A Spectrogram is Worth 16 x 16 Words: ViTs for Spoken Language

Pooja Sethi [1] Jared Weissberg [1] Kaushal Alate [1]

[1]Department of Computer Science, Stanford University

Motivation & Background

- The Audio Spectrogram Transformer (AST) [2] showed that **CNNs are not necessary** for end-to-end audio classification. A convolution-free, purely attention-based architecture has benefits.
- AST naturally supports variable length inputs without any architecture changes. It has fewer parameters and converges faster during training.
- AST achieves SOTA results on three audio classification benchmarks: AudioSet, ESC-50, and Speech Commands (recognizing single words). But, AST's **performance on spoken language tasks is under-explored.**

Problem Statement

- Can a purely attention-based architecture achieve competitive performance on spoken language tasks?
- We evaluate AST for **spoken language classification** as well as introduce a novel convolution-free automatic speech recognition (**ASR**) architecture.

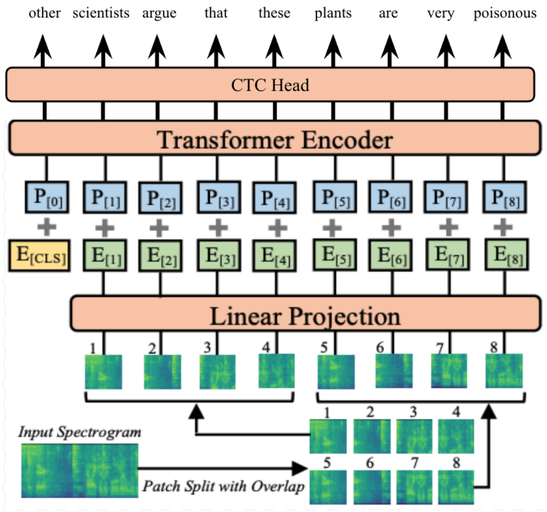


Figure 1. Our novel architecture for ASR, `ASTForCTC`.

Tasks & Datasets

Table 1. We tested two types of classification tasks (SER and music genre) and ASR.

Task	Description
Speech Emotion Recognition (SER)	
RAVDESS	1440 files (24 actors) of 7 emotions
IEMOCAP	12 hours (5 actors) of 5 emotions
Music Genre Classification	
GTZAN	100 30-second tracks per genre x 10 genres
Automatic Speech Recognition	
Google FLEURS (en_us)	10 hours train; 2 hours for validation/test

Experiments & Methods

- **Speech Classification:** We added a linear layer head (`ASTForClassification`) to the pre-trained AST model and fine-tuned with Adam, weight decay, and a linear learning rate scheduler for either 5 or 20 epochs. The metric we measured was accuracy (total correct classifications). We benchmark emotion recognition with the SOTA model Att-Net [4] for RAVDESS and DS-CNN for IEMMOCAP. We benchmark music genre classification with the SOTA model MUSER (MUsic SEquence Representation) [1].

- **ASR:** We added a CTC head (`ASTForCTC`) to the pre-trained AST model and built a custom ASTProcessor that combines the ASTFeatureExtractor for audio processing with a Wav2Vec2CTCTokenizer for label processing. We used the same learning rate and scheduler as for classification. The metric we measured was the Word Error Rate (WER).

Results: Emotion Recognition, Music Genre, and ASR

Task	Baseline	AST (5 Epochs)	AST (20 Epochs)
Music Genre Classification (GTZAN)	0.825	0.865	**0.900**
SER (IEMOCAP)	**0.788**	0.640	0.7
SER (RAVDESS)	0.800	0.760	**0.816**

Table 2. Accuracy results for different classification tasks. Best results are in bold.

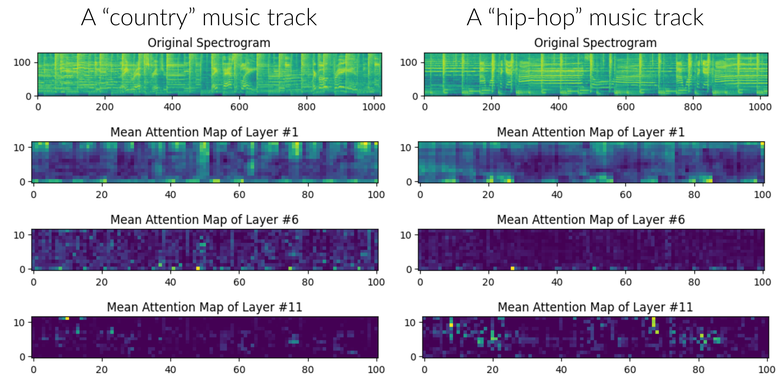
Prediction	Reference	pwc	rwc	ratio
some people believe that experiencing many artificially induced lucid dreams often enough can be very exhausting	some people believe that experiencing many artificially induced lucid dreams often enough can be very exhausting	16	16	1.0
permits must be reserved in advance you must have a permit to stay overnight at sirenasss	permits must be reserved in advance you must have a permit to stay overnight at sirena	16	16	1.0

Table 3. ASR training predictions. pwc is the predicted word count, and rwc is the reference word count.

Prediction	Reference	pwc	rwc	ratio
a oy al al un ald wo hernm os an sth ldg hrndlsoni wn weshey iaodnighisssssssss	in some areas boiling water for a minute is enough in others several minutes are needed	16	16	1.0
aoas alsi toilnsd tans pan cailehtan s.iic tafnendan inirds...	oliver sacks in his paper the president's speech indicated how people...	28	28	1.0

Table 4. ASR test predictions.

Analysis: Spectrograms & Attention Maps



The figures above show spectrograms and attention maps for music tracks of two distinct genres. The differences illustrate how the AST model learns to pay attention to different regions for classifying different kinds of music.

Conclusions

Speech Classification

- The AST outperformed the SOTA baseline for emotional recognition on the RAVDESS dataset, but performed worse than the baseline on the IEMOCAP dataset. The AST model achieved SOTA performance on music genre classification.

ASR

- AST performed surprisingly well despite being trained in an extremely low-resource setting (10 hours of data).
- We obtained a WER of 0.093 on the FLEURS train set and 1.085 on the test set, which is suggestive of overfitting.
- The AST model seems good at detecting the number of words in an utterance – the ratio of the number of words predicted to the number of actual words was 1.00 in the train set and 1.05 in the test set.
- We observed a few hallucinations toward the end of the sentence; e.g., adding extra "s"'s to the ends of predicted words (see Table 3).

Next Steps

- Compute and time allowing, we are trying to train `ASTForCTC` from scratch on Librispeech (10,000 hours of data).

Seleted References

[1] Tianyu Chen, Yuan Xie, Shuai Zhang, Shaohan Huang, Haoyi Zhou, and Jianxin Li. Learning music sequence representation from text supervision. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4583–4587, 2022.

[2] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021.

[3] Zhengwei Huang, Ming Dong, Qi rong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn.