Seeing voices and hearing voices:
Learning discriminative embeddings using cross-modal self-supervision

Soo-Whan Chung[1], Hong-Goo Kang[1], Joon Son Chung[2]

[1]Department of Electrical & Electronic Engineering, Yonsei University, South Korea
[2]Naver Corporation, South Korea

`joonson.chung@navercorp.com`

Abstract

The goal of this work is to train discriminative cross-modal embeddings without access to manually annotated data. Recent advances in self-supervised learning have shown that effective representations can be learnt from natural cross-modal synchrony. We build on earlier work to train embeddings that are more discriminative for uni-modal downstream tasks. To this end, we propose a novel training strategy that not only optimises metrics across modalities, but also enforces intra-class feature separation within each of the modalities. The effectiveness of the method is demonstrated on two downstream tasks: lip reading using the features trained on audio-visual synchronisation, and speaker recognition using the features trained for cross-modal biometric matching. The proposed method outperforms state-of-the-art self-supervised baselines by a significant margin.

Index Terms: self-supervised learning, metric learning, cross-modal, speaker recognition, lip reading.

1. Introduction

How do we learn and understand the world around us? In contrast to the current paradigm in machine learning, humans accumulate most of their knowledge through experience rather than taught with rules and examples. Literature in perceptual studies suggest that humans rely on hearing sounds to learn about various concepts in the visual domain and vice-versa [1, 2]. For example, most people would think that it is strange to hear a female voice from a man's face, but this is something that they usually learn through experience, not through explicit teaching.

Deep learning has led to notable success of many fields, including image recognition [3, 4], speech recognition [5, 6] and machine translation [7, 8]. Fully labelled datasets already exist in these popular fields of research, but there are multiple challenges to hand-crafting datasets for every task and application – (i) the cost of annotation can be prohibitively expensive; (ii) human annotation of data often has privacy issues, particularly those regarding biometrics; (iii) some tasks such as speaker verification can be extremely challenging even for human experts; (iv) other tasks such as emotion recognition is inherently ambiguous and cannot be clearly divided into categories.

The Internet provides abundant resources that can be used to learn semantics and concepts, but these have not been utilised in the prevailing supervised learning framework. Recently, self-supervised learning has received a growing amount of attention with applications in many research areas [9, 10, 11, 12]. Most of the previous work on self-supervision are trained to reconstruct the input data [13] or to predict withheld parts of the data,

such as inpainting missing part of images [14], predicting relative position of image patches [15] and colourising RGB images from only grey-scale images [16].

Of closer relevance to this paper is *cross-modal* self-supervision, where the supervision comes from the correspondence between naturally co-occurring data streams, such as sound and images. Earlier works in this area have used pairwise objectives to force corresponding embeddings to be closer together and non-corresponding embeddings to be further apart using binary classification loss [17, 18, 19] or contrastive loss [20, 21, 22, 23]. Our recent work [24] has demonstrated that the performance on various cross-modal tasks can be improved by re-formulating the problem as a multi-way matching task. The use of multiple negatives has also been discussed in the context of supervised metric learning [25, 26] where they have also discovered advantages over pairwise objectives.

These networks are used not only for the cross-modal matching task that they have been trained on (such as audio-video synchronisation [20, 24] or cross-modal biometrics [27]), but also for downstream tasks such as lip reading [20, 24] and speaker recognition [28]. This shows that the self-supervised training allows representative speech and speaker features to be learnt in the individual modalities.

The goal of this work is to learn embeddings that are discriminative for the uni-modal downstream tasks, as well as for the primary cross-modal tasks that the networks have been trained on. The work naturally builds on [24, 28] which show that the embeddings learnt via self-supervised proxy task demonstrate promising performance on these downstream tasks. We propose a novel training function that not only penalises the distances between a video segment and non-corresponding audio segments, but also between the different video segments as well as between the different audio segments. While the previous works enforce intra-class separation and inter-class compactness *across* modalities, they do not penalise or optimise metrics *within* modalities. Therefore, the proposed cross-domain discriminative loss explicitly enforces intra-class separation *within* modalities in addition to the aforementioned criterion, helping the network to learn more discriminative embeddings.

