Perfect Match: Self-Supervised Embeddings for Cross-Modal Retrieval

Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang

***Abstract*—This paper proposes a new strategy for learning effective cross-modal joint embeddings using self-supervision. We set up the problem as one of cross-modal retrieval, where the objective is to find the most relevant data in one domain given input in another. The method builds on the recent advances in learning representations from cross-modal self-supervision using contrastive or binary cross-entropy loss functions. To investigate the robustness of the proposed learning strategy across multi-modal applications, we perform experiments for two applications – audio-visual synchronisation and cross-modal biometrics. The audio-visual synchronisation task requires temporal correspondence between modalities to obtain joint representation of phonemes and visemes, and the cross-modal biometrics task requires common speakers representations given their face images and audio tracks. Experiments show that the performance of systems trained using proposed method far exceed that of existing methods on both tasks, whilst allowing significantly faster training.**

***Index Terms*—Cross-modal, multi-modal, self-supervision, embedding, retrieval.**

I. Introduction

SUPERVISED learning with deep neural networks (DNN) has brought phenomenal advances to various fields of research such as image recognition [1], [2], speech recognition [3], [4], machine translation [5], [6]. The performance of such DNN-based systems relies heavily on the quality and quantity of annotated databases that are tailored to the particular applications. In many popular fields of research, large-scale labelled datasets already exist [7], [8], however it is prohibitively expensive to manually collect and annotate databases for every application. There is a plethora of data on the Internet that are not used in machine learning due to the lack of such labels. Unsupervised or semi-supervised learning strategies are sometimes used when it is not feasible to obtain labelled training data for some applications – in particular, self-supervision allows a model to learn the natural characteristics of the data itself or relationship between different modalities, and hence capitalise on the raw data without any manual annotations.

Manuscript received September 16, 2019; revised March 24, 2020; accepted April 5, 2020. Date of publication April 14, 2020; date of current version June 24, 2020. This work was supported by Naver Corporation. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong He. *(Corresponding author: Hong Goo Kang.)*

Soo-Whan Chung and Hong-Goo Kang are with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea (e-mail: jsh6293@dsp.yonsei.ac.kr; hgkang@yonsei.ac.kr).

Joon Son Chung is with the Naver Corporation, Seongnam-si, Gyeonggi-do 13561, South Korea (e-mail: joonson.chung@navercorp.com).

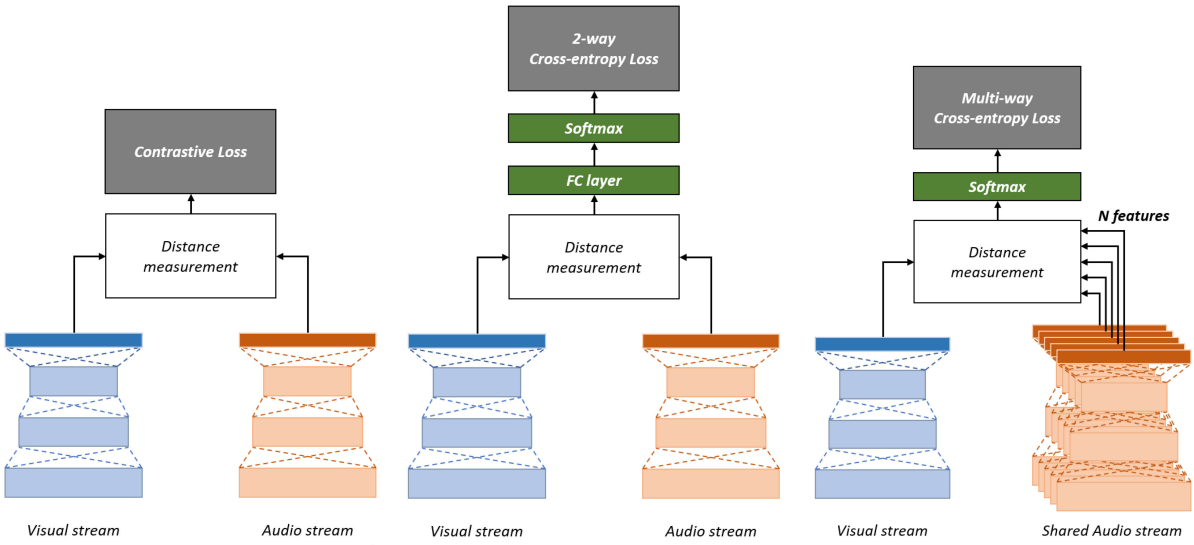
Digital Object Identifier 10.1109/JSTSP.2020.2987720

Self-supervised learning has received a growing amount of interest in the recent years, with a number of applications in many research areas [9]–[12]. One of the earlier adaptations of such idea is the work on auto-encoders [13], and there are more recent works on learning representations via data imputation such as context prediction by inpainting missing images [14] or colourising RGB images from only grey-scale images [15]. Recently, the use of *cross-modal* self-supervision has proved particularly popular, where the supervision comes from the correspondence between two or more naturally co-occurring streams, such as sound and images.

Traditional literature in this field computes correlations by measuring relevance between modalities, e.g. Canonical Correlation Analysis (CCA) [16], Partial Least Square (PLS) [17], [18], a Bilinear Model (BLM) [19], in which it is difficult to learn joint information on mismatched latent domains. With the advances in deep learning, recent papers have proposed methods for jointly training audio and video representations for source localisation [20], [21], cross-modal retrieval [20], audio-visual synchronisation [22], [23], recognition [24], [25] and cross-modal biometrics [26], [27]. These applications commonly adopt two-stream architectures and they are trained to predict whether the cross-modal inputs are matching or not with contrastive loss [22], [24] or with binary classification objective [20], [21], [25]. Although these training strategies show promising results on cross-modal retrieval tasks, the binary objectives do not directly address the task of retrieval.

