Data Augmentation For Wav2Vec2

Russell Wang
University of California, Berkeley
russell.wang@berkeley.edu

Eason Wang
University of California, Berkeley
esn.w@berkeley.edu

Abstract

Machine learning models often require a massive amount of data to perform well, but data acquisition and labeling can be quite expensive and time-consuming. Several architectures have been proposed to address these issues, and one of them is Wav2Vec2 in the speech domain. Wav2Vec2 is a self-supervised learning method that achieves impressive word error rate for speech recognition with only 10min labeled speech data. Although the architecture has already employed an advanced data augmentation technique during the fine-tuning stage, we investigate two additional data augmentation approaches to further improve the Wav2Vec2 performance in the low data regime. One of them is employing standard data augmentation techniques including speed adjustment, pitch adjustment and random noise. The other one is to simulate speech sound that as if the speaker is wearing a mask when uttering.

1 Introduction

Data-driven approaches often require a massive amount of data to perform well. However, data labeling is quite expensive and time-consuming. Besides, in some domains, data are hard to acquire so it is desirable to have techniques that deal with data scarcity in machine learning. Several approaches have been proposed such as self-supervised learning and semi-supervised learning as well as sophisticated data augmentation techniques. Self-supervised learning pretrains on a large unlabeled dataset to learn useful features before learning specific tasks. Recently, a state-of-the-art self-supervised speech recognition architecture, Wav2Vec2 [1] has been proposed, which can achieve 4.8/8.2 WER on the Librispeech clean/other test sets with unsupervised pretraining and only fine-tuning on 10min of labeled data. Although the architecture has already employed an modified version of SpecAugment [13], an advanced data augmentation technique, during the fine-tuning stage, there is still potentially room for improvement by using a few more data augmentations in the time domain.

In this project, we investigate two approaches of data augmentation to improve Wav2Vec2 further, especially in the low data regime. First, we experiment with standard data augmentation techniques in speech, including random speed adjustment, random pitch adjustment and random noise addition. In order to make the resulting audio still sound real, we only use small distortion factors. The second approach is to transform clear speech when the speaker is not wearing a mask to speech that as if the speaker is wearing a mask (we will call this "masked speech" later throughout the paper). People often lose fricatives or other features when they speak under masks, but the speech is still understandable and does not lose the original meaning. Therefore, we consider that as a valid and promising way to augment speech data. We find that using each data augmentation technique separately will boost the model performance, but incorporating all of them will cause the performance to degrade. This will be further discussed in later sections.

Fall 2021 EE225D Project, please see https://github.com/rwang97/ee225-project for details.

2 Related Work

Self-supervised Learning. Self-supervised learning is one of the few ways to address data scarcity issues. The general algorithm of self-supervised learning is to firstly train on a large unlabeled dataset via metric learning or contrastive learning framework to learn useful features, which will be used to fine-tune on a new dataset that is often small in size. In the image domain, one popular line of work is SimCLR [4], which proposes a framework for learning visual representations using contrastive learning. Similarly in the speech domain, Wav2Vec [12] takes raw audio as input and produces a representation that can be fed into a speech recognition system. The model is trained with a contrastive loss that tries to distinguish a true future audio sample from negatives. A more recent research Wav2Vec2 [1] proposes a framework that masks the speech input in the latent space and performs a contrastive task that is defined over a quantization of the jointly learned latent representations. In our project, we will use WavVec2 as our baseline to test out our approaches.

Semi-supervised Learning. Semi-supervised learning is another approach to mitigate data scarcity. Consistency regularization and pseudo labeling are in general the two approaches to semi-supervised learning. Framworks such as ReMixMatch [2] and FixMatch [15] use either one of the approaches or a combination of them to achieve state-of-the-art performance in image recognition. As for speech, a recent work [19] also uses semi-supervised learning to reach incredible word error rate in speech recognition. It uses noisy student training with SpecAugment using pretrained Conformer models.

Data Augmentation. Data augmentation is a different technique which does not use unlabeled data at all. Instead, it generates an augmented version of the existing labeled data as if the augmentation is coming from real data distribution. In this way, we can generate artificial data to train machine learning models. Data augmentation has been explored in great depth in images. Besides the standard ones such as flipping, cropping, brightness adjustment and color adjustment, many sophisticated data augmentation techniques can be used such as Cutout [8], AutoAugment [5], RandAugment [6], CTAugment [8]. Data augmentation in speech is quite different from that in the image domain. One way is to augment artificial data for low-resource speech recognition [14]. Another proposed way is to apply speed perturbation on raw audios [10]. More recently, SpecAugment [13] was proposed to combine augmentation policies including warping the features, masking blocks of frequency channels, and masking blocks of time steps. In our project, we will investigate simple techniques such as speed perturbation and pitch adjustment, as well as masked speech transformation, which we believe no one has worked on before.

3 Methodology

3.1 Standard Augmentation

Speed Adjustment. The first augmentation we experiment is randomly adjusting the speed of the raw audio. We introduce a factor range of $(0.75, 1.25)$ meaning that the audio can be slowed down by a factor of 0.75 or sped up by a factor of 1.25, and everything in between is also valid.

Pitch Adjustment. We also apply random pitch adjustment to the raw audios. Similarly we have a range of $(-3, 3)$ meaning that the audio can be randomly lowered by 3 semitones to randomly increased by 3 semitones. The reason why we do not choose a higher number is that we observe the pitch is either too high or too low, and does not resemble real data at all.

Random Noise. Random noise is sampled from standard normal distribution and will be added to the raw audio. We apply a scaling factor that is a random number sampled in (0.0001, 0.0003) to the noise as we do not want to make the sound too noisy.

Result. We apply the three augmentations altogether instead of only applying one to the raw audios because we want more variations of augmented data. The generated speech sounds real with perturbed speed, scaled pitch and small noise, so we can safely use them in training.

3.2 Masked Speech Generation

Data. We have found two research papers [3, 18] with datasets on speech under masks. In [3], the dataset was recorded by a female English native speaker in five mask conditions, including no mask, under a surgical mask, under a cloth mask without filter, under a cloth mask with filter, and a cloth mask with transparent plastic window. We processed this data by replicating no-mask data four times in order to create one-to-one correspondence between no mask and each type of mask condition. There was some missing data but we had 622 pairs of samples in total. The second dataset comes from [18], which investigates three mask conditions: no mask, transparent mask, and disposable face mask and two speaking styles: conversational speech and clear speech. We used the same strategy to pair the audios and eventually have 480 pairs of samples.

Besides the two public dataset, we also recorded our own masked and unmasked speech using CMU arctic database. We have 330 pairs of recordings and in total we have 1432 pairs of speech with mask and without mask.

Data Processing. In order to train our network, which will be discussed next, we first need to resample our audio recordings to 16kHz, and more importantly we need to align masked and unmasked audios. These pairs of audios are recorded separately, so even though they are speaking the same sentence, the tones and intonations are impossible to be exactly the same. Therefore, we try to align each sound unit so that the model can calculate the loss correctly. We use dynamic time warping (DTW) to achieve this alignment, and we aligned longer samples to shorter samples to prevent introducing artifacts if we instead lengthen shorter ones. Since it is costly to do DTW in the time domain, we firstly extract features of mel frequency cepstral coefficients (MFCCs) from the two audios, and then use librosa [11] library to perform DTW and find the optimal warping path, which is then used to warp the longer audios to align with the shorter ones.