The effectiveness of this method is demonstrated on three experiments each using different modalities: (1) audio-only speaker verification where the task is to verify whether or not two speech segments come from the same person, (2) audio-visual biometric matching which aims to determine whether the face image and the speech segment belongs to the same person, and (3) video-only lip reading which recognises words from silent speech. The models trained using the proposed criterion is able to produce powerful representations of both streams which are effective for all of these tasks, where we demonstrate performance that exceed state-of-the-art self-supervised baselines.

2. Training functions

This section outlines training strategies for learning cross-modal embeddings. Existing works including AVE-Net [18], SyncNet [20] are described, followed by the cross-modal matching strategy [24] and our proposed variants.

2.1. Network architecture

Our architecture consists of two sub-networks – one that ingests face images as input, and another that takes mel-filterbank features of speech segments. Both sub-networks are based on the VGG-M architecture [29] which strikes a good trade-off between efficiency and performance. The input representations and network architectures are identical for the both tasks – audio-visual synchronisation and cross-modal biometrics. The detailed architecture is given in Table 1.

Audio stream. The inputs to the audio stream are 40-dimensional mel-filterbank coefficients in logarithm scale from sliding a hamming window of width 25ms with a stride of 10ms. The input size is T frames in the time-direction, where the value of T depends on the task. The network is based on the VGG-M CNN model, but the filter sizes are modified for the audio input. The layer configurations are the same as [28]. The network has a temporal stride of 4 in order to synchronise the sampling rate between the audio and the video streams.

Visual stream. Visual stream ingests a video of cropped face, with a resolution of 224×224 and a frame rate of 25 fps. The visual stream is also based on the VGG-M, but the first layer is a 3D convolution layer with a filter size of $5 \times 7 \times 7$. The feature extractor therefore has a temporal receptive field of 5 video frames and stride of 1 frame.

Table 1: *Architecture of two-stream networks for audio-visual synchronisation*

(a) Audio stream			(b) Visual stream		
conv1	[3 × 3]	64	conv1	[5 × 7 × 7]	96
pool1	[1 × 1]		pool1	[1 × 3 × 3]	
conv2	[3 × 3]	192	conv2	[1 × 5 × 5]	256
pool2	[3 × 3]		pool2	[1 × 3 × 3]	
conv3	[3 × 3]	384	conv3	[1 × 3 × 3]	256
conv4	[3 × 3]	256	conv4	[1 × 3 × 3]	256
conv5	[3 × 3]	256	conv5	[1 × 3 × 3]	256
pool5	[3 × 3]		pool5	[1 × 3 × 3]	
conv6	[3 × 3]	512	conv6	[1 × 6 × 6]	512
fc7	512		fc7	512	
fc8	128		fc8	128	

2.2. Loss functions

We describe the proposed training strategies for the cross-modal embeddings, and compare it to the existing state-of-the-art methods for audio-visual correspondence. Figure 2 gives an overview of the structure and training criteria of the existing and proposed approaches.

Pairwise objectives. SyncNet [20] and AVE-Net [18] learn cross-modal correspondence using pairwise objectives. The objective is to learn representations in the joint embedding space such that the corresponding pairs are close together in the embedding space, and non-corresponding pairs are far apart. The networks are trained with contrastive loss [20] or with binary classification loss on the pairwise distances [18]. For all training functions introduced in this section, it is assumed that all non-matching pairs are *negative* examples from different classes (e.g. different identity in the cross-modal biometrics task).

Multi-way matching. While most of the work in cross-modal learning has used pairwise objectives, [24] proposed a network trained via a *multi-way matching* task. The pairwise losses only enforce that an embedding is far from one particular negative embedding, not all negatives. This algorithm adopts *relative* similarity of the matching pair over all non-matching pairs, leading to more stable learning. The method is related to the GE2E loss [25] and multi-class N-pair loss [26] proposed in supervised metric learning.

