On Combining Global and Localized Self-Supervised Models of Speech

Anonymous Authors[1]

Abstract

Self supervised learning involves learning general-purpose representations that can be useful in a variety of downstream tasks. In this work, we study the application of speech-embeddings derived from popular self-supervised learning frameworks such as wav2vec-2.0 and HuBERT over four different speech-classification tasks such as sentiment classification, command detection, emotion classification and depression detection. We distinguish between and discuss self-supervised training tasks that induce *localized* and *global* features of speech based on their temporal granularity: noting that self-supervised representation learning frameworks based on the masked language-modeling objective – such as wav2vec-2.0 and HuBERT – induce *localized* embeddings, we define a self-supervised learning framework based on SimSiam for learning global features of speech. Through our evaluations, we find that these global representations are better suited for tasks such as depression detection and emotion classification while the *localized* embeddings of speech can be very useful in tasks such as speech-command detection; we also find that our proposed model outperforms TRILL – a popular model for learning global representations. Finally, we also propose and confirm empirically that combining the global and localized representations of speech helps obtain better performance across a range of downstream tasks than each of the individual embedding methods.

1. Introduction

Self-supervised learning frameworks aim to learn general purpose representations from large amounts of data through strategically defined training tasks that do not require per-instance manual annotations. Popular self-supervised learning frameworks include: contrastive learning frameworks (Chen et al., 2020b;c; He et al., 2020; Henaff, 2020), clustering frameworks (Asano et al., 2019; Caron et al., 2020) and non-contrastive learning frameworks (Grill et al., 2020; Chen & He, 2021). In this work we will explore the use of simple siamese representation learning (SimSiam) (Chen & He, 2021), a non-contrastive representation learning framework, for speech.

As part of the self-supervised learning setup, models are trained on cleverly defined pretext tasks to obtain general purpose representations which can then be used for the downstream tasks. In the context of learning self-supervised representations of speech, we can categorize the pretext tasks as:

- **Localized** pretext tasks focus on learning segment-level representations of speech conditional on remaining contextual segments of the speech. As an example, self-supervised learning frameworks over raw speech waves that use masked prediction as the pretext task (Baevski et al., 2020; Hsu et al., 2021; Shor et al., 2020) are said to induce localized representations of speech as the task involves learning to represent and identify segments of speech conditional on the remaining segments; these localized representations of speech can be used to obtain state-of-the-art performance in automatic speech recognition (ASR) (Van den Oord et al., 2018; Baevski et al., 2020; Hsu et al., 2021). In summary, examples of localized pretext tasks for speech include: predicting the content of the unseen regions (Chung & Glass, 2020; Liu et al., 2020), language model-style pre-training (Chung et al., 2019), predicting discrete targets of masked regions (Baevski et al., 2019; 2020; Hsu et al., 2021), combining self-supervised learning with adversarial training (Baevski et al., 2021).
- **Global** pretext tasks focus on learning global representations of entire speech wave. For example, (Saeed et al., 2021) proposes to contrastively learn representations of mel-spectrograms derived from speech segments – on the other hand, wav2vec-2.0 contrastively learns representations of speech segments conditional on remaining segments of the raw speech. In a similar

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022. Do not distribute.

vein, (Niizumi et al., 2021) proposes a self-supervised learning strategy based on BYOL (Grill et al., 2020) for learning representations of spectrograms derived from speech segments without requiring the use of negative samples (i.e. without contrastive learning). Some other examples of self-supervised learning frameworks based on global pretext tasks (Chen et al., 2021; Shor et al., 2020; Jiang et al., 2020; Shor et al., 2021) are usually 1) applied to mel-spectrograms; and 2) use contrastive learning frameworks which rely on negative samples, thus requiring large batch sizes to train the models.

