Does Visual Self-Supervision Improve Learning of Speech Representations?

Abhinav Shukla, Stavros Petridis *Member, IEEE* and Maja Pantic, *Fellow, IEEE*

Abstract—Self-supervised learning has attracted plenty of recent research interest. However, most works are typically unimodal and there has been limited work that studies the interaction between audio and visual modalities for self-supervised learning. This work (1) investigates visual self-supervision via face reconstruction to guide the learning of audio representations; (2) proposes two audio-only self-supervision approaches for speech representation learning; (3) shows that a multi-task combination of the proposed visual and audio self-supervision is beneficial for learning richer features that are more robust in noisy conditions; (4) shows that self-supervised pretraining leads to a superior weight initialization, which is especially useful to prevent overfitting and lead to faster model convergence on smaller sized datasets. We evaluate our audio representations for emotion and speech recognition, achieving state of the art performance for both problems. Our results demonstrate the potential of visual self-supervision for audio feature learning and suggest that joint visual and audio self-supervision leads to more informative speech representations.

Index Terms—Self-supervised learning, Representation learning, Generative modeling, Audiovisual speech, Emotion recognition, Speech recognition, Cross-modal supervision.

1 Introduction

Deep neural networks trained in a supervised manner are a popular contemporary choice for various speech related tasks such as automatic speech recognition (ASR), emotion recognition and age/gender recognition. However they are a double-edged sword by virtue of providing extremely good performance given that large scale annotated data is available, which is usually expensive. For problems like emotion recognition, reliably annotated data is also extremely scarce and even modern datasets are very limited in size. Transfer learning approaches attempt to solve this problem by domain adaptation (e.g. using supervised ImageNet pretraining for visual tasks), but even they need a large amount of annotated data for the primary supervised task and generalization is not guaranteed. Self-supervised learning is a recent and rapidly developing area of machine learning which might offer a potential solution to this problem. In this work, we present a method for visually guided self-supervised learning of speech features that outperforms baseline self-supervised methods and also outperforms fully supervised pretraining on the evaluated downstream tasks.

Self-supervision is an interesting way to attempt to combat the paucity of labeled data by capturing the intrinsic structure of the unlabelled data. The idea behind self-supervision is to find a 'pretext task / proxy task' for the network to learn that does not require any explicit labeling, but instead the data's inherent structure *provides* the labels. During training, the network is tasked with predicting these implicit labels, which could be of various kinds. For instance, predicting the next element or a randomly masked element of a known sequence given the history/context is a popular pretext task. The key idea is that the whole sequence is already available as an unlabeled data sample, and we are just choosing an intrinsic property (here the value of the element to be predicted) as the label for the proxy supervised learning problem. This 'label' is provided to us for free by the data and does not require any sort of external annotation. These pretext tasks may also model and span across multiple modalities (e.g. predicting the data or features of one modality from another). This is especially relevant in the context of speech and emotion recognition where we are interested in modeling complementary multimodal information, especially in audio and video.

In this work, we investigate self-supervised learning for audio representations. Audio representations are a cornerstone of speech and affect recognition. Most audio-related applications involve the analysis of a speech signal using either handcrafted low level descriptors or through a supervised (or fine tuned) neural network which directly predicts the labels of interest. However, self-supervised learning may offer better representations for these applications, especially in cases where labeled data is hard to come by and unlabeled audio data is readily available. We look into how self-supervision can be used to produce robust audio features.

First, we examine the state-of-the-art in self-supervised audio feature learning which we use as baselines. We then propose a novel visual self-supervised method and two novel audio-only self-supervised methods for learning audio features.

Most existing self-supervised learning approaches are unimodal. The few existing cross-modal approaches typically have some interaction between the modalities in the latent space by pretext tasks like clustering but they do not produce an intuitive interaction between the two modalities. By contrast, our work proposes audio features that are explicitly guided by lip movements and facial expressions' reconstruction (see Fig. 1). We impicitly

- *A. Shukla, S. Petridis and M. Pantic are with the iBUG group in the Department of Computing at Imperial College London. E-mail: a.shukla@imperial.ac.uk*
- *S. Petridis is also with the Samsung AI Centre, Cambridge, UK.*

Manuscript under review.

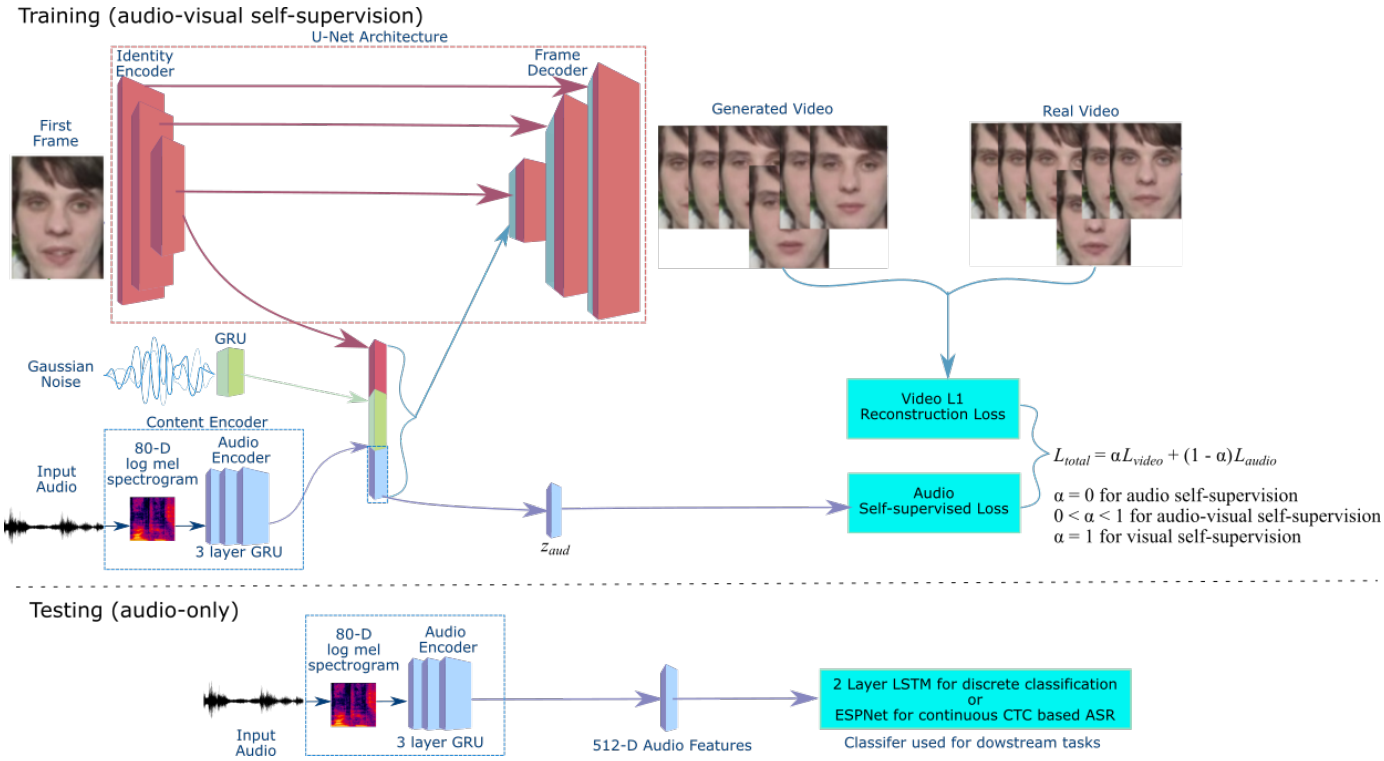


Fig. 1. An overview of our proposed model for visually guided self-supervised audio representation learning. During training, we generate a video from a still face image and the corresponding audio and optimize the reconstruction loss. An optional audio self-supervised loss can be added to the total to enable multi-modal self-supervision. During testing, we use the audio encoder to extract features for (or finetune on) downstream audio-only tasks.

capture visual information related to lip movements and facial expressions in the audio features. The visual modality is needed only during training and our audio features can be evaluated on audio-only datasets.