For audio-to-video matching, the learning criterion compares one input feature from the audio stream to multiple features from the video stream. The task can be set up as any N-way feature matching task, with the objective function:

$$L_{AV} = -\frac{1}{N} \sum_{j=1}^{N} \log \frac{S(\mathbf{x}_A^j, \mathbf{x}_V^j)}{\sum_{k=1}^{N} S(\mathbf{x}_A^j, \mathbf{x}_V^k)}, \quad (1)$$

where $\mathbf{x}_A$ and $\mathbf{x}_V$ are the audio and video embeddings respectively, and N is the batch size. For each video input, Euclidean distances between the video feature and N audio features are computed, and the network is trained on the *inverse* of this distance. Therefore the similarity function S can be written as:

$$S(\mathbf{x}_a, \mathbf{x}_b) = \exp(\|\mathbf{x}_a - \mathbf{x}_b\|_2^{-1}). \quad (2)$$

The overall multi-way matching loss is:

$$L_{\text{MWM}} = L_{AV} + L_{VA}. \quad (3)$$

The objective encourages the relative distance between the corresponding pairs of the audio and the videos features to be closer than the non-corresponding pairs.

Multi-way matching with angular criterion. In metric learning, it has been observed that networks trained with the cosine distance metric often generalise better in unseen conditions compared to those trained with the Euclidean distance [30, 25]. However, the use of cosine distance combined with the softmax loss has not led to effective learning since the distances are bound to $[-1, 1]$ resulting in small entropy. To resolve this, some previous works have used fixed scale parameters [31, 32] to increase the dynamic range of the input to the softmax layer. In the context of metric learning, [25] uses *learnable* scale parameters trained together with the network. We adopt this strategy, and replace the Euclidean metric in [24] with the cosine distance with learnable scale. The score function is defined as below, where w and b are learnable parameters.

$$S(\mathbf{x}_a, \mathbf{x}_b) = \exp(w \cdot \cos(\mathbf{x}_a, \mathbf{x}_b) + b). \quad (4)$$

Cross-domain discriminative loss (CDDL). While the existing training objectives optimise relative distances *across* modalities, they do not explicitly enforce discriminative power *within* modalities. In addition to the cross-modal matching objective in [24], we introduce additional criterion that enforces the distances of different inputs from the same modality to be far apart, which enforces intra-class separation within the modality. This is particularly relevant to self-supervised representation learning, since the learnt representations are not only used for the
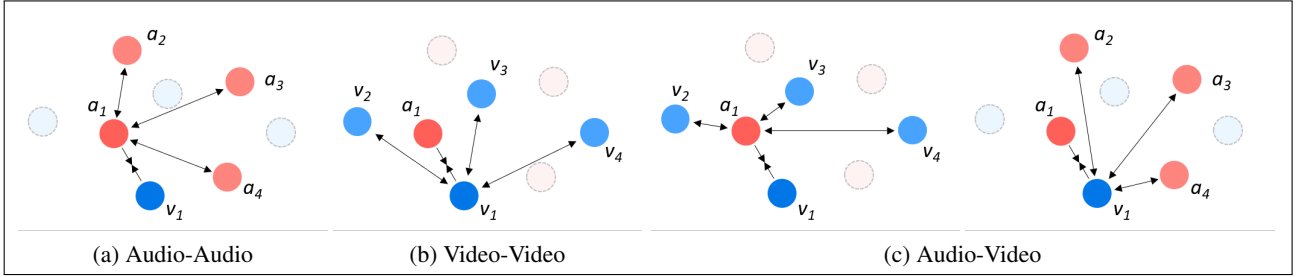
Figure 1: *Cross-domain discriminative loss*

cross-modal task, but also for uni-modal downstream tasks such as speaker verification.