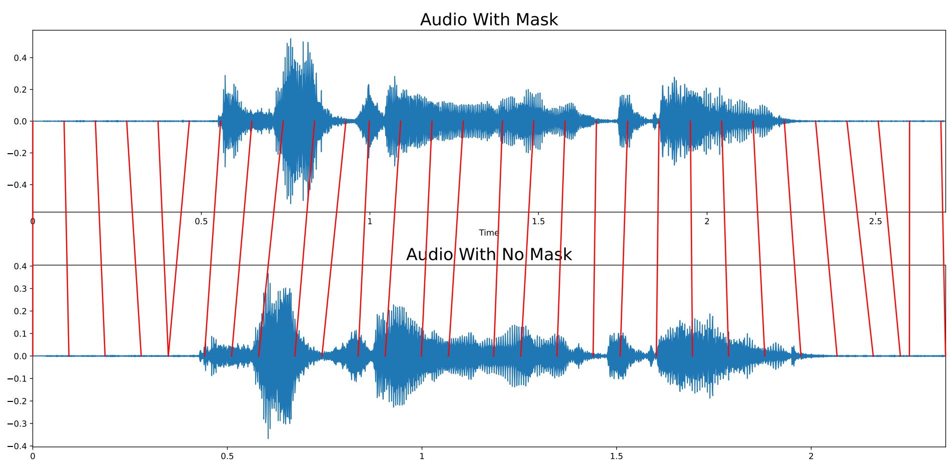

Figure 1: Example of Audio Alignment

Model. To transform unmasked speech to masked speech, what we need is a network that takes in an audio and outputs another audio, and ideally the input and output are relevant and similar. We have looked into the literature and found the area of speech enhancement and speech denoising to fit our goal. Specifically, we looked at *Real Time Speech Enhancement in the Waveform Domain (DEMUCS)* [7], which takes in a speech with background noise and outputs the same speech with noise removed.

$$L = \frac{1}{T}[\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_1 + L_{\text{stft}}(\boldsymbol{y}, \hat{\boldsymbol{y}})] \tag{1}$$

DEMUCS is based on an encoder-decoder architecture with skip-connections and the loss function is L1 loss over the waveforms plus a multi-resolution STFT loss over the spectrogram magnitudes,

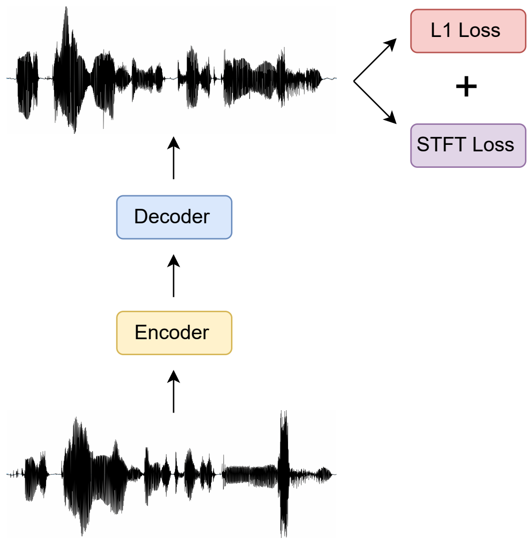

Figure 2: General encoder-decoder architecture for DEMUCS

which was originally proposed in [16, 17]. Because of the L1 loss, we need to perform speech alignment as discussed above. In our project, we leave the model architecture to be the same, and only modify the input audio to be unmasked speech, and let the model output masked speech. However, we have also tried a few different hyperparameters to improve the output sound quality. Specifically, we lower the spectral convergence factor when calculating the STFT loss.

Result. Generated masked speech sounds rough and has much noise and pop sounds. This is reasonable because the dataset quality is not perfect, especially our own recordings were not recorded in an extremely quiet environment. Speech alignment also introduces some artifact which makes the network harder to learn. Additionally, the generated speech also has a much lower volume. We think this is expected because speech under mask tends to produce sounds with lower volume due the mask covering. There is still lots of room for improvement and we will leave that to future work.

4 Evaluation

Experimental Setup. To evaluate the effectiveness of our data augmentation techniques, we generate the augmentations using either standard augmentation or masked speech augmentation on Libri-Light [9] 10min labeled data. We feed the augmented data plus the original data into Wav2Vec2 pretrained model to do fine-tuning. For testing, we use another 9h labeled data from Libri-Light and use the fine-tuned model plus a 4-gram language model to perform speech recognition.

Data Type	Word Error Rate
Baseline: Original (10min)	8.641
Original + Masked Augmentation (20min)	**8.493**
Original + Standard Augmentation (20min)	**8.277**
Original + Masked + Standard (30min)	8.692

Table 1: Word error rate of baseline and of using different data augmentation techniques

Baseline. The baseline we use is the original Wav2Vec2 model pretrained on Librispeech 960h training data. Due to time constraint and computing resource limitations, we did not go through the pretraining stage, and instead download the model directly from the author's official repository [12]. We take the pretrained model and fine-tune on the 10min labeled data and perform speech recognition on the 9h test data. The word error rate is 8.641 for the baseline.

