, positive point x_p, and negative point x_n such that $y_a = y_p$ and $y_a \neq y_n$. Following the embedding goal, 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. Formally, a triplet must satisfy:

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

where T is the set of all possible triplets of the training set, α is the margin enforced between the positive and negative pairs, and d is the Euclidean distance in the embedding space. Subsequently, we define the loss to be minimized as:

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

in which

$$l(x_a, x_p, x_n; f) = [d(f(x_a), f(x_p)) - d(f(x_a), f(x_n)) + \alpha]_+ \quad (3)$$

In spite of its advantages, the triplet loss training is empirical and depends on the training data, the initialization, and sampling methods. To guarantee good performance, one needs to make sure that training data come from the same distribution of the test data [24] or to train on a massive training dataset [30]. Hence, we tackle the problem from the multimodal point of view by using a superior face embedding network to regularize the speaker embedding space, thus guiding the training process to a better minima.

2.2 Maximum mean discrepancy (MMD)

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. From $\mathcal{F}$, the measure of discrepancy between p and q are be estimated as:

$$\text{MMD}[X, Y] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{j=1}^{n} f(y_j) \right) \quad (4)$$

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 $\text{MMD}[\mathcal{F}, X, Y] = 0$ if and only $if p = q$ [15].

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:

$$\begin{aligned} \text{MMD}^2[X, Y] &= \left\| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(y_j) \right\|^2 \quad (5) \\ &= \sum_{i,j=1}^{m} \frac{k(x_i, x_j)}{m^2} - 2 \sum_{i,j=1}^{m,n} \frac{k(x_i, y_j)}{mn} + \sum_{i,j=1}^{n} \frac{k(y_i, y_j)}{n^2} \end{aligned}$$

Intuitively, 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 [31].

Originally proposed as a statistic measure between 2 distributions, MMD is widely used as the loss for domain adaptation[1, 11, 27]. 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 $\text{MMD}(f^s(x^s), f^t(x^t))$, 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.

3 LEARNING SPEAKER TURN EMBEDDING WITH CROSSMODAL ADAPTATION

In audio-visual (or any multimodal data in general) settings, data contain 2 corresponding streams $\{(x_i^A, x_i^V, y_i)\}$. Our goal is to use the visual stream x_i^V and its embedding function f^V to assist learning the embedding function of the auditory stream x_i^A. The overall scheme is illustrated in Fig. 1.

First, as shown in the experiments as well as in literature, the face embedding f^V can achieve significantly lower error than its counterpart in the acoustic domain. Therefore f^V is learned independently beforehand and is frozen during the speaker turn embedding training process. In general any learning method can be used as long as the embedding space is equipped with Euclidean distance and shares the same final dimension with the acoustic counterpart.

To learn the projection f^A for speaker turn embedding, we define a loss as the sum of two terms: the loss to learn the metric, and the domain discrepancy loss. Concretely, we have:

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

As with f^V, $\mathcal{L}^A(f^A)$ can be any loss to learn the embedding. In this case we use the triplet loss, which can be defined from the generic loss in Eq. 2 as:

$$\mathcal{L}^A(f^A) = \frac{1}{|T^A|} \sum_{(x_a^A, x_p^A, x_n^A) \in T^A} l(x_a^A, x_p^A, x_n^A; f^A) \tag{7}$$

Given that the two embedding spaces can be constrained to lie within the same hypersphere, one can measure the discrepancy between the distributions of face embedded features $f^V(x_i^V)$ and auditory embedded features $f^A(x_j^A)$ using Eq. 5 as:

$$\mathcal{L}^{V \to A}(f^A) = \text{MMD}(\{f^V(x_i^V)\}, \{f^A(x_j^A)\}) \tag{8}$$

Based on Eq. 8, 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 together as in coupled multimodal matching [20, 23], minimizing Eq. 8 only regulates the statistical properties of the whole population in an unsupervised fashion. 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. In practice, we choose the associated kernel ϕ to be Radial Basis Function kernel, *i.e.* $k(u,v) = \exp(-d(u,v)^2)/\sigma$ in Eq. 5.

4 EXPERIMENTS

We first describe the datasets and evaluation protocols before discussing the implementation details and the experimental results. Our codes and models are publicly available[1].

4.1 Datasets

REPERE [13]. This standard dataset features programs including news, debates, and talk shows from two French TV channels along with annotations available through the REPERE challenge. All segments with face and voice from the same identity are collected. The resulting data is split into training and test sets, with 208 and 98 identities in each set respectively.

ETAPE [14]. This standard dataset contains 29 hours of audio-only news 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 [5]. The models learned with REPERE will be applied on ETAPE to benchmark their generalization ability. 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.

4.2 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 the threshold is varied.

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

- Weighted cluster purity (WCP) [32] and entropy (WCE): 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 and entropy are calculated for each cluster and averaged to get WCP and WCE.
- Operator clicks index (OCI-k) [16]: This is the total number of clicks required to manually label all clusters. For 1 cluster, 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. The cluster clicks are then added to produce the overall OCI-k performance measure.