In an unsupervised setting, it is not possible to obtain *positive* pairs from the same class. Since the features for both modalities are represented in a joint embedding space, the positive pairs are taken from the other modality, as in the cross-modal matching task. This can be formulated as:

$$L_{AA,V} = -\frac{1}{N} \sum_{j=1}^{N} \log \frac{S(\mathbf{x}_A^j, \mathbf{x}_V^j)}{S(\mathbf{x}_A^j, \mathbf{x}_V^j) + \sum_{k \neq j}^{N} S(\mathbf{x}_A^k, \mathbf{x}_A^j)},$$ (5)

where the score function of Equation (4) is used. The use of positives from the other modality is comparable to mining *hard positives* in metric learning. The overall loss function is defined as below, and is also illustrated in Figure 1.

$$L_{\text{CDDL}} = L_{AV} + L_{VA} + L_{AA,V} + L_{VV,A}$$ (6)

3. Experiments

The models are trained two tasks and evaluated on three different tasks.

3.1. Cross-modal biometrics

While face and voice recognition are both widely used to verify a person's identity, the research in each field have been mostly independent due to their heterogeneous feature characteristics [33, 34]. There has been a number of recent works that have demonstrated that a joint embedding of face and voice characteristics can be learnt for cross-modal biometric matching or verification [35, 36, 37, 27, 28].

In this section, the training setup mimics that of the biometrics task of [24], in which the models are trained for cross-modal biometric matching. We evaluated the trained model on two tasks: cross-modal biometric verification and audio-only speaker recognition.

Training representations. The cross-modal matching network is trained on the VoxCeleb2 dataset which consists of 5,994 different speakers with 1,092,009 clips in the training set and 118 speakers with 36,237 clips in test. The dataset is collected from YouTube clips, and the speech data comes together with face-cropped video and identity metadata. The identity labels are not used during training – the training strategy uses audio-visual co-occurrences as the only form of supervision. We therefore assume that the audio-image pair from the same video are of the same identity, and audio-image pairs from different videos are of different identities.

The network is trained with an auxiliary *content* loss on the synchronisation task, since [28] has shown that this helps

Table 2: *Identity verification results on the VoxCeleb1 test set. Lower is better.* **IL** *denotes the primary identity loss,* **CL** *denotes the auxiliary content loss. For more details on the joint training, refer to [28].* **CBM***: Cross-modal Biometric Matching;* **SV***: Speaker Verification.*

Method	CBM EER	SV EER
IL only [28]	24.8%	23.15%
IL + CL [28]	23.1%	22.59%
IL + CL + Disent. loss [28]	18.9%	22.09%
IL + CL (Angular)	14.9%	20.25%
IL + CL + CDDL (Angular)	**13.8%**	**17.52%**

to improve performance and reduced over-fitting on the identity task by regularising features in the common encoders. Except for the training functions proposed in Section 2, the implementation details such as network specification, batch size and clip duration are the same as [28] so that the results can be compared directly.

The two stream network generates $N \times T$ identity vectors from each stream, where N is the batch size and T is the number of features in the time domain. During training, the T identity vectors from the audio stream are then averaged into a single vector, while a single identity vector is selected from the face stream at random. The rationale for this choice is as follows: speaker recognition requires an input of sufficient length to capture voice characteristics, but face characteristics can be obtained from a single image; we randomly select one image feature rather than average over many to prevent the model from lip reading (learning the distribution of phonemes over the time sequence).

Evaluation on cross-modal biometric matching. The objective of the task is to determine whether the face image and the speech segment comes from the same person. Following [28], 300,000 pairs are audio-image pairs randomly generated from the test set of VoxCeleb2 [38], of which 10,000 are positive and the rest are negative.

Evaluation on speaker verification. The task is to verify whether or not two speech segments come from the same person. We extract identity embeddings for the data in the VoxCeleb1 *test* set, which consists of 40 speakers [39]. The negative cosine distance between embeddings is calculated directly, then is used as the similarity score between verification pairs.