In this paper, we propose a novel training strategy for cross-modal learning, where we learn powerful cross-modal embeddings through a multi-way matching task. In particular, we combine the similarity-based methods (*e.g.* L2 distance loss) used to learn joint embeddings across modalities, with a multi-class cross-entropy loss; this way, the training objective naturally lends itself to cross-modal retrieval where the task is to find the *most* relevant sample in one domain to a query in another modality, rather than training for a proxy task such as binary classification. We propose a new training strategy in which the network is trained for the multi-way matching task without explicit class labels, whilst still benefiting from the favourable learning characteristics of the cross-entropy loss.

We demonstrate the effectiveness of the proposed method in two different cross-modal retrieval tasks. The first is audio-visual synchronisation, where the objective is to locate the most relevant audio segment given a short video clip. The models trained for multi-way matching is able to produce powerful representations of the auditory and visual information that can



(a) SyncNet (b) AVE-Net, L3-Net (c) Proposed model

Fig. 1. Comparison between the existing and proposed training strategies.

be applied to other tasks – we also demonstrate that the learnt embeddings show better performance on a visual speech recognition task compared to the representations learnt via pairwise objectives. The second application is cross-modal biometrics, where the objective is to predict who is speaking from a face image and vice-versa. Here, the network learns to identify person characteristics across modalities, even between identities with same gender, nationality and age. The training objective using multi-way matching is better suited to the retrieval task compared to the methods that use binary decision as a proxy task.

The paper is organised as follows. In Section II, we describe the proposed self-supervised learning strategy. Section III describes the experiments on the audio-video synchronisation and an application of the learnt embedding to a downstream task. Section IV demonstrates the effectiveness of the proposed method on cross-modal biometrics, and the conclusion follows in Section V.

II. Cross-Modal Embeddings for Self-Supervision

In this section, we describe training strategies for the cross-modal matching task, and compare it to the existing state-of-the-art methods for audio-visual correspondence, including AVE-Net [20] and SyncNet [22]. The two baselines are trained as a pairwise correspondence task, whereas the proposed method is set up as a multi-way matching task. Fig. 1 gives an overview of the structure and training criteria of the existing and proposed approaches.

A. Baseline Methods

There are a number of works based on joint embedding of multi-modal data [28]–[30], including a number of more recent works on self-supervision [24], [25], [31]. The deep learning-based approaches are used to learn representations for each modality through two-stream networks whose the outputs are of the same dimension regardless of the input size of each modality. The networks measure similarity or distance between the embedded features extracted from each stream. For a positive (matching) pair, the distance is minimised such that the representation from both modalities are close together in the joint embedding space, and this is maximised for a negative (non-matching) pair.

Baseline – SyncNet: SyncNet [22] depicted in Fig. 1a performs synchronisation between lip movement and speech signal using a contrastive loss which was originally proposed for training Siamese networks [32]. The contrastive loss function is defined as:

$$E = \frac{1}{2N}\sum_{n=1}^{N}(y_n)d_n^2 + (1-y_n)\max(margin - d_n, 0)^2$$
$$d_n = ||v_n - a_n||_2, y_n \in \{0,1\} \quad (1)$$

where y_n is the binary similarity metric between inputs, d_n is distance, v_n, a_n are visual and audio embedding features, respectively. N is the number of samples. Contrastive loss uses an anchor and a margin to minimise embedding differences between synchronous pairs and to maximise distances between out-of-sync pairs, where the training characteristics are affected by the manually tuned margin value.

Baseline – AVE-Net: Fig. 1b depicts L3-Net [31] and AVE-Net [20]. L3-Net and AVE-Net are trained to classify and localise sound sources in video frames. Both methods are similar to SyncNet in that they are trained to identify whether a pair is matching or not.

The embedded vectors are L2 normalised, then the Euclidean distance between the two normalised embeddings are computed, before being passed through a fully-connected layer and finally binary cross-entropy loss (Equation 2).

$$E = -\frac{1}{2N}\sum_{n=1}^{N} y_n \log f(d_n) + (1-y_n)\log f(1-d_n), \quad (2)$$

where d_n is the distance between the cross-modal embeddings, $f(\cdot)$ is a fully-connected layer and y_n is the binary similarity metric. Binary cross-entropy criterion naturally handles the distance between genuine or false pair to maximise posterior probability in a logistic decision task. The fully-connected layer essentially learns the threshold on the distance above which the features are deemed not to correspond and it has the advantage over SyncNet in that it does not require manual tuning of the margin parameter.

B. Proposed Strategy

Unlike previous methods that use pairwise losses, the proposed embeddings are learnt here via a multi-way matching task. Pairwise losses only enforce that an embedding is far from one particular negative embedding, not all negatives. The proposed algorithm enforces relative similarity of the matching pair over all non-matching pairs, leading to more stable learning. The use of multiple negatives has also been explored in supervised metric learning, where they have found advantages over pairwise methods [33], [34].