Intuitively, localized features of speech effectively represent fine details of speech segments conditional on remaining speech segments – the state-of-the-art performance in ASR obtained by the use of localized features support this intuition. On the other hand, global features of speech represent more temporally stable elements of speech and as we show in our results, these global representations of speech are better suited than localized representations to downstream tasks such as emotion classification and depression detection. In fact, TRILL(Shor et al., 2020) was proposed to learn such non-semantic representations of speech. Accordingly, the localized representations of speech depend on the speech duration and change quickly throughout a speech sample, whereas global representations do not. In general, a global representation of speech can be obtained from the localized features of speech – such as those obtained from BERT-style training – by averaging across the constituent speech segments. We note that we can draw a parallel to text-representation learning: for example, BERT training over text corpora learns context-sensitive embeddings of words whereas some downstream tasks require sentence embeddings; see (Reimers & Gurevych, 2019) for a novel fine-tuning method over BERT for deriving sentence embeddings from pre-trained BERT.

In this paper, we will propose SimSiam-speech, a simple siamese self-supervised learning framework for speech to learn global features of speech. We show that these global features achieve better performance than localized features on some downstream tasks. We will also show that combining these global features with localized features will achieve improved performance on various downstream tasks.

The main contributions of this work are as follows:

- We introduce a SimSiam-based framework for speech (SimSiam-Speech): Most previous self-supervised learning work on speech has been based on frameworks such as masked prediction similar to BERT (Kenton & Toutanova, 2019), simple contrastive learning (Chen et al., 2020b) and bootstrap your own latent (Grill et al., 2020). In this work, we have extended the SimSiam framework for speech representation learning.
- Building on the intuition that global and localized features of speech have different temporal granularity, we propose to combine these two representations to obtain better performance on downstream tasks. In our evaluations, we report on and analyse the application of localized and global – both, independently and together – self supervised representations of speech on a diverse set of speech classification tasks.

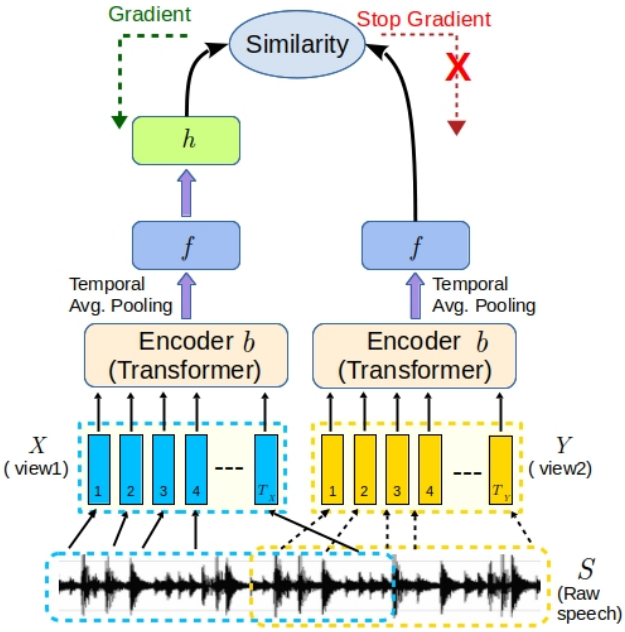

Figure 1. Outline of the proposed SimSiam framework for speech. Blocks with the same color have weight sharing

2. Proposed Approach

SimSiam Framework for Speech builds on top of SimSiam (see Figure 2 in Appendix A) and is shown in Figure 1: given a raw speech waveform S as input, two random augmented sequences of speech segments – X and Y – are used to enforce invariance across a sequence of speech augmentations described below. We obtain the representations of speech by average pooling the encoder outputs across the time steps; accordingly, these representations are used as input to the Projector f. We refer to these representations as global representations of speech and apply them in downstream speech classification tasks.