We summarize our research contributions as follows:

1) We investigate visual self-supervision for learning audio features. We propose a novel method for visually-guided self-supervised learning of speech representations by face reconstruction (L1). The proposed speech features, which are correlated with **lip movements** and **facial expressions** due to being driven by video generation, outperform existing audio-only self-supervision approaches for speech and emotion recognition.
2) We propose two new audio-only self-supervised methods (Odd One Out and Arrow of Time). These methods are both inspired from visual self-supervised learning and are based on temporal order verification as the pretext task. Both of them offer competitive performance on the tested datasets.
3) We combine the proposed audio-only and video-only supervision methods by multi-task learning. We find that the encoder trained in a multi-modal regime encodes richer information about the speech signal and yields the most effective representation that attains the best performance among all tested methods.
4) We show that pretraining by audio-visual self-supervision produces a better weight initialization for downstream tasks than does training from scratch. This results in faster training and convergence for a variety of hyperparameters for downstream tasks.
5) We show that the proposed visually-guided audio features are more robust for various levels of noise.

2 Related Work

To position our work with respect to existing literature and to highlight its novelty, we review prior work in: (1) Self-supervised learning, including audio, visual and cross modal methods; (2) Audiovisual speech recognition and methods that exploit both modalities for speech related tasks like emotion recognition.

2.1 Self-Supervised Learning

Self-supervised learning is a rapidly developing field in machine learning, with the promise of being able to learn useful representations from unlabeled data. Perhaps the most seminal and widespread applications of self-supervised learning have come in natural language processing. Extremely popular recent works like ELMo [1] and BERT [2] are based on predicting the next token of text based on the history or context. Self-supervised learning of visual features has also attracted a lot of research interest, whereas self-supervised learning of audio representations has received less attention so far. There have also been a few works on cross-modal self-supervised learning. We briefly survey these trends in the subsequent sub-sections.

2.1.1 Self-supervised video feature learning

There have been numerous recent works on visual self-supervised representation learning. Gidaris et. al. [3] predict rotations for unlabeled images that have been rotated by a known amount, which drives the features to encode information about the object shape and appearance. Other works try to predict the relative location of patches [4], temporal order of frames in a video [5], or audio-visual synchronization [6], [7]. BigBiGAN is a recent method proposed for adversarial self-supervised representation learning [8]. The work shows that more accurate and realistic reconstructions tend to produce better visual features for downstream tasks. Cycle consistency is also a concept that has been explored for visual feature learning [9]. DeepCluster [10] was an interesting idea which focused on clustering in the latent space based on iteratively improving labels provided by the model being trained. NoisyStudent [11] was a follow up work on a similar concept, the idea being that the predictions of the model from a previous training epoch could be used as labels for the current epoch. S4L (Self-Supervised Semi-Supervised Learning) is another recent work which combines self-supervised learning with a small amount of labeled data to learn richer representations [12]. Contrastive learning is a recent trend in self-supervised learning that is focused on separating representations of positive and negative pairs in the latent space. MoCo [13] is an important work in this area, and is based on distancing a positive pair from a large memory bank of negative examples. PIRL [14] extends this idea to produce image representations that are invariant of the chosen pretext task. A more detailed overview of self-supervised methods for visual feature learning can be found in [15] We draw inspiration from visual self-supervised learning for the learning of audio features. In this work, we apply concepts from visual self-supervision to develop two audio-only self-supervised methods (see section 4) and a cross-modal self-supervised method based on visual reconstruction (see section 3).

2.1.2 Self-supervised audio feature learning

There has also been a wave of recent work on self supervised audio-only representation learning. CPC (Contrast Predictive Coding) [16] and APC (Autoregressive Predictive Coding) [17] are similar approaches that model the next token of a speech segment given the history. Another method called LIM (Local Info Max) [18] is based on maximizing the MI (mutual information) among randomly chosen windows in an unsupervised way to learn speaker embeddings. Wav2vec [19] is also an unsupervised pre-training method used in the context of speech recognition. Self supervised audio features have also been proposed for mobile devices [20]. Another very relevant recent work is PASE (Problem Agnostic Speech Encoder) [21], which aims to learn multi-task speech representations from raw audio by predicting a number of handcrafted features such as MFCCs, prosody and waveform. SeCoSt [22] is a teacher-student self-supervised approach very similar to the ones in the visual domain that iteratively use the predictions of one epoch as the labels for the next one. Phase prediction [23] has also been proposed as an audio-based pretext task. WaveNet [24] is a generative model for raw audio waveforms that can be used for generic audio representations. There has also been a new version of CPC proposed for audio for multiple languages [25]. We compare our proposed methods with the best performing audio-only self-supervised baselines in recent literature. A detailed description of the baselines can be found in section 6 and the results can be found in section 7.

2.1.3 Self-supervised cross-modal learning

A few works also exploit the relationship between modalities, such as by predicting cyclic transitions [26], the relationship between ambient sound and vision [27], and cross-modal prediction based fusion [28]. XDC [29] extends the idea of clustering as a pretext tasks across modalities, with the cluster predictions for video coming from audio and vice versa. Piergiovanni et. al. [30] propose a method that shares representations across modalities via distillation and finds better loss functions using evolutionary algorithms. Contrastive Multiview Coding [31] is a method that projects representations of views from different modalities closer in the latent space for positive examples and further for negative examples. In this process, the encoders for each view (modality) learn useful representations. Patrick et. al. [32] extend the contrastive learning concept from MoCo [13] to a multi-modal setting. Morgado et. al. [33] combine audio-visual instance discrimination with cross-modal agreement for self-supervised learning. Zhu et. al. [34] present a detailed survey of deep audio-visual learning including cross-modal self-supervised learning. All of these works have shown that it is possible to learn robust multi-task representations from a large amount of unlabeled data that is inexpensive to obtain. We propose a novel self-supervised method based on cross-modal reconstruction to learn audio features. Our method is based on speech-driven facial reconstruction and is explained in detail in section 3.

2.2 Audiovisual Speech and Emotion Recognition

Audiovisual speech data is extremely common and the usage of complementary information from both modalities is a popular concept in many fields of research. The McGurk effect [35] was the classic example that demonstrated the audio-visual nature of human perception of speech. The visual modality contains information that offers robustness in circumstances where the audio modality may be corrupted with noise [36].

Audiovisual emotion recognition has also seen a significant amount of recent research efforts. Automatic affect recognition has a variety of applications in various fields; from detecting depression [37], to more emotionally relevant advertising [38], [39]. A lot of contemporary affect analysis approaches are based on deep neural networks that study both the visual and audio modalities [40], [41]. However a big problem in emotion recognition is the lack of reliably annotated data for large datasets, which we try to address (implicitly) in this paper.

3 Visual Self-Supervision for Speech Representation Learning

The proposed method is illustrated in Fig. 1 and is based on our prior work on visually guided speech representation

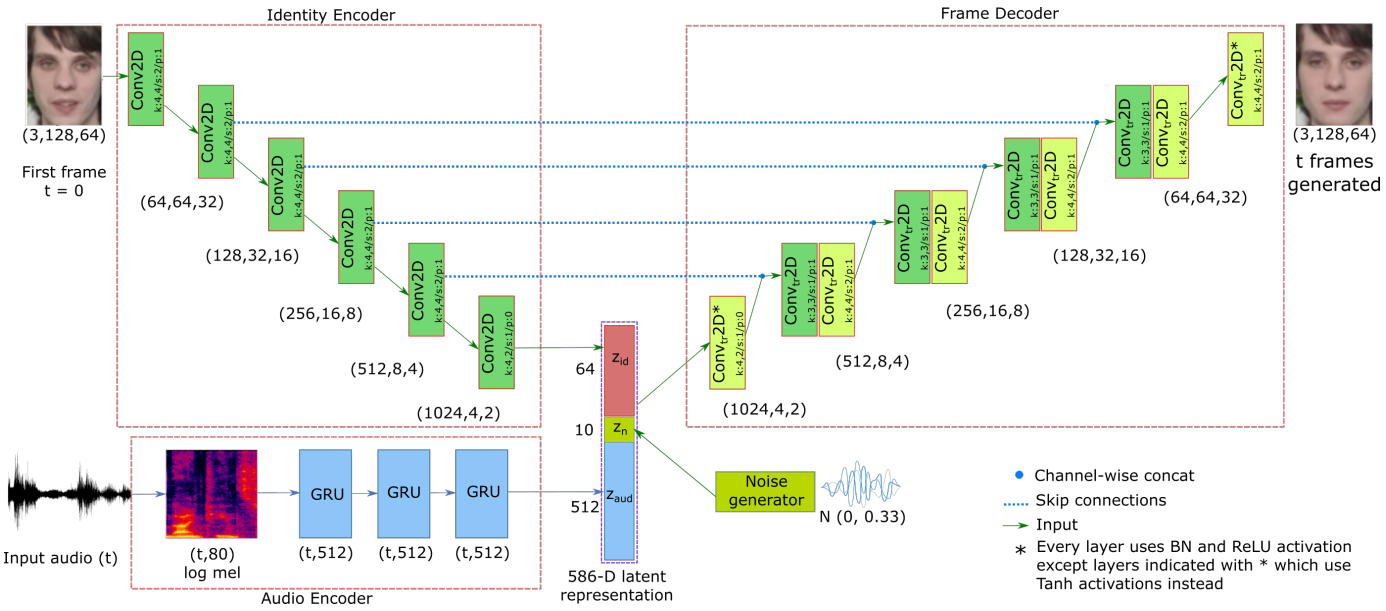


Fig. 2. A detailed illustration of our the encoder-decoder model we use for video reconstruction. From an unlabeled sample of audiovisual speech, we use the audio and the first frame of the video (*t = 0*) to generate a video with *t* frames. The model contains: (1) an identity encoder which produces a 64-D identity embedding; (2) an audio encoder which converts the input audio (t frames of 80 dimensional log mel spectrograms) into a 512-D audio embedding; (3) a frame decoder which generates video from the concatenated latent representation using transposed convolutions.

learning through speech-driven facial animation [42], [43]. The model is a temporal encoder-decoder which takes a still image of a face (frame from a 25 fps video) and an audio signal as inputs and generates video frames from these. The model itself can be conceptually divided into three subnetworks (see Fig. 1 and Fig. 2), namely the content/audio encoder (3 layer GRU), the identity encoder (6 layer 2D CNN) and the frame decoder (U-Net [44] architecture with skip connections from the identity encoder).

The architecture of the content encoder is a 3 layer GRU with log mel spectrograms as input (closely following [17]), as shown in Fig. 2. Its purpose is to convert the input audio into a latent space audio feature vector z_{aud}. Similarly, the identity encoder (see Fig. 2 top-left), which is made of 6 (Conv2D - BatchNorm - ReLU) blocks, reduces a 64x128 input image (which is the first video frame of the audiovisual speech segment) to a 64x1 feature vector z_{id}.

We also use a noise generator (see Fig. 1) capable of producing noise that is temporally coherent. A 10 dimensional vector is sampled from a Gaussian distribution with mean 0 and variance of 0.33 and passed through a single-layer GRU to produce the noise sequence. This latent representation z_n accounts for randomness in the face synthesis process (such as the generation of random sequential behaviour like blinks [45]), which leads to a more realistic facial reconstruction.

The latent representation is the concatenation of z_{aud}, z_{id} and z_n (as shown in Fig. 2). This results in a 586 dimensional embedding. This embedding then goes through the frame decoder (see Fig. 2 top-right), which is a CNN that uses strided transposed convolutions to produce the video frames. The skip connections to the identity encoder help in preserving subject identity.

An L1 reconstruction loss between a random frame from the generated video and the corresponding frame from the real video is used to train the network. The L1 loss on the pixel level is commonly used in facial reconstruction as opposed to the L2 loss which typically produces blurrier reconstructions. We use the Adam optimizer with a learning rate of 0.06 that is decayed by a factor of 0.98 every 10 epochs. Essentially, our model aims to predict the video modality (face reconstruction) given only the audio modality and speaker identity information from the first frame. In this process, the audio encoder is driven to produce useful **speech features that correlate with mouth and facial movements** (because we need to generate these lip and facial movements using only the audio information, so the features z_{aud} must encode this in order to reduce the L1 loss). After this process of visually guided self-supervised pretraining, we simply use the trained audio encoder as a pretrained model for audio-only downstream tasks. The features extracted from this model are especially interesting to evaluate on tasks like speech recognition and emotion recognition. This is because these features are explicitly trained (guided by the visual modality) to contain information related to lip movements (highly correlated with speech) and facial expressions (highly correlated with emotion).

3.1 Audio Encoder Architecture

The audio encoder (see Fig. 2 bottom-left) is a log mel spectrogram encoder (closely following [17]). The log mel spectrogram is computed with 80 frequency bins, a window width of 25ms and a stride of 10ms, which is a standard choice for processing speech signals. This (t, 80) dimensional input then goes through the encoder which is a 3 layer GRU network with each layer having a hidden size of 512 followed by a fully connected layer which converts it into a feature with dimensionality (t, 512). This specific architecture of the audio encoder with 3 GRU layers is the exact same as used in [17]. We chose this for simplicity and

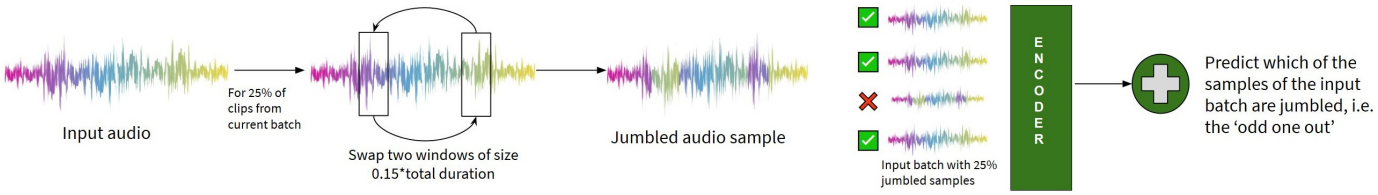


Fig. 3. An overview of the proposed Odd One Out networks for audio representation learning. The input audio is jumbled as shown for 25% of the input batch. The audio encoder is then trained on the self supervised task of predicting which clip is the 'odd one out'.

to enable direct comparison with this baseline which uses a similar audio input as us (80 dimensional log mel features).

We use the above described architecture (see Fig. 2) as the audio encoder in the proposed models in the following sections (for both visual and audio self-supervision). We then use the trained encoder to extract features from the evaluation datasets.

4 Audio Self-Supervision for Speech Representation Learning

This section introduces two audio-only self-supervised methods that we propose for speech representation learning. The concept and inspiration behind both of them is similar, which is temporal order verification for audio as a pretext task. The first proposed method is inspired by a work for video representation learning called the 'Arrow of Time' [46], and the second one by a similar work called 'Odd One Out' [5].

4.1 Audio feature learning with the Arrow of Time

The temporal order of a sequence carries a lot of potentially useful information about its structure. For video, Wei et. al. [46] proposed the Arrow of Time as a self supervised method that predicts whether a given video sequence is being played forwards or backwards. This helps the encoder that is predicting this pretext task label to learn features that correspond to object semantics and other visually correlated physical characteristics like gravity, forces etc. which may be useful for generic visual feature learning. We adapt the Arrow of Time (henceforth abbreviated as AoT) method for speech signals. The problem reduces to predicting whether a given audio clip is being played forwards or backwards. While learning to predict the task, the encoder learns useful audio features that differentiate between certain phonemes and how they sound when played forward vs backward. In order to predict the direction of the Arrow of Time, the encoder must capture useful characteristics about the phonemes themselves. In our implementation, we simply flip the temporal order of half of the sequences of an input batch (make them play backwards), and train the audio encoder with the supervised task of predicting the binary class problem (forward or backward). We use the encoder architecture described in Section 3.1.

4.2 Odd One Out networks for Audio

Odd One Out networks for video [5] are based on predicting which one out of multiple sets of ordered sequences of frames is in jumbled order (temporally incorrect order). The intuition behind such a method being able to learn useful features is very similar to that of Arrow of Time. Being able to predict temporal order should drive the encoder to learn generic useful features about the data. We adapt this idea to the audio modality in a straightforward way as well. For a given input batch of audio clips, we jumble 25% of the clips. The jumbling is performed by selecting at random two windows of a length of 15% of the total audio duration and swapping them. The encoder is then tasked with predicting which element in the input batch is the 'Odd One Out', and is optimized using cross entropy loss. Fig. 3 illustrates the training procedure for Odd One Out networks for audio representation learning. We use the same encoder architecture as before (Section 3.1).

5 Audio-visual Self-supervision for Speech Representation Learning

We combine the proposed audio and visual self-supervision methods by making the encoder jointly predict the visual self-supervision task and the audio self-supervision task. Since we used the same encoder architecture for both the visual and audio tasks, this is straightforward to accomplish. In the pipeline shown in Fig. 1 for visual self-supervision, we also use the optional prediction for the audio-only self-supervised task (either AoT or L1). This leads to two losses being calculated, one for visual and one for audio self-supervision.

The total loss L_{total} is the weighted sum of the L1 reconstruction loss from visual self-supervision L_{video} and the cross entropy loss from the audio-only self supervision L_{audio}. α is the weight factor which controls how much of the loss term comes from which type of supervision. The total loss is given by the equation:

$$L_{total} = \alpha L_{video} + (1 - \alpha) L_{audio} \quad (1)$$

We have two possible multimodal self-supervised models being trained depending audio self-supervision type, namely: L1 + AoT and L1 + Odd.

6 Datasets and Baselines

6.1 Datasets

This section introduces the various audio-only and audiovisual datasets that were used in the work either for pretraining or evaluating the baseline and proposed models. For all datasets, we divide the data into training, validation and test sets with all samples from each speaker belonging

to a particular set only. Table 1 summarizes the statistics for all the datasets used in this work.

The CREMA-D dataset [47] contains a diverse set of 91 actors who utter 12 sentences multiples times each with a different level of intensity for each of 6 basic emotional labels (anger, fear, disgust, neutral, happy, sad). We use Crema as a discrete emotion recognition evaluation dataset.

The Ravdess dataset [48] contains 1440 samples of 24 different actors who acted out two sentences with 8 different basic emotions (anger, calm, sad, neutral, happy, disgusted, surprised, fear) and two different intensity levels. We use Ravdess also as a discrete emotion recognition evaluation dataset.

The IEMOCAP dataset [49] contains dyadic conversations between 10 speakers for a total of 12 hours of audiovisual data. The discrete emotion labels comprise of 8 categories (anger, happiness, sadness, neutral, excitement, frustration, fear, surprise), however we only consider the first 4 categories for our experiments (anger, happiness, sadness, neutral). This is due to much higher inter annotator agreement for these categories, and this portion of the dataset has been similarly used in prior studies [50]. This partition also leaves us with around 6.5 hours of data instead of the original 12 hours. We use IEMOCAP as another discrete emotion recognition evaluation dataset.

The SPC (Speech Commands) dataset [51] contains 64,727 total utterances of 30 different words by 1,881 speakers. We use SPC as a speech recognition evaluation dataset.

The GRID dataset [52] contains audio-visual speech recordings of subjects in full frontal view. It has 33 speakers, each of whom speak 1000 sentences containing six words. Every sentence in the GRID dataset follows a particular format for every word: [command/colour/preposition/letter/digit/adverb]. An example sentence is "Bin blue at F 1 now". We use GRID as an ASR evaluation dataset, and use only the audio modality for WER (word error rate) evaluation.

The LRW dataset [53] is a large, in-the-wild dataset of 500 different isolated words primarily from BBC recordings. It is an audiovisual speech dataset and is thus appropriate for training our methods. We use a subset of LRW that has only nearly frontal videos (with yaw, pitch and roll restricted to a maximum of 10 degrees), in order to have a cleaner supervisory signal from the visual modality. This filtering leaves us with a total of around 40 hours of usable data. We use LRW as the self-supervised pretraining dataset for all baseline and proposed methods.

6.2 Baselines

Since the aim of our work is to yield self-supervised audio features, we compare against other baselines focusing on the same goal. The three methods we compare against are CPC [16], APC [17] and PASE [21].

Contrast Predictive Coding (CPC) [16] is a technique that tries to model a density ratio to maximize mutual information (MI) between the target signal (random raw audio window) and the context (current raw audio window). By maximizing the MI, the method can extract the underlying latent variables that the two different parts of the signal have in common.

Dataset	Train	Val	Test
GRID	31639 / 26.4	6999 / 5.80	9976 / 8.31
LRW	112658 / 36.3	5870 / 1.90	5980 / 1.90
CREMA-D	11594 / 9.70	819 / 0.70	820 / 0.68
Ravdess	1509 / 1.76	415 / 0.48	519 / 0.60
SPC	51094 / 14.2	6798 / 1.88	6835 / 1.89
IEMOCAP	3548 / 4.28	793 / 0.95	942 / 1.31

Table 1
The number of samples and duration (number / time in hours) of speech data in the training, validation and test sets of each dataset.

Autoregressive Predictive Coding (APC) [17] is similar to CPC, however the key difference is that APC directly tries to predict the immediate future part of the signal based on the history whereas CPC tries to maximize mutual information between the target (future) and the context (present). The input features for APC are 80 dimensional log mel spectrograms with a window size of 25 ms and a step size of 10 ms. The model tries to predict the log mel spectrograms for the future windows given the history.

PASE [21] is a raw audio encoder trained in a self supervised way to predict various different handcrafted features such as MFCC, prosody, waveform etc. While predicting these multiple tasks, the encoder learns a very robust and multi-task representation for raw audio that these tasks exemplify (e.g. prosody for emotion).

We also compare our methods against 39 dimensional MFCCs (13 coefficients, 13 deltas, and 13 delta-deltas) which act as baseline features used for supervised learning for audio.