The learning criterion takes one input feature from the visual stream and multiple features from the audio stream, where the audio stream with shared parameters extract multiple features. This can be set up as any multi-way feature matching task. Euclidean distances between the audio and video features are computed, resulting in M distances. The network is then trained with a cross-entropy loss on the *inverse* of this distance. The proposed multi-way cross-entropy loss with softmax function is defined as:

$$E = -\frac{1}{2N}\sum_{n=1}^{N}\sum_{m=1}^{M} y_{n,m}\log(p_{n,m})$$

$$p_{n,m} = \frac{\exp(d_{n,m}^{-1})}{\sum_{m=1}^{M}\exp(d_{n,m}^{-1})} \tag{3}$$

where $d_{n,m}, y_{n,m}$ are negative similarity between embedding pairs and similarity metrics, respectively. The softmax function has similar effects to hard negative mining, since the hardest negative would most affect the gradients. M indicates the number of candidates pairs to be compared, and N is the amount of samples. The proposed algorithm maximises the posterior probability of the most suitable pair, thus it chooses the embedding pairs having similar distributions. The proposed learning strategy is shown in Fig. 1c.

III. Audio-Visual Synchronisation

In this section, we compare the performance of the proposed system to existing methods for lip synchronisation and a related audio-visual application.

Task definition: Audio-visual synchronisation is a common problem in film production and broadcasting since audio and video are usually recorded using different devices. In many cases, clapperboards are used to solve the synchronisation problem, but this method cannot be used to fix the synchronisation on pre-recorded audio and video. The issue can be solved by learning a joint embedding between the audio and the video, where the embeddings are implicit representations of the phonemes and visemes that can be captured from the audio and video inputs.

Audio-to-video synchronisation can be seen as a cross-modal retrieval task, where the temporal offset is found by selecting an audio segment from a set, given a video segment. This is done by computing the distance between a learnt video feature (from a 5-frame window) and a set of audio features.

TABLE I
Architecture of Two-Stream Networks for Audio-Visual Synchronisation

(a) Audio stream			(b) Visual stream		
conv1	$[3 \times 3]$	64	conv1	$[5 \times 7 \times 7]$	96
pool1	$[1 \times 1]$		pool1	$[1 \times 3 \times 3]$	
conv2	$[3 \times 3]$	192	conv2	$[1 \times 5 \times 5]$	256
pool2	$[3 \times 3]$		pool2	$[1 \times 3 \times 3]$	
conv3	$[3 \times 3]$	384	conv3	$[1 \times 3 \times 3]$	256
conv4	$[3 \times 3]$	256	conv4	$[1 \times 3 \times 3]$	256
conv5	$[3 \times 3]$	256	conv5	$[1 \times 3 \times 3]$	256
pool5	$[3 \times 3]$		pool5	$[1 \times 3 \times 3]$	
conv6	$[3 \times 3]$	512	conv6	$[1 \times 6 \times 6]$	512
fc	256		fc	256	

A. Network Architecture

The architecture of the audio and the video streams is described in this section. The inputs and the layer configurations are the same as SyncNet [22], so that the performance using the new training strategy can be compared to the existing methods. The network ingests 0.2-second clips of both audio and video inputs.

1) Audio Stream: The inputs to the audio stream are 13-dimensional Mel-frequency cepstral coefficients (MFCCs), extracted at every 10 ms with 25 ms frame length. Since the audio data is extracted from the video, there are natural environmental factors such as background noise and distortions in speech. The input size is 20 frames in the time-direction, and 13 cepstral coefficients in the other direction (so the input image is 13×20 pixels). The network is based on the VGG-M [35] CNN model, but the filter sizes are modified for the audio input size as shown in Table Ia.

2) Visual Stream: Visual stream input takes a video of a cropped face, with a resolution of 224×224 and a frame rate of 25 fps. The network ingests 5 stacked RGB frames at once, containing the visual information over the 0.2-second time frame. The visual stream network is also based on the VGG-M [35], but the first layer has a filter size of $5 \times 7 \times 7$ instead of 7×7 of the regular VGG-M, in order to capture the motion information over the 5 frames. The detailed architecture of visual stream is described in Table Ib.

3) Similarity Metric: For the synchronisation task, it is necessary to compare similarity or distance between the embeddings such that the best alignment can be found. Audio and visual embeddings include common linguistic information to be useful for corresponding one modality to another. Every embeddings taken from different offsets or clips should have unique representations. We use pairwise Euclidean distance to measure the difference of embeddings; the Euclidean distance metric has advantages over cosine distance for modelling distributions in that the embedding space is not limited to the unit sphere.

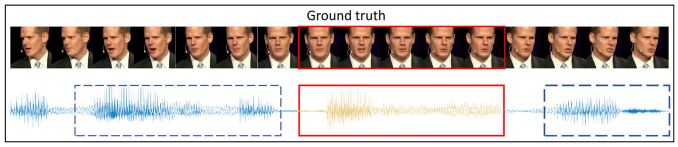
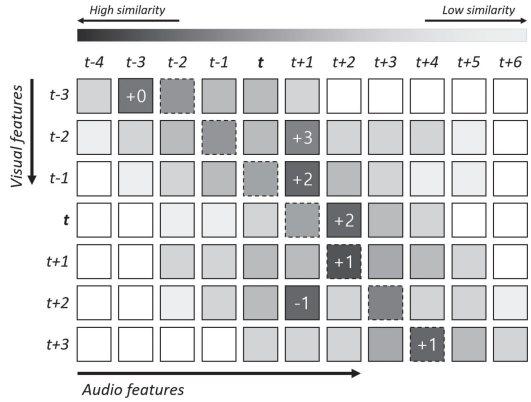


Fig. 2. Sampling strategy for self-supervised learning. The red rectangle highlights the audio segments that correspond to the talking face above, the blue dotted rectangles show non-matching audio segments.