4.3 Implementation details

Face embedding. Our face model is built using the ResNet-34 architecture[17] trained on the CASIA-WebFaces dataset [34]. A DPM face detector [10] is run to extract a tight bounding box around each face with no further preprocessing except for random flipping. ResNet-34 is first trained to predict 10,575 identities. After convergence, the last layer is removed and the weights are frozen. Then the last embedding layer with an embedding dimension of $h = 128$ is learned using the face tracks of the REPERE training set.

Speaker turn embedding. Our implementation of *TristouNet* consists of a bidirectional LSTM with the size of hidden states to be 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

[1] https://gitlab.idiap.ch/software/CTL-AV-Identification

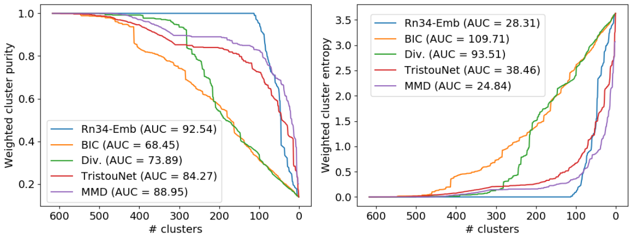

Figure 2: Hierarchical clustering evaluated on REPERE. (L) weighted cluster purity. (R) weighted cluster entropy.

to the LSTM, 13 Mel-Frequency Cepstral Coefficients (MFCC) are extracted with energy and their first and second derivatives.

All speaker turn embedding networks are trained using RMSProp optimizer [33] with a 10^{-3} learning rate, a fixed $\alpha = 0.2$, while $\sigma = 0.25, \lambda = 1.0$ are chosen using validation on the training set.

Baselines. We compare our speaker turn embedding with 3 approaches: Bayesian Information Criterion (BIC) [8], Gaussian divergence (Div.) [2], and the original *TristouNet* [5].

4.4 Experimental results

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. 2. 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. merges longer audio segments with enough data so it achieves higher purity. As small segments get progressively merged, the performance of BIC and Div. quickly deteriorate due to the lack of good voice statistics.

Our domain adaptation approach with MMD surpasses *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 why traditional approaches BIC and Div methods appear much worse with the clustering metrics. More experiments on full diarization are needed in future works.

Tab. 1 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 case and with the ideal number of clusters. This in practice can decrease the effort of human annotation by around 7.5%.

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.2. 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 adapting from visual embedding, we can improve *TristouNet* with a relative improvement of 3.7% of EER. It is also important to note that our models are trained on an independent training set (REPERE vs. ETAPE).

Table 1: Result of OCI-k metric on the REPERE test set. 'Min (# clusters)' 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 [8]	451 (390)	525
Div. [2]	330 (289)	521
TristouNet [5]	216 (119)	226
MMD	202 (94)	209

Table 2: EER reported on all pairs of 3746 sequences in ETAPE dev set.

	BIC[8]	Div.[2]	*TristouNet*[5]	MMD
EER	32.4	28.9	16.1	15.5

Table 3: Performance when training data are limited. EER is reported on ETAPE dev set. OCI-k is reported on REPERE.

	100%		60%		30%	
	[5]	MMD	[5]	MMD	[5]	MMD
Min OCI-k	216	202	274	229	249	213
OCI-k@98	226	209	285	231	263	221
EER	16.1	15.5	19.1	18.41	16.86	16.52

Results with limited data. To benchmark the generalization of our approach, the same verification and clustering protocols are applied when the amount of training audio data is reduced and reported in Tab. 3. In all 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 clustering protocol. On the other hand, our visual guided system is only affected slightly. 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.

5 CONCLUSION

We have presented a domain adaptation approach to improve speaker turn embedding using knowledge from a source face embedding. By optimizing the maximum mean discrepancy, our method exploits the regularization of the two distributions of visual and auditory features within the common hypersphere. 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, where backchannels and short answers/utterances are very frequent. One additional advantage of our work 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.

Acknowledgement. This work was supported by the EU projects EUMSSI (FP7-611057) and MuMMER (H2020-688147).

REFERENCES

[1] M. Baktashmotlagh, M. Harandi, and M. Salzmann. Distribution-matching embedding for visual domain adaptation. *Journal of Machine Learning Research*, 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*, 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] 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.

[5] H. Bredin. TristouNet: Triplet Loss for Speaker Turn Embedding. In *ICASSP*, New Orleans, USA, 2017. IEEE.

[6] 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.

[7] P. Chakravarty, J. Zegers, T. Tuytelaars, et al. Active speaker detection with audio-visual co-training. In *ICMI*. 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] C. Dubout and F. Fleuret. Deformable part models with individual part scaling. In *BMVC*, 2013.

[11] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[12] 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.

[13] 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.

[14] 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.

[15] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. *NIPS*, 2007.

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

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

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

[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.

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

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

[22] 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.

[23] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[25] 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.

[26] 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.

[27] A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. *arXiv preprint arXiv:1603.06432*, 2016.

[28] 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.

[29] 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.

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

[31] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2001.

[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.

[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.

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