7 Experiments and Results

This section presents the details of all experiments that we perform to rigorously validate our proposed method. We present all results for speech and emotion recognition from the extracted features in Table 2, for both visual and audio self-supervision for all variants of the models. We also show the results with the combination of the visual and audio self-supervision approaches using multi-task learning. We present numerous ablation studies such as the variation of model performance with change in pretraining set size and noise level. We also compare the frozen encoders (trained on the extracted features) with their finetuned and fully supervised (trained from scratch) equivalents.

7.1 Experimental Setup

We evaluate all extracted features on: (i) Discrete Emotion Recognition and (ii) Automatic Speech Recognition (ASR).

For the **emotion recognition** task, we first perform self-supervised pretraining on LRW as described, and then use the pretrained models as feature extractors on the CREMA, Ravdess and IEMOCAP datasets. Once we have these features, we then train an LSTM model for the emotion classification task. We opted to use an LSTM for simplicity, however this can be replaced by any model that can classify variable length sequences into discrete categories (such as BiGRUs, TCNs, LiGRUs [54]). For our experiments, we use a 2 layer LSTM with 256 units in each layer. The initial learning rate is 0.0001 and is decayed by a factor of 0.1

Self Supervised Methods			**Emotion Recognition**			**Speech Recognition**	
Pretraining Dataset			LRW	LRW	LRW	LRW	LRW
Evaluation Dataset			**CREMA-D**	**Ravdess**	**IEMOCAP**	**GRID**	**SPC**
Classifier for *(t, dim)* features			LSTM	LSTM	LSTM	ESPNet	LSTM
Labels			6 emotions	8 emotions	4 emotions	ASR Text	30 words
Method	Supervision	Dim.	Accuracy (↑)	Accuracy (↑)	Accuracy (↑)	WER (↓)	Accuracy (↑)
MFCC	-	39	41.50	28.32	42.06	4.7	91.06
CPC [16]	Audio	256	34.31	29.05	39.71	10.2	74.37
PASE [21]	Audio	100	43.16	30.05	42.47	5.8	89.1
APC [17]	Audio	512	41.30	34.36	41.19	5.5	87.7
AoT	Audio	512	**48.78**	**39.50**	44.17	5.6	86.45
Odd	Audio	512	**48.29**	**39.49**	**45.14**	**5.1**	**89.29**
L1	Visual	512	**51.09**	**46.05**	**46.34**	**4.5**	**90.05**
L1 + AoT	Audio+Visual	512	49.27	**45.86**	**47.38**	4.1	**92.49**
L1 + Odd	Audio+Visual	512	**53.17**	42.77	**47.91**	**3.8**	**92.28**

Table 2
Results for all baseline and proposed methods for discrete emotion recognition (on CREMA, Ravdess and IEMOCAP), and speech recognition (on GRID and SPC). All methods are used as frozen feature extractors before training a classifier on the downstream task. Results in bold indicate the best performance for a particular type of supervision during pretraining.

every 30 epochs. We train the LSTM for 100 epochs and use the checkpoint from the epoch which gives the best validation accuracy for evaluation on the test set. We pass the last hidden state of the LSTM to a linear layer with size equal to the number of target classes (6 for CREMA, 8 for Ravdess, 4 for IEMOCAP) followed by a Softmax layer with a cross entropy loss for emotion classification. This exact same process (self-supervised feature extraction + LSTM training) is followed for all the methods being compared (as shown at the bottom of Fig. 1).

For the **speech recognition** task, we use the GRID and SPC datasets to evaluate our methods. For the SPC dataset which is a spoken word classification task with 30 different possible labels, we use the exact same protocol as described for emotion recognition (self-supervised feature extraction + LSTM training). We use the same parameters and learning schedule for the LSTM. However, for the GRID dataset, we have a continuous ASR task instead of classification (i.e. we need to decode the full sentence for every utterance instead of just assigning it a class label). Thus we need to change the evaluation pipeline in order to do WER (word error rate) evaluation instead of classification. For this, we use the extracted features converted to Kaldi format and employ the ESPNet [55] toolkit for the end-to-end ASR training. We use a hybrid CTC/attention based ASR model with the default ESPNet parameters with a BLSTM encoder (as used similarly in [56]) with 320 units and location aware attention. We train the model for 15 epochs. For decoding, we use a beam search with a beamsize of 20 and a CTC weight of 0.1.

7.2 Results with Visual Self-Supervision (L1)

Our method for visual self-supervision by face reconstruction from audio is based on an L1 reconstruction loss, and is indicated as **L1** throughout the results in Table 2. For emotion recognition, irrespective of dataset, our method performs better than any audio self-supervised baseline. On CREMA, L1 achieves an accuracy of 51.09%. The best performing baseline is PASE which achieves an accuracy of 43.16%. For Ravdess, APC is the best baseline with an accuracy of 34.36%, but L1 with an accuracy of 46.05% significantly outperforms this. The same trend can be seen for IEMOCAP, with L1 again being the best performing method with an accuracy of 46.34%. For speech recognition, L1 is again the best performing method with a WER of 4.5, which is slightly better than the result attained when using MFCCs (WER 4.7). For SPC, L1 is again the best self-supervised method with an accuracy of 90.05%, which is closest to the performance by MFCCs (optimised for speech recognition) at 91.06%.

In summary, the proposed method for visual self-supervision leads to features that significantly outperform those from baseline audio self-supervised methods for both emotion recognition and speech recognition.

7.3 Results with Audio-only Self-Supervision (AoT and Odd)

Our methods for audio-only self-supervision: Arrow of Time and Odd One Out, are respectively indicated as **AoT** and **Odd** throughout the discussion of the results. For emotion recognition, AoT is the best performing audio-only self-supervised method on CREMA (achieving an emotion recognition accuracy of 48.78%), while Odd is the best on IEMOCAP (achieving an emotion recognition accuracy of 45.14%). Both jointly perform the best on Ravdess (AoT: 39.50%; Odd: 39.49%). For speech recognition, Odd is the best method both on SPC (89.29%) and GRID (WER 5.1). PASE is the closest competing self-supervised method for ASR except MFCCs.

When comparing between the two proposed methods, Odd and AoT seem to be very close in performance on emotion recognition, but Odd seems to slightly outperform AoT on speech recognition (likely due to being a more refined pretext task). Both methods outperform baselines

for audio-only self-supervision, however when compared to the L1 method using visual self-supervision, they fall short for all evaluated unimodal experiments. This leads to the observation that the proposed visual self-supervision approach yields better features than all proposed audio-only self-supervised approaches. There is also a performance gap between the proposed unimodal self-supervised methods and MFCCs for speech recognition. We attempt to bridge this gap and yield better features using a multimodal combination of the proposed methods using multi-task learning.

7.4 Results with Audiovisual Self-Supervision (L1 + AoT and L1 + Odd)

In order to determine the optimal weights for each modality for multi-task learning, we tune the parameter α (equation 1) on the validation sets of the CREMA, Ravdess and SPC datasets (introduced in section 6) for a range of values. The results for the tuning are in Table 3. From the table, we observe that the best value of α is 0.67 for both L1 + AoT and L1 + Odd. Thus, we use the models trained with this value when evaluating on the test sets in all experiments.