# of context frames	5	7	9	11
Averaged offset	+2.00	+1.67	+1.40	+1.14
Predicted offset	+2	+2	+1	+1
Oracle offset	+1			

Fig. 3. Example of context windows computing distance between audio and visual segments for offset prediction. Darker shades indicate higher similarity without window. Synchronisation with context window may improve accuracy by averaging estimated offsets.

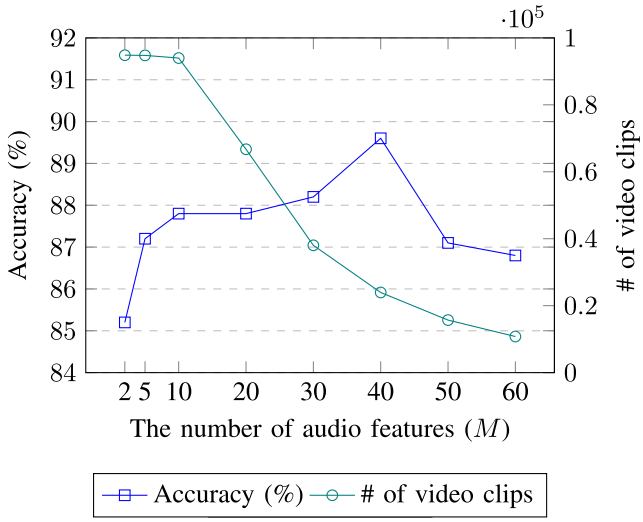


Fig. 4. Synchronisation accuracy and the number of available video clips according to *M*.

B. Experiments

Dataset: The network is trained on the pre-train set of the Lip Reading Sentences 2 (LRS2) [36] dataset. The LRS2 dataset contains 96,318 clips for training, and 1,243 for test. LRS2 dataset has no constraints on the vocabulary and each sentence is up to 100 characters or 6 seconds in length. There is a trade-off between the number of classes (or candidate audio features) M and the number of available video clips for training, since longer video clips are required to train networks with larger M (the candidate audio clips are sampled without overlap). This is because the positive (matching) and negative (non-matching) pairs are both sampled from the same clip, so that the network learns phonetic information as opposed to biometric information. The sampling strategy is illustrated in Fig. 2. The $M-1$ negative audio segments are sampled with random temporal offset without overlapped segments. We run experiments with different values of M in order to find the optimal value, and report the accuracy and the number of available video clips in Fig. 4.

Evaluation protocol: The task is to determine the correct synchronisation within a ± 15 frame window, and the synchronisation is determined to be correct if the predicted offset is within 1 video frame of the ground truth. A random prediction would therefore yield 9.7% accuracy.

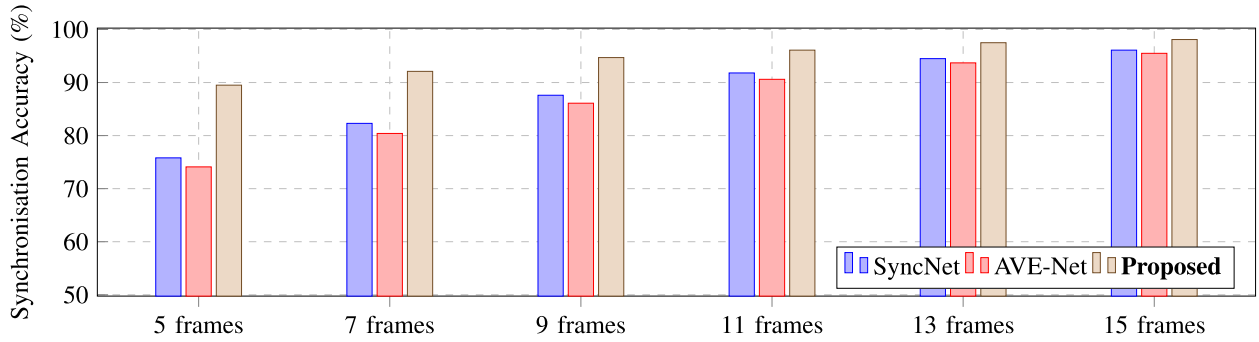
We assume that the two streams are synchronised when the distances between features are minimised. However as [22] suggests, one visual feature might not be enough to determine the correct offset, since not all samples contain discriminative information – for instance, there may be some 5-frame video segments in which nothing is said, or there may be phonemes that are repeated within a short time frame. Therefore, we also conduct experiments with the context window of more than 5 video frames, in which case we average the distances across multiple video samples (with a temporal stride of 1 frame).

For example in Fig. 3, it can be seen that whilst the offset values computed using single audio and video features are quite noisy, the offset value becomes more accurate as the context increases.

Results: The results of experiments are given in Fig. 5 using the network trained with $M = 40$ since this value of M gives the best performance in Fig. 4. The performance of the proposed method far exceeds the baseline trained with pair-wise objectives. In particular, for # frames $= 5$ (*i.e.* no context beyond the receptive field), there is a significant increase in synchronisation performance from 75.8% to 89.5%. Since our proposed learning approach directly addresses the task of searching for the most relevant segments among candidates, it is advantageous compared to the system only learning with a pairwise matching criterion.

C. Application: Visual Speech Recognition

The network learns a powerful embedding of the phonetic information contained in the input video. The objective of this experiment is to show that the embeddings learnt by the matching network are effective for a downstream task, in this case, visual speech recognition. This is demonstrated on a word-level speech recognition task, and we compare the performance using the embeddings learnt by the proposed self-supervised method to networks trained end-to-end with full supervision.



