This study focuses on audio and speech classification, with the goal of reducing the requirement for huge quantities of labelled data for the Audio Spectrogram Transformers by the use of self-supervised learning using unlabeled data. Using unlabeled recordings from AudioSet and Librispeech, the authors pre-train the AST model utilizing joint discriminative and generative masked spectrogram patch modelling (MSPM). The input audio waveform is first transformed into a series of 128-dimensional log Mel filterbank features calculated every 10ms with a 25ms Hanning window. This outputs a spectrogram that is fed into the AST. This is then split into a series of 16x16 patches. These patches are flattened and embedded as a 1D patch with a linear projection layer. A trainable positional embedding is introduced to enable the model to capture the spatial structure of the 2D audio spectrogram, since the Transformer architecture does not capture the input order information and the patch sequence is not in a temporal order. Next, the Transformer receives the resultant sequence. There are several levels of encoders and decoders in a Transformer. The patch representation is the result of using a Transformer encoder. The audio clip level representation is obtained using a mean pooling across the series of patch representations, and a linear head is used for classification during fine-tuning and inference. The structure remains very consistent with the original AST design. Two modifications were made by the authors to allow for self-supervised learning. First, the authors used a mean-pooling representation of all patches for audio clips. To better summarize the audio clip, the self-supervised pretraining framework applies supervision to each individual patch representation and then takes the mean of those representations. Second, the authors split the patch without overlap during pretraining so that the model would not be able to exploit overlapping edges as a shortcut for the task prediction rather than learning a meaningful representation. Like the initial AST, the patches are divided with an overlap in the fine-tuning and inference phases. Despite being pretrained on fixed-length audio data, AST can handle input of varying lengths by interpolating or truncating the spatial embedding to match the audio requirements of the task. As was previously discussed, AST divides the spectrogram into

patches as part of the pretraining process. The authors furthermore mask spectrogram patches instead of complete time frames during pretraining, enabling the model to pick up on the temporal and frequency structure of the data. This is the Masked Patch Sampling. For this purpose, the authors use cluster factor C to adjust the degree of clustering among masked patches. A p patis picked ch at random, and then we mask the square that contains the patch's coordinates. With a greater C, the model is compelled to learn global spectrogram structure, whereas, with a lower C, it is compelled to learn local structure. During pretraining, we use a random C ~ [3, 5] to ensure that the model picks up on both local and global structures. The authors use a joint discriminative and generative goal for pretraining. We replace the patch embedding of each patch that requires masking with a learnable mask embedding. To the patch embeddings, we add positional embeddings that are then sent into the Transformer encoder. The output from the Transformer encoding device is obtained for each individual masked patch. Then, we feed the output into a classification head and a reconstruction head, and we receive the classification and reconstruction token. It is hypothesized that reconstruction will be quite near to the masked patch, and the model is able to correctly pair the masked patch and the class. The InfoNCE loss is used for the discriminative goal, whereas the mean squared error (MSE) loss is used for the generative objective. Then, we add up these losses using a weight. We update the weights of the AST model to minimize this loss.

This research focuses on audio and speech classification, with the objective of decreasing the need for large amounts of labelled data for the Audio Spectrogram Transformers by the use of self-supervised learning using unlabeled data. The issue is well-motivated. The authors discuss how, although transformers outperform CNN, they

need a large amount of labeled data. Making this procedure self-supervised is therefore a brilliant solution to the issue of a lack of labeled data. This work, which is based on the AST architecture, has two significant contributions by the authors. MSPM is a novel patch-based joint discriminative and generative self-supervised learning framework proposed by the authors. When compared to prior supervised pretrained AST, the SSAST model with MSPM pretraining achieves the same or better results. Both MSPM and SSAST establish new standards in the audio and speech domains, with MSPM being the first self-supervised learning framework based on patches and SSAST being the first self-supervised pure self-attention based audio categorization model. Second, the authors demonstrate that compared to pretraining with datasets from a single domain, pretraining with both speech and audio datasets significantly increases the models' generalization capacity and leads to improved performance. The SSAST model therefore achieves high quality results in downstream audio and speech tasks. In most cases, prior work simply pretrains on datasets from a single domain. Extensive experimentations demonstrate that the proposed MSPM self-supervised pre-training architecture considerably enhances AST performance across the board for downstream tasks, by an average of 60.9%. It is clear that the proposed MSPM can replace supervised pretraining, which requires a substantial quantity of labeled data, since the SSAST model achieves similar or better results than earlier supervised pretrained models and has higher generalization capabilities. The research also demonstrates that pretraining the model using data from both the speech and audio domains yields better results than using data from a single domain, and that the model with both generative and discriminative goals leads to a higher performance than using a single objective.

Thus , we discover that MSPM facilitates AST model scaling, such that bigger ASTs consistently outperform smaller ones when subjected to MSPM pretraining. Considering MSPM's adaptability to patch shape, more exploration into frame-based AST would be a fascinating topic for researchers. Replacement of static patch masking with dynamic masking, as advocated by Fan et al. in "Mask Attention Networks: Rethinking and Strengthen Transformer," could be one approach to enhancing masking performance.

This study paves the way for more research on unsupervised learning and self-supervised learning in the audio and speech domains. Because there is no need for labeling, it also opens up the possibility of using data from the wild.