When comparing the results with other results in Table 2, we can see a clear improvement using audiovisual self-supervision. L1 + AoT and L1 + Odd significantly outperform all other methods in every experiment. For emotion recognition, L1 + Odd is the best-performing method on CREMA (accuracy of 53.17%) and IEMOCAP (accuracy of 47.91%), while L1 + AoT is the best-performing method on Ravdess (accuracy of 45.86%). For speech recognition, L1 + AoT is the best-performing method on SPC (accuracy of 92.49%) while L1 + Odd is the best-performing method on GRID (WER 3.8). The significant result here is for speech recognition, in which these methods outperform MFCCs, which neither unimodal method had done. This points to the presence of complementary information being encoded by the two types of supervision from the two modalities which leads to very good generalized audio representations. In summary, multimodal self-supervision methods clearly outperform any unimodal self-supervision method.

7.5 Performance in various levels of noise

In order to further rigorously validate our proposed models for robustness, we investigate the performance under various levels of noise. We create noisy versions of the CREMA and SPC datasets by adding babble noise from the NOISEX database [57], while varying the SNR from -5 dB to 20 dB in steps of 5 dB. We perform a comparison between the best performing proposed methods: (i) L1 using visual self-supervision, (ii) Odd using audio-only self-supervision, and (iii) L1 + Odd for bimodal self-supervision. We examine how the performance varies for the three methods as the level of added noise changes in the evaluation datasets.

The results for the experiments with added noise can be seen in Fig 4. For both datasets, we observe that the audiovisual combination outperforms unimodal methods. For emotion recognition on CREMA, the audio features from the clean dataset give the best performance, and there is a linear degradation of performance with the increase in noise. Visual self-supervision is more effective in almost all scenarios, which may be expected as visual features are unaffected by auditory noise and can still drive robust learning of audio features. A similar conclusion can be reached for speech recognition as well. Yet audio-visual self-supervision leads to the best audio representations across all noise labels. The results for speech recognition are also significantly more robust to noise than those for emotion recognition. We can see this from the decrease in performance with increase in noise, which is not very sharp until extremely high noise levels at -5 dB and 0 dB. We can also notice a sharper degradation for the audio-only results when compared to both the video-only and audiovisual results. This suggests that audio features obtained by visual or audio-visual self-supervision are more robust to noise compared to those obtained by audio-only self-supervision.

7.6 Performance with various sizes of the pretraining set

It is also an interesting experiment to see how model performance varies with the amount of data used for self-supervised pretraining. One of the most important advantages of a self-supervised learning approach is the ability to use an arbitrarily high amount of unlabeled data to learn a good representation. But there is still a tradeoff to be made with training time and model performance. We used a subset of the LRW dataset with a total pretraining size of 112658 samples (36.3 hours of audiovisual speech) to train our full model. For this experiment, we investigate what happens to our model if we only use a fraction of the total available data for pretraining. We use 0.2, 0.4, 0.6 and 0.8 times the dataset size for pretraining. We compare the performance between: (i) L1 for visual self-supervision, (ii) Odd for audio-only self-supervision, and (iii) L1 + Odd for audio-visual self-supervision. As before, we evaluate on the CREMA dataset for emotion recognition and the SPC dataset for speech recognition.

The results for the experiments with different pretraining set sizes can be seen in Fig. 5. For emotion recognition on CREMA, there is a steady and somewhat linear degradation in performance with the reduction in data in the pretraining. The best performance drops from an accuracy of 53.17% with the full dataset to an accuracy of 43.91% for 20% of the dataset. For speech recognition on SPC, there is a slower degradation for both visual and audiovisual methods, however there is a very sharp degradation for the audio method with lesser training data. With 20% of the training data, the performance for the audio-only method drops to an accuracy of 51.67%, which is a massive gap from the accuracy of 76.60% for the audiovisual method in the same setting. This goes on to show that the method with combined audio and visual self-supervision offers more robustness with varied amounts of pretraining data. Another observation is that a larger amount of pretraining data helps learn better features: the gain is significant initially however starts to plateau after a point (80% of the data). Fig. 5 also shows that with just 40% / 60% of the LRW dataset for pretraining, the audio-visual self-supervised methods achieve similar performance ($\pm$42% for emotion, $\pm$89% for ASR) as PASE, CPC and APC which use the full pretraining set (see Table 2). This is a very interesting result because it shows that our proposed

MTL Weight Tuning for Audio and Visual Tasks				**Emotion Recognition**		**Speech Recognition**
Pretraining Dataset				LRW	LRW	LRW
Evaluation Dataset				**CREMA-D**	**Ravdess**	**SPC**
Classifier for *(t, dim)* features				LSTM	LSTM	LSTM
Labels				6 emotions	8 emotions	30 words
Method	Video weight (α)	Audio weight ($1-\alpha$)	Dim.	Accuracy (↑)	Accuracy (↑)	Accuracy (↑)
L1 + AoT	0.17	0.83	512	46.22	38.61	88.74
L1 + AoT	0.33	0.67	512	47.91	40.18	89.28
L1 + AoT	0.50	0.50	512	51.50	43.03	90.36
L1 + AoT	**0.67**	**0.33**	512	**51.77**	**44.39**	**91.94**
L1 + AoT	0.83	0.17	512	48.93	40.40	90.79
L1 + Odd	0.17	0.83	512	48.91	42.11	89.97
L1 + Odd	0.33	0.67	512	47.48	39.81	88.39
L1 + Odd	0.50	0.50	512	50.73	43.26	90.78
L1 + Odd	**0.67**	**0.33**	512	**52.81**	**44.32**	**92.17**
L1 + Odd	0.83	0.17	512	51.17	42.41	91.31

Table 3
Comparison of different MTL weights. All results are accuracies on the validation sets of the evaluation datasets.

(a) Emotion recognition accuracy on CREMA under noise

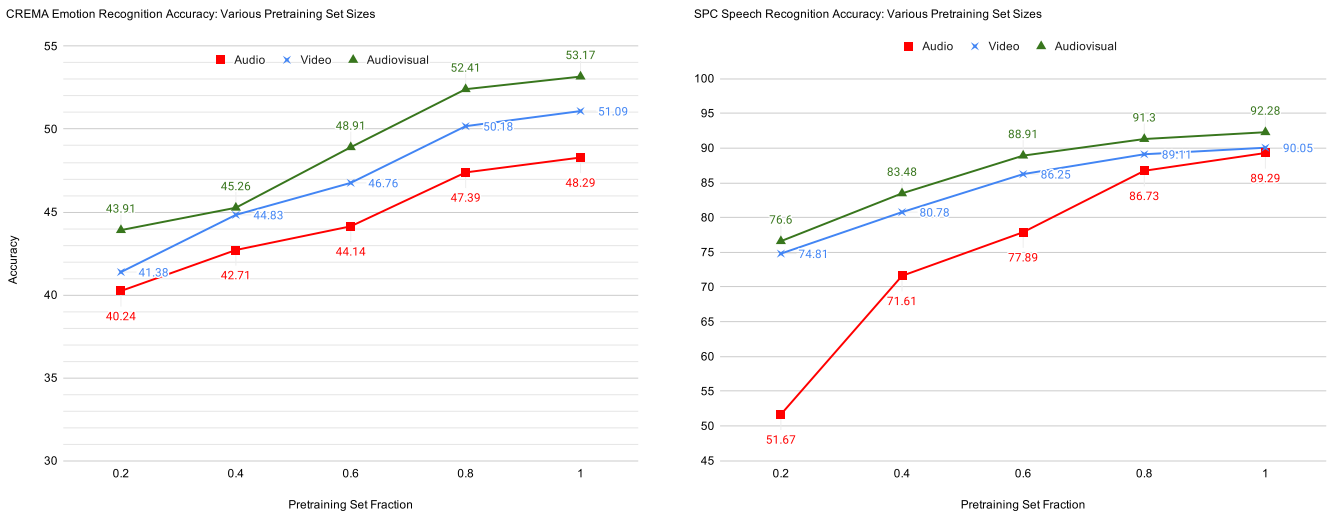
(b) Speech recognition accuracy on SPC under noise

Fig. 4. Comparison of performance of our methods trained using audio-only (Odd), video-only (L1) and audio-visual (L1 + Odd) self-supervision under various levels of artificially introduced babble noise. Best viewed under zoom.

self-supervised methods require lesser pretraining data than other tested self-supervised methods to achieve competitive performance.