# Frames	5	7	9	11	13	15
SyncNet	75.8%	82.3%	87.6%	91.8%	94.5%	96.1%
AVE-Net	74.1%	80.4%	86.1%	90.6%	93.7%	95.5%
Proposed	**89.5%**	**92.1%**	**94.7%**	**96.1%**	**97.5%**	**98.1%**

Fig. 5. Synchronisation accuracy. **# Frames**: the number of visual frames for which the distances are averaged over.

TABLE II
ARCHITECTURE OF LIP READING TC-5 NETWORK

Lip reading network TC-5			
Front-end feature extractor	conv1	$[3 \times 3]$	64
	pool1	$[1 \times 1]$	
	conv2	$[3 \times 3]$	192
	pool2	$[3 \times 3]$	
	conv3	$[3 \times 3]$	384
	conv4	$[3 \times 3]$	256
	conv5	$[3 \times 3]$	256
	pool5	$[3 \times 3]$	
	conv6	$[3 \times 3]$	512
Backend classifier	conv7	$[5 \times 1 \times 1]$	
	conv8	$[5 \times 1 \times 1]$	
	fc	512	

TABLE III
WORD ACCURACY OF VISUAL SPEECH RECOGNITION USING VARIOUS ARCHITECTURES AND TRAINING METHODS

Architecture	Method	Top-1	Top-5
MT-5 [38]	E2E	66.8%	94.6%
LF-5 [38]	E2E	66.0%	93.3%
LSTM-5 [38]	E2E	65.4%	94.3%
TC-5	E2E	**71.5%**	**95.9%**
TC-5	PT - SyncNet	67.8%	94.3%
TC-5	PT - AVE-Net	66.7%	94.0%
TC-5	PT - **Proposed**	**71.6%**	**95.2%**

Dataset: We train and evaluate the models on the Lip Reading in the Wild (LRW) [37] dataset, which consists of word-level speech and video segments extracted from the British television. The dataset has a vocabulary size of 500, and contains over 500,000 utterances, of which 25,000 are reserved for testing. The utterances are spoken by hundreds of different speakers.

Architecture: The lip reading structure we use can be divided into a front-end feature extractor and a back-end classifier. The front-end architecture is taken from the visual stream of the network described in Section I. We propose a 2-layer temporal convolution back-end, followed by a 500-way softmax classification layer. This network structure is summarised in Table II and is referred to as **TC-5**. **'TC'** indicates temporal convolution model and **'5'** indicates the receptive frame size of the feature extractor in the temporal dimension. The baseline models Multiple Towers **(MT-5)**, Late Fusion **(LT-5)** and Long Short-Term Memory **(LSTM-5)** are models are from in [38].

The performance of the **TC-5** model exceeds the network designs proposed in [38] when trained end-to-end (E2E). The visual features are extracted in advance for the 'pre-trained' experiments (PT), and only the back-end layers are trained for the 500-way classification task – the feature extractor is not fine-tuned with full supervision.

Results: Table III summarises the results of the lip reading tasks. The recognition accuracy of the proposed TC-5 model is higher than that the architectures (MT-5, LF-5, LSTM-5) introduced by [38], trained end-to-end with full supervision. To evaluate the effectiveness of embedding strategy, we use the pre-trained (PT) networks trained with self-supervision (i.e. SyncNet, AVE-Net and the proposed method) as front-end visual extractors. For these experiments, only the back-end classifier is trained for the classification task with supervision. The front-end feature extractor is not fine-tuned together with the back-end classifier, since the objective of this experiment is to demonstrate the effectiveness of the self-supervised representations. It is noteworthy that the performance of the feature extractor trained with the self-supervised method matches that of the end-to-end trained network without any fine-tuning. In addition, the performance of the lip reading network trained using the proposed embeddings far exceeds the equivalent with SyncNet and AVE-Net.

IV. CROSS-MODAL BIOMETRICS

The objective of this section is to demonstrate the effectiveness of the learning strategy for the application of cross-modal biometrics. We compare the performance of the proposed system to existing methods for cross-modal identity retrieval and verification.

Task definition: Face and voice are both widely used to recognise a person's identity, but the research in each field have been evolved independently until recently due to the difficulty in building an unified statistical model as a result of their heterogeneous feature characteristics [39]–[41]. With the advances

TABLE IV
ARCHITECTURE OF TWO-STREAM NETWORKS FOR AUDIO-VISUAL BIOMETRICS

(a) **Audio stream**			(b) **Visual stream**		
conv1	$[5 \times 7]$	96	conv1	$[7 \times 7]$	96
pool1	$[1 \times 3]$		pool1	$[3 \times 3]$	
conv2	$[5 \times 5]$	256	conv2	$[5 \times 5]$	192
pool2	$[3 \times 3]$		pool2	$[3 \times 3]$	
conv3	$[3 \times 3]$	384	conv3	$[3 \times 3]$	384
conv4	$[3 \times 3]$	256	conv4	$[3 \times 3]$	256
conv5	$[3 \times 3]$	256	conv5	$[3 \times 3]$	256
pool5	$[3 \times 3]$		pool5	$[3 \times 3]$	
conv6	$[4 \times 1]$	512	conv6	$[6 \times 6]$	4096
pool6	$[1 \times 5]$		conv7	$[1 \times 1]$	4096
fc	512		fc	512	

in deep learning based methods, works on joint embedding or cross-modal learning have become increasingly popular [42]–[44]. For cross-modal person identification, it is important to derive common characteristic of the face and the corresponding voice signal. Thus, the main objective of the network is to extract embeddings in the latent domain such that the distance between embeddings of the same speaker is lower than that of different speakers. Earlier work [26] trained cross-modal embeddings in an unsupervised manner by introducing a criterion of contrastive loss (CL) and a curriculum-based hard negative mining (CHNM) technique [45].