Standard Augmentation. We generate 10min augmentations of the original data using the standard augmentation, and feed 20min (augmentation + original) into the same pretrained Wav2Vec2 model to do fine-tuning. Then we perform speech recognition on the same 9h test data. With standard augmentation, we achieve 8.277 word error rate which is better than our baseline of 8.641. It is expected as the augmented data have a high quality and sound very close to real data. These augmentations help the model learn from more variations of the data.

Masked Speech Augmentation. We generate 10min augmentations of the original data using the masked speech augmentation, and feed 20min (augmentation + original) into the same pretrained Wav2Vec2 model to do fine-tuning. Then we perform speech recognition on the same 9h test data. The word error rate with masked speech augmentation is 8.493, which is only marginally better than the baseline. Even though the generated speech from the denoiser network is very rough, Wav2Vec2 model still picks up some additional useful information from the masked speech.

Standard & Masked Speech Augmentation. We take both the augmentations plus the original data (30min in total) to fine-tune the Wav2Vec pretrained model. Surprisingly, when we use everything, we have the worst performing model, with a word error rate of 8.692. One possible reason is that if we use too much augmented or fake data, we cause the model to learn more towards the augmentations instead of real data. In our case, we use 20min of augmentations and only 10min of real data. Not only the data ratio is unbalanced, but also we are treating augmentations and real data equally, which can make training worse.

5 Conclusion and Future Work

Motivated by improving Wav2Vec2 further in the low data regime, we propose and investigate two data augmentation techniques. The first we use is standard augmentation, which applies speed adjustment, pitch adjustment and random noise addition to augment the raw audios. The second approach is to generate speech that sounds like as if the speaker is wearing a mask. This is achieved by using a encoder-decoder denoiser network that has very similar settings to ours.

As a result, we find that data augmentation can help in general, but adding too much augmented data leads to an imbalanced data ratio that eventually hurts model performance. Specifically looking at masked speech augmentation, it only provides a marginal improvement because the generated speech is very rough with a relatively bad quality than the standard augmentations. However, it could be a promising direction for data augmentation in speech. For our future work, it is tempting to try out different architectures or frameworks to generate masked speech. Style transfer is one of the options because the input and output of the style transfer networks also have similar properties. Besides a different architecture, a large high-quality dataset is always helpful if we want to generate a high-quality masked speech. With good data augmentation strategies, we can potentially push the limit of our machine learning models especially when the data is scarce.

6 Individual Contribution

Russell:

1. Worked on implementing standard data augmentation and masked speech augmentation.
2. Trained the denoiser network and evaluated wav2vec2 with augmented data.
3. Recorded masked and unmasked data for training denoiser model.

Eason:

1. Set up Google Cloud Platform for computational resources.
2. Worked on setting up baseline of wav2vec2.
3. Recorded masked and unmasked audio pairs for training denoiser model.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[2] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring, 2020.

[3] Violet A. Brown, Kristin J. Van Engen, and Jonathan E. Peelle. Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. 6(1), July 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020.

[5] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019.

[6] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019.

[7] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. ISCA, October 2020.

[8] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

[9] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, and et al. Libri-light: A benchmark for asr with limited or no supervision. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[10] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *INTERSPEECH*, 2015.

[11] Brian McFee, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, , Dana, Kyungyun Lee, Oriol Nieto, Dan Ellis, Jack Mason, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, Viktorandreevichmorozov, Keunwoo Choi, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, Nullmightybofo, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Adam Weiss, Matt Vollrath, Taewoon Kim, and , Thassilo. librosa/librosa: 0.8.1rc2, 2021.

[12] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[13] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, Sep 2019.

[14] Anton Ragni, Kate Knill, Shakti Prasad Rath, and Mark John Francis Gales. Data augmentation for low resource languages. In *INTERSPEECH*, 2014.

[15] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence, 2020.

[16] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Probability density distillation with generative adversarial networks for high-quality parallel waveform generation. ISCA, September 2019.

[17] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram, 2020.

[18] Hoyoung Yi, Ashly Pingsterhaus, and Woonyoung Song. Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. 12, June 2021.

[19] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition, 2020.