7.7 Comparison with finetuned and supervised versions of encoders

All the results presented thus far were with the frozen versions of the encoders, i.e. the encoders with their fixed weights were used as feature extractors on the evaluation datasets before training a classifier. However, these encoders can also be fine-tuned to the target dataset. We present results for finetuning in Table 4. We use the weights from the L1 + Odd model as the initialization for the encoder before performing training on the target datasets in an end-to-end manner. We vary both the learning rate of the encoder (Enc LR) and that of the LSTM classifier (Cls LR). We set the Enc LR to 10e-6 or 10e-4, and the Cls LR to 10e-4 or 10e-3, which gives us 4 different sets of hyperparameters to compare. Note that the frozen version of the encoders presented in Table 2 can be interpreted to have an encoder learning rate of 0, and we present those as baseline results in Table 4 which we aim to improve upon by finetuning. We observe that finetuning is able to give us significantly better results depending on the problem setting. The best result that we get on CREMA is an accuracy of 58.90%, which is better than the 53.17% that is attained using just the frozen encoder. For SPC, we get an accuracy of 93.56% which is also the best result seen so far in all experiments. The largest observed



(a) Emotion recognition accuracy on CREMA

(b) Speech recognition accuracy on SPC

Fig. 5. Comparison of performance of our methods trained using audio-only (Odd), video-only (L1) and audio-visual (L1 + Odd) self-supervision for various sizes of the pretraining set (fraction of total). Total samples in LRW pretraining set = 112658. Best viewed under zoom.

gain comes for the Ravdess dataset, on which we get an accuracy of 64.35%, which represents a gain of nearly 20% from the frozen variation.

It is also interesting to compare our methods with a supervised version of an encoder with the same architecture trained from scratch directly on the target dataset (see 4). The only difference is the weight initialization, for which we use random initialization for each layer. We compare the supervised version to the finetuned version for the exact same hyperparameter sets (Enc LR and Cls LR). We find that for the first two parameter sets, the supervised model is not able to learn any useful features at all and attains performance close to chance. For the other parameter sets with Enc LR = 10e-4, the supervised model does converge and offers good performance. However this is still significantly worse than that obtained by the finetuned version (see 4). This is clear evidence to support that our self-supervised pretraining yields a much better weight initialization that is likely to converge for a wider variety of hyperparameters while training on downstream tasks that have smaller datasets.

We observe from Fig. 6 that training a fully supervised model from scratch without self-supervised pretraining is more susceptible to overfitting and non-convergence for certain hyperparameters. It also results in significantly slower training (as can be seen from Fig. 6). We are able to achieve much better performance by finetuning our model for every tested parameter set, despite training for only half the number of epochs.

8 Discussion

There are a number of interesting observations from the experiments.

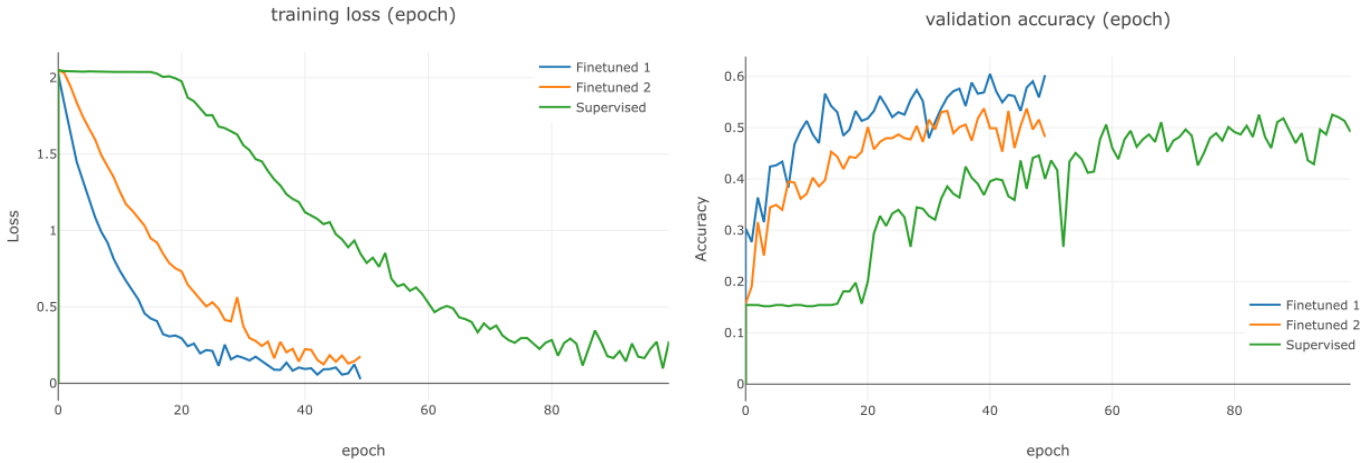
The audio-only self-supervised methods outperform the existing self-supervised baselines. There is also a clear observation that visual self-supervision is vastly superior when compared to audio-only self-supervision, both for baseline and proposed methods. This is largely due to the fact that the audio features obtained by visual self-supervision are closely related to the useful information present in lip movements and facial expressions (because they must encode this information for accurate facial reconstruction during pretraining). This property is also especially useful for emotion recognition due to the correlation between emotion and facial expression information, and for speech recognition due to the information from lip movements.

It is also clear that the models that have been trained using a combination of audio and visual self-supervision are able to encode complementary information from each modality to yield the best possible representations among all tested methods in this work. These representations are also the most robust in the presence of various levels of noise in the data, and offer the best performance independently of the size of the pretraining set. This is perhaps the most useful finding of this work, with the implication being that any problem using any sort of speech data can benefit greatly from using visual supervision from available audiovisual speech datasets to enhance the target representations.

Another very useful finding of the work is that finetuning of the pretrained audiovisual self-supervised models offers not only better performance, but faster training and convergence for a variety of hyperparameters when compared to training a fully supervised model from scratch. This could be useful in setting a strong baseline for other speech related problems and cutting down training time on downstream tasks on small datasets, which is a typical problem setting in various domains. We will make our pretrained models publicly available as to enable the community to commence further research on these problems.

8.1 Limitations

A current limitation to our work is the fact that we use a nearly-frontal subset of the LRW dataset (with yaw, pitch



(a) Training loss per epoch

(b) Validation accuracy per epoch

Fig. 6. Learning curves for the finetuned and supervised versions of the models when evaluating on the Ravdess dataset for emotion recognition. It is clear that the finetuned versions of the model start from a superior weight initialization and offer faster convergence. Finetuning also provides better overall performance with half the epochs.