In this paper, we conduct similar experiments with [26] and replace the training criterion by the proposed self-supervision method described in Section II. The objective is to learn a joint embedding of face images and voices using self-supervision without any identity labels. The evaluation involves two different sub-tasks – cross-modal retrieval and verification.

A. Network Architecture

The architecture and hyper-parameters are deliberately similar to [26] so that the experimental results can be compared directly to the previous work. The architectural details are described in Table IV.

1) Audio Stream: The audio stream ingests 2-seconds clip since it requires enough length to generate representative embeddings of person identities from speech. The inputs to the audio stream are the 40-dimensional Mel-filterbank coefficients in logarithm scale, extracted at 10 ms interval with 25 ms frame length. Therefore, the input audio feature has 200 frames in temporal axis and 40 coefficients on spectral axis; the input dimension to the CNN is 40×200. The network structure is based on the VGG-M model but the filter sizes are modified to be matched with the audio input size. Detailed configuration of audio stream is described in Table IVa.

2) Visual Stream: The visual input takes cropped face images of 224×224 pixels. The face image is randomly selected from the video clip. The image has 3 channels (RGB), so the input image size is $3 \times 224 \times 224$. The visual stream is also based on the VGG-M network with some changes in kernel size. Parameter settings for visual stream network is described in Table IVb.

3) Similarity Metric: The representations from the audio and image streams are mapped onto a common latent embedding space. We use the Euclidean distance to measure the distance between embeddings [46], [47]. This distance is minimised for the positive (same identity) pairs but maximised for negative (different identity) pairs.

B. Experiments

Dataset: For training the models, we use the same subset of the VoxCeleb1 [48] dataset that is used by [26]. The training set that consists of 901 speakers with 105,751 video clips. The validation set for the unseen-unheard case consists of 12,734 clips from 100 different speakers and the test set of 30,496 clips from 250 speakers. For some of the experiments, we also use VoxCeleb2 [49] dataset which consists of with 5,994 different speakers with 1,092,009 clips in the training set and 118 speakers with 36,237 clips in test. The VoxCeleb datasets are large-scale audio-visual databases extracted from YouTube clips, and the speech data comes together with face-cropped video and identity labels. The corpus therefore includes high reverberation with some background noise and interfering sounds and its degree of distortion varies by recording environments.

We acquire face image and audio segments from video and audio streams respectively. Unlike the synchronisation task described in Section III, the image and audio are sampled from independent points in time so that the network learns representations about the identity rather than phonetic or lexical information. The identity labels are not used during training – i.e. the training scheme uses co-occurrences as the only form of supervision.

Evaluation protocols: We propose two different sub-tasks relating to cross-modal biometrics – cross-modal retrieval and verification.

The objective of the retrieval sub-task is to find the corresponding face image given a speech segment and vice versa. This is evaluated as a 10-way matching task, where the network has to select the matching identity from randomly selected examples. Fig. 6 depicts an example of the cross-modal retrieval (forced matching) task.

The objective of the verification sub-task is to determine whether the face image and the speech segment comes from the same person. The evaluation metrics are the area under curve (AUC) of the receiver operating characteristic (ROC) curve and the equal error rate (EER).

Baseline: The baseline network for the verification task [26] has been trained with contrastive loss (CL) and curriculum hard negative mining (CHNM). The audio and image streams of the model marked as '**pretrained**' have been trained with full supervision on VGGFace [50] and VoxCeleb [48] datasets respectively.

Cross-modal retrieval: The baseline models have been trained using the contrastive loss with a number of different difficulties for the hard negative mining. **CL50**, **CL75** and **CL100** denotes the models in which the negatives are randomly mined from the hardest 50%, 75% and 100% respectively; **CL-Semihard** uses semi-hard negative mining strategy proposed by [51] in place of the fixed hard negative difficulty. The proposed model is trained both with and without L2 normalisation

Fig. 6. Examples of the 10-way cross-modal forced matching task. Upper figure illustrates face-to-voice matching, and bottom is for the voice-to-face matching task. The left-most column is the query.

of the embeddings. The L2 normalised embeddings are scaled up by a fixed factor of 5 to accelerate training with the cross-entropy loss. The number of candidate pairs $M = 200$ is used. For the models trained using contrastive loss, only the results with L2 normalisation is shown since the network does not converge without normalisation. All experiments are run at least 2 times and the performance values for each epoch are averaged.

The network is trained for 100 epochs, which takes around a week using a Tesla P40 accelerator. The results in Fig. 7a show that the proposed method converges much faster than the baselines whilst also giving better final performance, since the network has been trained on the matching task that is better in line with the task objective.

The baselines used for experiments on the VoxCeleb2 are same as that used for the retrieval sub-task. The results in Fig. 7b also demonstrate that the proposed method converges much faster than the baselines, whilst showing competitive results despite the fact that the baseline network has been trained on an objective that is more closely related to the verification task.

On the VoxCeleb1 unseen-unheard test set, the trained model shows strong results compared to that reported by the authors of [26] using the same initialisation method [52], training data and supervision (Table V). The proposed method is competitive even against the baseline model pre-trained with full supervision, whilst trained without any labels.