We evaluate this on the VoxCeleb1 [39] test set. Cosine similarity between a pair of audio embeddings is used as the scoring function.
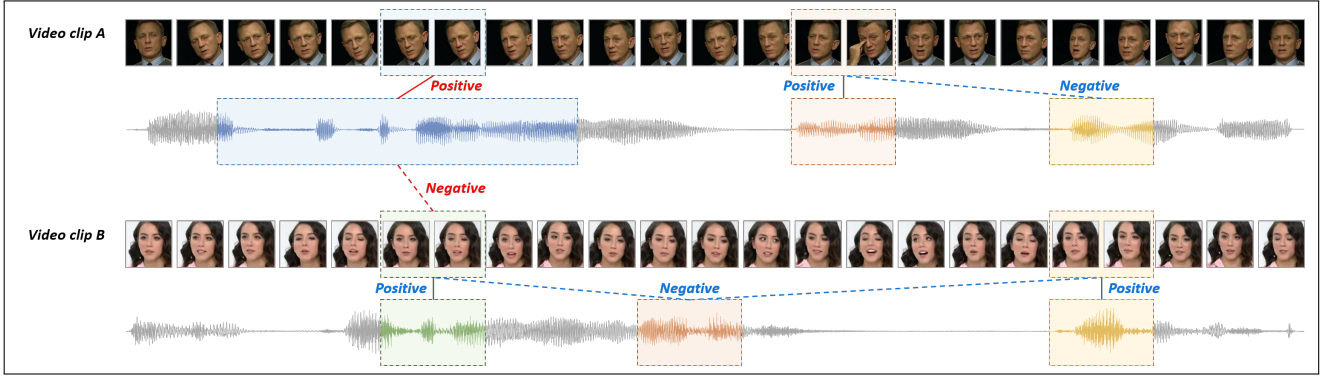
Figure 2: *Sampling strategy for self-supervised training. The* red *lines represent the pairs for cross-modal biometrics, the* blue *lines represent the pairs for lip synchronisation.*

3.2. Visual speech recognition

Speech and lip motion contain complementary information that is often used together by humans to interpret what is being said. In filming and broadcasting, audio-visual synchronisation is a common problem, but the solution to this problem using the speech *content* has only recently received attention.

In order to synchronise the two streams, the *phonetic* content at a particular time in one data stream must be matched to that in the other stream. Audio-to-video synchronisation is therefore naturally a cross-modal retrieval task, where the offset is found by selecting an audio segment from a set, given a video segment.

In this experiment, the models are trained on the cross-modal synchronisation as the proxy task and the embeddings evaluated on the video-only downstream task of visual speech recognition (lip reading).

Training representations. The self-supervised methods are trained on the *pre-train* set of the Lip Reading Sentences 2 (LRS2) [40] dataset. The pre-train set contains 96,318 video clips containing continuous speech of various lengths. The text labels and the metadata in the LRS2 dataset are not used in any way to train the model.

Following [24], audio and video segments for the matching pairs are taken at the same timestep from a facetrack; the segments for non-matching pairs are taken at different timesteps but still from the same facetrack. This is to encourage the network to learn phonetic information as opposed to biometric information.

Evaluation on lip reading. The effectiveness of the trained embeddings are evaluated on the word-level classification task of the Lip Reading in the Wild (LRW) [20] dataset. The dataset has a vocabulary size of 500 and over 500,000 word-level utterances, of which 25,000 are reserved for testing. The utterances are spoken by hundreds of different speakers.

For the fully supervised baselines, the VGG-M based networks are trained end-to-end on the LRW dataset. For the pre-trained (PT) methods, the front-end feature extractors are trained with self-supervision using the methods described in Section 2, and only the back-end classifier is trained with a classification loss with supervision.

The classifiers are trained for 500-way classification and we report top-1 and top-10 accuracies on the test set. The model specifications are identical to that described in Table 2 of [24]

to enable direct comparison of the results.