Frozen vs Finetuned vs Supervised				**Emotion Recognition**		**ASR**
Pretraining Dataset				LRW	LRW	LRW
Evaluation Dataset				**CREMA-D**	**Ravdess**	**SPC**
Classifier for *(t, dim)* features				LSTM	LSTM	LSTM
Labels				6 emotions	8 emotions	30 words
Method	Enc LR	Cls LR	Epochs	Accuracy (↑)	Accuracy (↑)	Accuracy (↑)
Mel encoder (Supervised)	10e-6	10e-3	100	12.43	15.22	5.31
Mel encoder (Supervised)	10e-6	10e-4	100	17.08	13.31	11.38
Mel encoder (Supervised)	10e-4	10e-3	100	50.21	44.68	91.80
Mel encoder (Supervised)	10e-4	10e-4	100	53.19	52.68	92.11
L1 + AoT (Frozen)	0	10e-4	100	49.27	45.86	92.49
L1 + Odd (Frozen)	0	10e-4	100	53.17	42.77	92.28
L1 + Odd (Finetuned)	10e-6	10e-3	50	50.49	43.93	92.64
L1 + Odd (Finetuned)	10e-6	10e-4	50	53.17	50.67	92.76
L1 + Odd (Finetuned)	10e-4	10e-3	50	58.78	57.99	93.41
L1 + Odd (Finetuned)	10e-4	10e-4	50	**58.90**	**64.35**	**93.56**

Table 4
Comparison between frozen, finetuned and fully supervised encoders for discrete emotion recognition (on CREMA and Ravdess) and speech recognition (on SPC).

and roll restricted to a maximum of 10 degrees each) for pretraining. This leaves out a large portion of the audio-visual dataset with profile faces which could also contain useful visual supervisory signals. There are also other larger datasets like AVSpeech [58] which could potentially yield better pretrained models. Another limitation is that we have used a very simple 2 layer LSTM with 256 hidden units as the classifier of choice for our audio classification tasks. This might not be the most optimal method or configuration. However this was chosen for simplicity. Other models such as BiGRUs, LiGRUs [54] or temporal convolutional networks (TCNs) in different configurations may yield even better results. Another possible limitation is the fact that the models in our work start from log mel spectrograms instead of raw audio (as input to the audio encoder). There is a static frequency domain transformation applied to raw audio to yield the spectrogram representation, however a more refined approach might be to use a set of trainable filters (e.g. as used in SincNet [59]) instead of static Mel filters. In summary, the methods presented in this paper show the principle that visual and bimodal self-supervision lead to much better performances than full supervision from scratch. However, more refined approaches may result in even better performances than those presented here.

8.2 Future work

In this work, we have considered the interaction between the audio and visual modalities and how visual self-supervision can benefit learning of audio features. There are also other modalities that could be considered, especially the text modality. Multimodal human language is comprised

of text, audio and video combined, and developing a self-supervised model that can capture the interactions between the three can be very useful. The visual pretext task that we focused on was facial reconstruction optimized by the L1 loss. This process leads to a very realistic facial animation, however this might not be the most desirable thing in order to learn the best features. Realistic reconstruction will need to capture a lot of additional information related to fine grained visual characteristics. A lot of this information might not be useful if our end goal is simply to learn useful audio representations. Although reconstruction does give us really good performance, the question remains open to what a good alternative or additional visual pretext task might be. This work has also focused on audio features alone. It is also interesting to see how we could use audio self-supervision to guide the learning of visual speech features in an analogous way (by predicting the audio waveform from only the visual modality, like in [60]). These visual features could then be used for problems like facial affect recognition and lipreading, or even combined with our proposed audio features.

ACKNOWLEDGMENTS

Abhinav Shukla's work was supported by a PhD scholarship from Samsung Electronics UK.

REFERENCES

[1] M. Peters, M. Neumann, M. Iyyer, M.and Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv:1802.05365*, 2018.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018.
[3] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv:1803.07728*, 2018.
[4] C. Doersch, A. Gupta, and A. Efros, "Unsupervised visual representation learning by context prediction," in *ICCV*, 2015, pp. 1422–1430.
[5] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," in *CVPR*, 2017, pp. 3636–3645.
[6] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018, pp. 7763–7774.
[7] A. Owens and A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *arXiv:1804.03641*, 2018.
[8] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 541–10 551.
[9] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2566–2576.
[10] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.
[11] Q. Xie, E. Hovy, M.-T. Luong, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *arXiv preprint arXiv:1911.04252*, 2019.
[12] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 1476–1485.
[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.
[14] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," *arXiv preprint arXiv:1912.01991*, 2019.
[15] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," *arXiv preprint arXiv:1901.09005*, 2019.
[16] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[17] Y. Chung, W. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv:1904.03240*, 2019.
[18] M. Ravanelli and Y. Bengio, "Learning speaker representations with mutual information," *arXiv:1812.00271*, 2018.
[19] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv:1904.05862*, 2019.
[20] M. Tagliasacchi, B. Gfeller, F. Quitry, and D. Roblek, "Self-supervised audio representation learning for mobile devices," *arXiv:1905.11796*, 2019.
[21] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv:1904.03416*, 2019.
[22] A. Kumar and V. K. Ithapu, "Secost: Sequential co-supervision for weakly labeled audio event detection," *arXiv preprint arXiv:1910.11789*, 2019.
[23] F. d. C. Quitry, M. Tagliasacchi, and D. Roblek, "Learning audio representations via phase prediction," *arXiv preprint arXiv:1910.11910*, 2019.
[24] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.
[25] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," *arXiv preprint arXiv:2002.02848*, 2020.
[26] H. Pham, P. Liang, T. Manzini, L. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *AAAI*, vol. 33, 2019, pp. 6892–6899.
[27] A. Owens, J. Wu, J. McDermott, W. Freeman, and A. Torralba, "Learning sight from sound: Ambient sound provides supervision for visual learning," *IJCV*, vol. 126, no. 10, pp. 1120–1137, 2018.
[28] S. Petridis and M. Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 45–58, 2015.
[29] H. Alwassel, D. Mahajan, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *arXiv preprint arXiv:1911.12667*, 2019.
[30] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," *arXiv preprint arXiv:2002.12177*, 2020.
[31] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.
[32] M. Patrick, Y. M. Asano, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi, "Multi-modal self-supervision from generalized data transformations," *arXiv preprint arXiv:2003.04298*, 2020.
[33] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," *arXiv preprint arXiv:2004.12943*, 2020.
[34] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep audio-visual learning: A survey," *arXiv preprint arXiv:2001.04758*, 2020.
[35] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.
[36] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.
[37] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–7.
[38] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian, "Affect recognition in ads with application to computational advertising," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1148–1156.
[39] A. Shukla, S. S. Gullapuram, H. Katti, M. Kankanhalli, S. Winkler, and R. Subramanian, "Recognition of advertisement emotions

with application to computational advertising," *IEEE Transactions on Affective Computing*, 2020.
[40] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, pp. 1–23, 2019.
[41] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
[42] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven facial animation with temporal gans," *Proceedings of the British Conference on Machine Vision (BMVC)*, 2018.
[43] A. Shukla, K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Visually guided self supervised learning of speech representations," *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2020.
[44] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.
[45] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with gans," *International Journal of Computer Vision*, pp. 1–16, 2019.
[46] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman, "Learning and using the arrow of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8052–8060.
[47] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.
[48] S. Livingstone and F. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.
[49] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.
[50] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-Based Systems*, vol. 161, pp. 124–133, 2018.
[51] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.
[52] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.
[53] J. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, 2016.
[54] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.
[55] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Interspeech*, 2018.
[56] S. Petridis, Y. Wang, P. Ma, Z. Li, and M. Pantic, "End-to-end visual speech recognition for small-scale datasets," *arXiv preprint arXiv:1904.01954*, 2019.
[57] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.
[58] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *SIGGRAPH*, 2018.
[59] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with sincnet," *IEEE SLT Workshop*, 2018.
[60] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," *Proc. Interspeech 2019*, pp. 4125–4129, 2019.