The use of VoxCeleb2 as well as VoxCeleb1 during training results in a further improvement on the verification performance.

Discussion: The improvement in performance can be attributed to two reasons.

First, the use of multiple negatives with the softmax function leads to more stable learning. Pairwise losses only enforce that an embedding is far from one particular negative embedding, not all negatives. Hence, at every iteration the embedding is being 'pushed' away from only a small number of randomly selected negatives but may still remain close to others, which may lead to instability in training. The proposed algorithm enforces relative similarity of the matching pair over all non-matching pairs, whilst the softmax function decides which negatives (hard negatives) should contribute most of the gradients.

Second, the benefit of the proposed method is that the task in training is the same as the task in deployment. Cross-modal biometric matching is naturally a retrieval task, where the task is to pick the matching audio to the image input. With the proposed loss function, the network is explicitly trained to do exactly this. As with most applications of deep learning, models usually perform best for the task that they have been trained for.

C. Analysis on Demographics

Face appearances and speaking styles (e.g., pitch, prosody or intonation) are influenced by a number of factors such as nationality, age, gender and any other environmental factors. People in a specific group may share similar peculiarities in the way in which they look and speak, so it is also expected to be captured in the cross-modal embeddings.

In a similar manner to [26], we conduct experiments to determine if the proposed model learns characteristic of persons beyond gender, nationality and age. Here, the negative test pairs are sampled only within the same demographic criteria. We examine which demographic category is the most influential to biometrics, then compare the performance of embedding models learnt by contrastive loss and the proposed strategy.

Dataset: Since this task is based on the trained model in previous section, VoxCeleb1 and VoxCeleb2 are used for training, and the unseen-unheard test set for testing. The test list for each of the scenario is provided by the authors of [26].

Evaluation protocol: The negative pairs are sampled from different identities but within common demographic group. Therefore, these characteristics are excluded as a cue of distinguishing between speakers. There are 'gender,' 'nationality' and 'age' in the demographic categories, and we conduct 5 experiments; holding (1) none, (2) gender, (3) nationality, (4) age, (5) all demographic categories. Lower AUC in these experiments means that the removal of these demographic information makes the cross-modal verification a more challenging task.

Results: Evaluation results are given in Fig. 8. The results are in line with [26], where gender is the most influential property compared to the other factors, and also with our earlier findings (Section IV-B) in that the proposed method consistently performs better than the baselines. In particular, the use of the larger VoxCeleb2 dataset brings a significant improvement in the challenging same gender experiment.

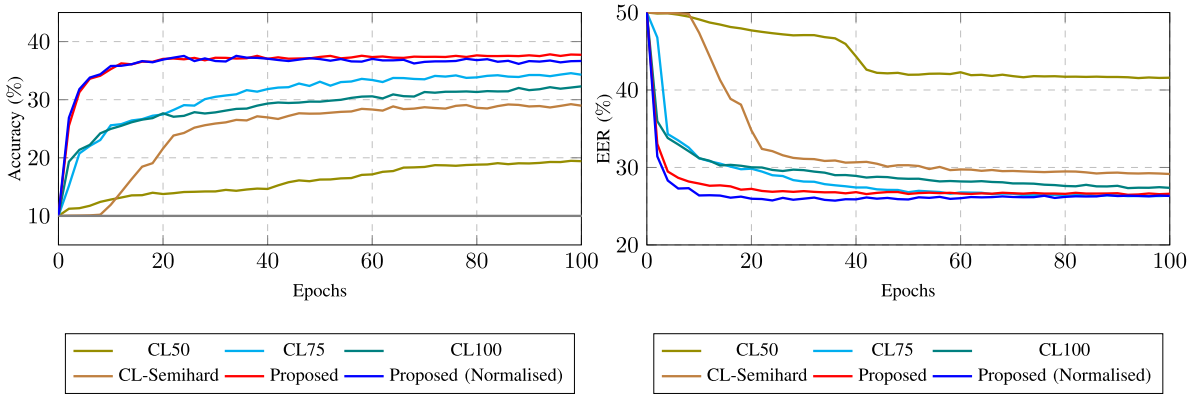
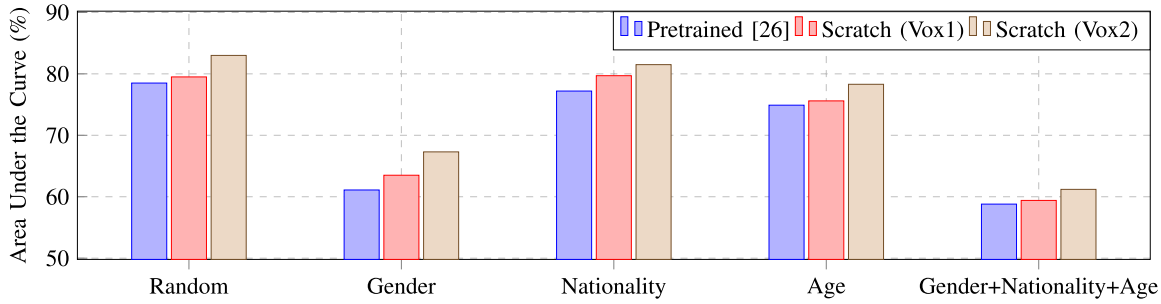


Fig. 7. Results over training epochs on the VoxCeleb2 test set. **Left**: 10-way matching classification accuracy (higher is better), **Right**: speaker verification EER curve (lower is better).