Given a raw speech waveform S with sampling frequency f_s, we obtain a sequence of segments $s_i \in \mathbb{R}^d$ such that s_i spans d/f_s seconds – in other words, each segment s_i contains d waveform amplitudes and the waveform S containing D amplitude samples gets partitioned into $T = D/d$ segments. We generate the two augmented views by first par-

titioning the sequence of segments $\{s_i\}_{i\in[1,T]}$ into two contiguous sequences $X = \{x_i\}_{i\in[1,T_X]}$ and $Y = \{y_i\}_{i\in[1,T_Y]}$ with an overlap of 50-80% segments as shown in Figure 1. The duration of the two augmented views X and Y is in the range of 2.0-2.5 seconds, and T_X and T_Y can be different. We can now define the outputs obtained after applying the encoder b (i.e. the transformer backbone), projector f and predictor h as $z_X = h(f(\hat{b}(X)))$ and $z_Y = \text{stopgrad}\{f(\hat{b}(X))\}$ where $\hat{b}$ refers to the average-pooled outputs obtained after applying the encoder b. The training objective involves maximizing the cosine similarity between z_X and z_Y as shown:

$$s(X,Y) = \frac{z_X}{\|z_X\|_2} \cdot \frac{z_Y}{\|z_Y\|_2}, \tag{1}$$

where $\|\cdot\|_2$ refers to l_2-norm. Finally, to have a symmetric loss as defined in (Chen & He, 2021), the loss function is defined as:

$$L = s(X,Y) + s(Y,X). \tag{2}$$

Augmentations: For each of the sub-sequences X and Y, we apply the following sequence of augmentations: Gaussian Noise, Shuffling, Masking, and adding silence segments. For details, check Appendix refaugs

Encoder network b: We use a multi-layer Transformer architecture (Kenton & Toutanova, 2019) as the encoder network. The transformer model consists of 12 transformer-encoder layers, each containing 12 attention heads. Each segment (i.e. x_i and y_i) is projected to a 768-dimensional vector before passing as input to the transformer; for example, the input and output of the transformer for sequence X is of shape $T_X \times 768$. The non-linear projection networks used within the transformer consist of one hidden layer having 2048 units with ReLU as the non-linearity. Finally, we use the standard self-attention (Vaswani et al., 2017) as the Multi-Head-Attention module with layer-norm normalization.

Projector (f) and Predictor (h) networks: We follow the architecture of the projector and predictor as defined in (Chen & He, 2021). The projector network takes a 768-dimensional global representation as input and has 2048-dimensional output vector while the prediction network has both input and output dimensions equal to 2048: the projector has two hidden layers, each having 2048 ReLU units and batch normalization (BN) (Ioffe & Szegedy, 2015) in between; the prediction network has one hidden layer with 512 ReLU units and BN.

Pre-training dataset: For a fair comparison with the previous self-supervised learning models, our SimsSiam-speech model is pre-trained using the LibriSpeech dataset (Panayotov et al., 2015). LibriSpeech dataset derived from public domain audio books contain 960 hours of speech data collected from 2338 different speakers (1128 female and 1210 male speakers). In this work, we discard utterances with duration less than 3 seconds as we select two speech segments each of duration ranging from 2.0-2.5 seconds, with an overlap of 50-80% in duration, to train our models.

Pre-training Details: Input speech is sampled at $f_s =$ 16kHz. We take raw speech waveform as input and use $d =$ 1000 for creating segments. Our models are pre-trained with a batch size of 480 using Adam optimizer with a learning rate of $3e^{-4}$ and a weight decay of $5e^{-5}$. Furthermore, the learning rate has a cosine decay schedule (Chen et al., 2020a).

3. Experiments

3.1. Downstream Tasks

We compare and evaluate the representations extracted from the pre-trained SimSiam-speech model on a diverse set of downstream tasks to test different aspects of the representations. The downstream tasks are summarized in Table 2 (In Appendix B) and are as follows:

1. **Sentiment classification:** We use CMU Multimodal Opinion-Level Sentiment Intensity (CMU-MOSI) (Zadeh et al., 2016) and (2) CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) (Zadeh et al., 2018) datasets to evaluate the pre-trained models for the task of sentiment classification.
2. **Low-vocabulary speech recognition:** We use speech commands (Warden, 2018) dataset which contains short spoken commands for the task of low-vocabulary speech recognition.
3. **Speech emotion recognition:** We use IEMOCAP (Busso et al., 2008) dataset for the downstream task of speech emotion recognition.
4. **Depression detection:** DAIC-WOZ (Gratch et al., 2014) dataset is used for the downstream task of depression detection.