Table 3: *Word accuracy of visual speech recognition using various architectures and training methods.* **R@K**: *Recall at K or top-K accuracy.*

Arch.	Method	R@1	R@10
MT-5	E2E [41]	66.8%	94.6%
TC-5	E2E [24]	71.5%	95.9%
TC-5	PT - AVE-Net [18]	66.7%	94.0%
TC-5	PT - SyncNet [20]	67.8%	94.3%
TC-5	PT - Multi-way [24]	71.6%	95.2%
TC-5	PT - Multi-way (Angular)	73.8%	96.4%
TC-5	PT - CDDL (Angular)	**75.9%**	**97.0%**

Results. The results on the lip reading experiments are reported in Table 3. When trained end-to-end with full supervision, the classification accuracy of the TC-5 model introduced in [24] is higher than that the existing networks of similar size such as the MT-5 model in [41]. The results show that our proposed models outperform the baselines by a significant margin. It is notable that the self-supervised methods can outperform the end-to-end trained models even with the large-scale fully labelled dataset.

4. Conclusions

In this paper, we proposed a new strategy for learning discriminative cross-modal embeddings without access to manually annotated data. The method enforces not only intra-class separation of features across modalities, but also within modalities helping the model to learn more discriminative embeddings. The model trained using the proposed objective outperforms existing methods on cross-modal biometric verification which is the task that it has been trained on, as well as on the downstream tasks such as speaker recognition and visual speech recognition.

5. Acknowledgements

6. References

[1] P. Belin, S. Fecteau, and C. Bedard, "Thinking the voice: neural correlates of voice perception," *Trends in cognitive sciences*, vol. 8, no. 3, pp. 129–135, 2004.

[2] L. Shams and R. Kim, "Crossmodal influences on visual perception," *Physics of life reviews*, vol. 7, no. 3, pp. 269–284, 2010.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[4] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. CVPR*, 2015, pp. 3367–3375.

[5] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*. IEEE, 2013, pp. 6645–6649.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*. IEEE, 2016, pp. 4960–4964.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[8] K. Cho, B. van Merrienboer, C. Gulçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014.

[9] N. Goncalves, J. Nikkilä, and R. Vigario, "Self-supervised mri tissue segmentation by discriminative clustering," *International journal of neural systems*, vol. 24, no. 01, p. 1450004, 2014.

[10] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. CVPR*, 2017, pp. 932–940.

[11] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proc. ICCV*, 2017, pp. 2051–2060.

[12] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[14] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. CVPR*, 2016, pp. 2536–2544.

[15] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. ICCV*, 2015, pp. 1422–1430.

[16] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. ECCV*. Springer, 2016, pp. 649–666.

[17] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. CVPR*, 2018.

[18] R. Arandjelović and A. Zisserman, "Objects that sound," in *Proc. ECCV*, 2018.

[19] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *Proc. CVPR*, 2018, pp. 4358–4366.

[20] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[21] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *NIPS*, 2018.

[22] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *CoRR*, vol. abs/1906.05849, 2019. [Online]. Available: http://arxiv.org/abs/1906.05849

[23] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proc. ECCV*, 2020.

[24] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Self-supervised embeddings for cross-modal retrieval," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[25] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*. IEEE, 2018, pp. 4879–4883.

[26] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016, pp. 1857–1865.

[27] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. ECCV*, 2018.

[28] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *Proc. ICASSP*, 2020.

[29] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC.*, 2014.

[30] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[31] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. CVPR*, 2018, pp. 5265–5274.

[32] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[33] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. ICASSP*, 2002.

[34] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2006.

[35] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *Proc. CVPR*, 2019, pp. 7539–7548.

[36] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Proc. ACCV*, 2018.

[37] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proc. CVPR*, 2018.

[38] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[39] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[40] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. CVPR*, 2017.

[41] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *CVIU*, 2018.