Method	Init.	Training data	Random	Gender	Nationality	Age	G+N+A
CL+CHNM [26]	Random	VoxCeleb1	63.5%	-	-	-	-
CL+CHNM [26]	Pretrained	VoxCeleb1	78.5%	61.1%	77.2%	74.9%	58.8%
Proposed	Random	VoxCeleb1	79.5%	63.5%	79.7%	75.6%	59.4%
Proposed	Random	VoxCeleb1+2	83.0%	67.3%	81.5%	78.3%	61.2%

Fig. 8. Analysis of the biometric matching task for different demographics groups. VoxCeleb1 unseen-unheard test list.

TABLE V
10-WAY FORCED MATCHING RESULTS ON THE VOXCELEB1 UNSEEN-UNHEARD LIST (HIGHER AUC IS BETTER); CL: CONTRASTIVE LOSS, CHNM: CURRICULUM HARD NEGATIVE MINING

Method	Init.	Training data	AUC	EER
CL + CHNM [26]	Random	VoxCeleb1	63.5%	39.2%
CL + CHNM [26]	Pretrained	VoxCeleb1	78.5%	29.6%
Proposed	Random	VoxCeleb1	79.5%	28.7%
Proposed	Random	VoxCeleb1+2	83.0%	25.4%

V. CONCLUSION

We proposed a new self-supervised training strategy for cross-modal matching and retrieval. The methods enable efficient training even without explicit class labels by using natural co-occurrences as the only source of supervision. In particular, our methods have shown favourable learning characteristics compared to existing pairwise methods by allowing multiple negatives to be used during training. Moreover, by re-formulating the problem as one of multi-way matching, the training procedure can mimic the deployment scenario.

Our proposed strategy was examined for audio-visual synchronisation and cross-modal biometric matching tasks, in which we have demonstrated superior performance compared to the existing state-of-the-art. We have also shown that the representation learnt through the synchronisation task is effective for a downstream task of visual speech recognition, where we have found the performance of our self-supervised embedding to match that of the model trained with full supervision.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[2] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3367–3375.
[3] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6645–6649.
[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4960–4964.
[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations*, 2015.
[6] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. Empirical Methods Nat. Lang. Process.*, 2014.
[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.
[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[9] N. Goncalves, J. Nikkilä, and R. Vigario, "Self-supervised mri tissue segmentation by discriminative clustering," *Int. J. Neural Syst.*, vol. 24, no. 01, 2014, Art. no. 1450004.
[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. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 932–940.
[11] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 2051–2060.
[12] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. 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. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2536–2544.
[15] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 649–666.
[16] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Multimedia Conf.*, 2010.
[17] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2011.
[18] Y. Chen, L. Wang, W. Wang, and Z. Zhang, "Continuum regression for cross-modal multimedia retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2012.
[19] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012.
[20] R. Arandjelović and A. Zisserman, "Objects that sound," in *Proc. Eur. Conf. Comput. Vision*, 2018.
[21] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4358–4366.
[22] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Workshop Multi-View Lip-Reading, Asian Conf. Comput. Vision*, 2016.
[23] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019.
[24] B. Korbar, D. Tran, and L. Torresani, "Co-training of audio and video representations from self-supervised temporal synchronization," 2018, *arXiv:1807.00230*.
[25] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018.
[26] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. Eur. Conf. Comput. Vision*, 2018.
[27] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," 2018, *arXiv:1805.05553*
[28] H. Bredin and G. Chollet, "Audiovisual speech synchrony measure: Application to biometrics," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 179–179, 2007.
[29] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, Nov. 2007.
[30] E. Marcheret, G. Potamianos, J. Vopicka, and V. Goel, "Detecting audio-visual synchrony using deep neural networks," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015.
[31] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *Proc. Int. Conf. Comput. Vision*, 2017.
[32] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2005, vol. 1, pp. 539–546.
[33] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4879–4883.
[34] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1857–1865.
[35] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vision Conf.*, 2014.
[36] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017.
[37] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vision*, 2016.
[38] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Comput. Vision Image Understanding*, vol. 173, pp. 76–85, 2018.
[39] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002.
[40] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proc. IEEE*, vol. 94, no. 11, pp. 2025–2044, Nov. 2006.
[41] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2009, pp. 498–505.
[42] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8427–8436.
[43] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 9299–9306.
[44] Y. Wen, M. Alismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," in *Proc. Int. Conf. Learn. Representations*, 2019.
[45] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 761–769.
[46] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Asian Conf. Comput. Vision*, 2010.
[47] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in *Proc. Odyssey*, 2010.
[48] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. INTERSPEECH*, 2017.
[49] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. INTERSPEECH*, 2018.
[50] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vision Conf.*, 2015.
[51] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 815–823.
[52] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010.

Soo-Whan Chung received the B.S. degree from Yonsei University, Seoul, South Korea, in 2016 and is currently working toward the combined M.S. and Ph.D. degrees at Yonsei University, all in electricla and electronic engineering. His current research interests include speech signal processing, source separation, audio-visual signal processing, and deep/machine learning.

Joon Son Chung received the D.Phil. in engineering science from the University of Oxford. He is currently a Research Scientist with NAVER Corporation, Seongnam-si, South Korea.

Hong-Goo Kang received the B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1989, 1991, and 1995, respectively. He was a Senior Member of Technical Staff with AT&T Labs-Research, from 1996 to 2002. In 2002, he joined the Department of Electrical and Electronic Engineering, Yonsei University, where he is currently a Professor. His current research interests include speech/audio signal processing, general signal processing, and deep/machine learning.