3.2. Models for Comparison

We compare the performance of the pre-trained SimSiam-speech model with various publicly available SOTA speech representation models i.e, Mockingjay (Liu et al., 2020), VQ-wav2vec (Baevski et al., 2019), Wav2vec-2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), TRILL (Shor et al., 2020). Note that VQ-wav2vec, wav2vec-2.0 and HuBERT takes raw speech waveform as input whereas Mockingjay and TRILL takes mel spectrogram as input. While Mockingjay, VQ-wav2vec, wav2vec-2.0 and HuBERT learn

Table 1. Performance (in terms of accuracy) on different downstream tasks by using the representations obtained from different pre-trained models. Details of tasks: MOSI and MOSEI for sentiment classification, Speech commands for command detection, IEMOCAP for emotion classification, and DIAC-WoZ for depression detection

Model	MOSI	MOSEI	Speech Commands	IEMOCAP	DAIC-WOZ
COVAREP	46.6	52.9	48.2	48.2	53.3
OpenSMILE	51.3	58.1	39.8	51.6	56.8
MockingJay	56.3	69.1	87.4	52.4	58.7
VQ-wav2vec	54.1	68.2	86.7	51.6	60.4
Wav2vec-2.0	60.8	70.4	89.6	55.6	61.9
HuBERT	61.9	71.1	90.2	54.9	61.3
TRILL	56.8	63.5	78.5	61.7	64.1
SimSiam-S (Ours)	59.0	65.8	85.1	65.3	66.4
Wav2vec-2.0 + HuBERT	62.3	71.3	90.5	56.5	62.2
Wav2vec-2.0 + TRILL	62.2	70.7	90.3	64.9	68.2
HuBERT + TRILL	63.7	71.4	90.9	64.2	67.8
TRILL + SimSiam-S	59.3	65.5	85.3	65.5	66.9
Wav2vec-2.0 + SimSiam-S	64.8	71.2	91.1	**67.6**	**71.2**
HuBERT + SimSiam-S	**65.9**	**71.8**	**91.4**	67.1	70.3

localized features, TRILL learns global features. Further, we also use the openSMILE (Eyben et al., 2010) and COVAREP (Degottex et al., 2014) features as a baseline. See Appendix D for details.

3.3. Evaluation Details

For each downstream dataset, we freeze the weights of the pre-trained models, and extract representations from the encoder for the downstream task. We apply a non-linear projection (i.e. linear projection followed by non-linearity) to these representations and then pass it as input to the softmax classifier. The number of hidden units are selected based on the performance on the validation set. For the case of combining the representations extracted from different pre-trained models, we simply concatenate the non-linear projections applied to the representations before passing it as input to the softmax layer.

3.4. Results

We provide results obtained by performing 5-fold cross validation on each dataset, unless otherwise mentioned.

Table 1 shows the performance of the proposed SimSiam-Speech (SimSiam-S) compared to previous self-supervised models across four diverse set of downstream tasks. For the downstream tasks of emotion recognition and depression detection, proposed SimSiam-S and TRILL performs better than the other pre-trained models, with SimSiam-S outperforming TRILL. For the downstream tasks of sentiment classification and low-vocabulary speech recognition, HuBERT and wav2vec-2.0 performs better. This shows that the models learning global representations (SimSiam-S and TRILL) perform better on some tasks while models learning localized representations (HuBERT and wav2vec-2.0) perform better on other tasks. We can also observe in Table 1 that combining global and localized representations (HuBERT + SimSiam-S and wav2vec-2.0 + SimSiam-S) achieve significant improvements in performance compared to the case of combining two global (wav2vec-2.0 + HuBERT) or two localized representations (TRILL + SimSiam-S). For all the downstream tasks, combining SimSiam-S with global representations (HuBERT or wav2vec-2.0) achieves the best performance.

4. Conclusion

In this work, we discuss the application of self-supervised embeddings to speech classification tasks as the downstream tasks. In doing so, we highlight the distinction between self-supervised learning frameworks that favor *global* versus *localized* embeddings of speech: in this study, pretext tasks which operate on smaller speech units are said to yield localized embeddings, while those operating on entire speech sequences are said to yield global embeddings. We define a self-supervised learning framework for learning global features of speech and propose combining both global and localized features of speech for obtaining more robust downstream performance. Our empirical experiments confirm that global embeddings of speech capture certain paralinguistic elements and can be particularly useful in tasks such as emotion classification and depression detection, while localized embeddings of speech can be effective in tasks such as sentiment classification and speech-command detection. Importantly, we find that *combining* the global and localized representation of speech leads to better performance across a range of tasks than each of the individual embedding methods alone.

References

Asano, Y., Rupprecht, C., and Vedaldi, A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2019.

Baevski, A., Schneider, S., and Auli, M. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*, 2019.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 34, 2020.

Baevski, A., Hsu, W.-N., Conneau, A., and Auli, M. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34, 2021.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., et al. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. *arXiv preprint arXiv:2110.05752*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Chung, Y.-A. and Glass, J. Improved speech representations with multi-target autoregressive predictive coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2353–2358, 2020.

Chung, Y.-A., Hsu, W.-N., Tang, H., and Glass, J. R. An unsupervised autoregressive model for speech representation learning. In *INTERSPEECH*, 2019.

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 960–964. IEEE, 2014.

Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. ACM conference on Multimedia*, pp. 1459–1462, 2010.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3123–3128, 2014.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Jiang, D., Li, W., Cao, M., Zou, W., and Li, X. Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. *arXiv preprint arXiv:2010.13991*, 2020.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., and Lee, H.-y. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6419–6423. IEEE, 2020.

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., and Kashino, K. Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Saeed, A., Grangier, D., and Zeghidour, N. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3875–3879. IEEE, 2021.

Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., de Chaumont Quitry, F., Tagliasacchi, M., Shavitt, I., Emanuel, D., and Haviv, Y. Towards learning a universal non-semantic representation of speech. *Proc. Interspeech 2020*, pp. 140–144, 2020.

Shor, J., Jansen, A., Han, W., Park, D., and Zhang, Y. Universal paralinguistic speech representations using self-supervised conformers. *arXiv preprint arXiv:2110.04621*, 2021.

Van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*. NeurIPS, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Warden, P. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31:82–88, 2016.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. 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)*, pp. 2236–2246, 2018.

A. Background

The SimSiam framework (Chen & He, 2021) is an extension of BYOL (Grill et al., 2020) for learning self-supervised representations of images. Like other recently proposed self-supervised learning frameworks, SimSiam induces representations invariant over the set of pre-defined data augmentations. The core components of the SimSiam framework are as shown in Figure 2: given an input image x, the two random augmented views – x_1 and x_2 – of the input x are first passed through a series of Encoder-Projector-Predictor, and Encoder-Projector networks respectively; the training objective involves maximizing the cosine similarity between these final representations. The predictor and the stop-gradient are essential to avoid model collapse, which we will also show in the ablation studies. Finally, the encoder representations are used as the general-purpose representations of images in downstream tasks.

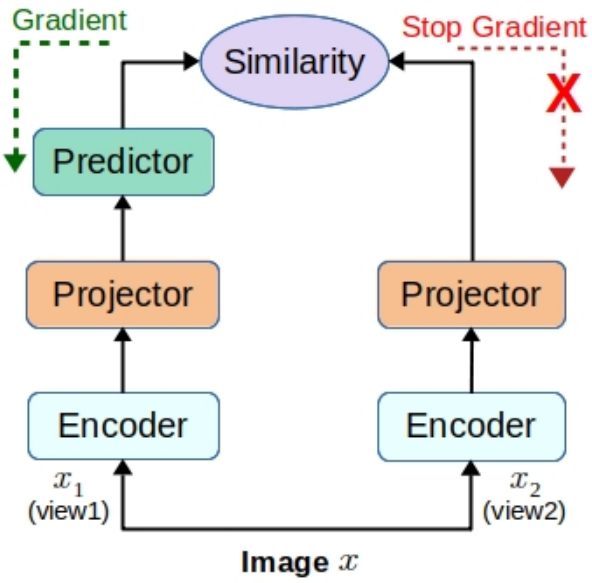

Figure 2. Outline of the SimSiam framework initially proposed for images (Chen & He, 2021). Blocks with the same color have weight sharing.

In summary, SimSiam is a simplified self-supervised learning framework which has been shown to achieve SOTA performance on image downstream tasks without relying on negative sample pairs, large batch sizes and momentum encoder. In this paper, we adapt the SimSiam framework to learn self-supervised global representations of speech.

B. Downstream Task Details

Table 2. Details of the datasets used for downstream tasks. #Class, #Sample and Avg. Len(s) refer to the number of classes, number of samples and average length in seconds, respectively

DATASET	TASK	#CLASS	#SAMPLE	AVG. LEN(S)
MOSI (ZADEH ET AL., 2016)	SENTIMENT	2	2199	4.2
MOSEI (ZADEH ET AL., 2018)	SENTIMENT	2	23453	7.28
SPEECH-COMMANDS (WARDEN, 2018)	COMMAND	12	100,503	1.0
IEMOCAP (BUSSO ET AL., 2008)	EMOTION	4	5531	4.5
DAIC (GRATCH ET AL., 2014)	DEPRESSION	2	219	960

C. Augmentations

Following augmentations are applied to the speech sequences:

1. **Gaussian Noise**: To each segment, apply additive white Gaussian noise (AWGN) with SNR between $0 - 10$ db and then normalize the amplitude of the resulting segment using min-max normalization.
2. **Shuffling**: Randomly select and shuffle $20 - 40\%$ of the segments.
3. **Masking**: Randomly select $20 - 40\%$ of the segments and replace the segments with an $\mathbb{R}^d$ vector sampled uniformly from the hypercube $[0.9, 1.1]^d$.
4. **Add silence segments**: Select N random locations and insert silence segments whose $\mathbb{R}^d$ vector has all components set equal to the lowest amplitude of the original speech waveform S. We choose N to be equal to 10% of the sequence length (i.e. T_X or T_Y).

D. Models for comparison

- **Mockingjay (Liu et al., 2020):** Mockingjay, a transformer-based model, was trained using BERT-like loss to predict masked frames.
- **VQ-wav2vec (Baevski et al., 2019):** VQ-wav2vec uses online k-means clustering to quantize the dense representations for training a deep convolutional network with a BERT-like loss.
- **Wav2vec-2.0 (Baevski et al., 2020):** Wav2vec-2.0, a transformer-based model, was trained to predict the discrete representation of the masked frames using a BERT-like loss.
- **HuBERT (Hsu et al., 2021):** HuBERT, an extension of wav2vec-2.0, utilizes an offline clustering step to provide aligned target labels to train transformer model using BERT-like loss.
- **TRILL (Shor et al., 2020):** TRILL, a transformer-based model, was trained on mel spectrograms using a triplet loss.

E. Ablation Studies

We conducted ablation studies to analyze the importance of stop-gradient, projector and predictor networks when SimSiam framework is extended to speech–which is a sequential data. We trained different SimSiam-Speech models by discarding one component at a time. All the models were pre-trained using the 960 hours of LibriSpeech dataset. It can be observed from Table 3 that all the three components i.e., projector, predictor and stop-gradient are required to avoid the model collapse. Even discarding a single component will lead to the model collapse. These observations are in agreement with those reported in (Chen & He, 2021).

Table 3. Ablation study results. Proj., Pred., SG refer to projector network, predictor network and stop-gradient, respectively. Results obtained for different SimSiam-Speech models trained by discarding one of the component

PROJ.	PRED.	SG	SPEECH COMMANDS	IEMOCAP
✗	✓	✓	6.8	19.1
✓	✗	✓	7.3	18.5
✓	✓	✗	6.5	20.8
✓	✓	✓	85.1	65.3