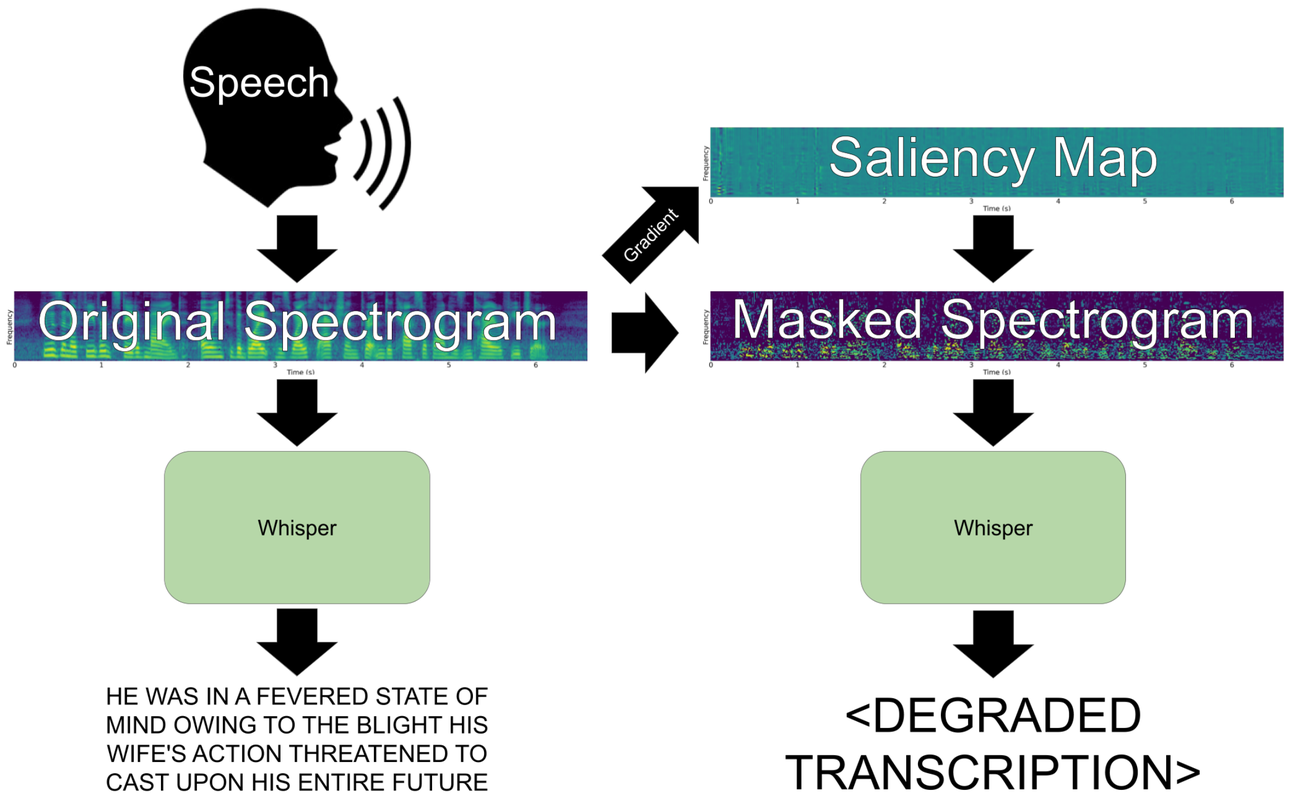
SalASR: Legilimency using SALiency highlighting for Automatic Speech Recognition

Jacob Johnson, Gurunath Parasaram, Rishanth Rajendhran

Research Question: Are Saliency Highlights an unfaithful explainability method for ASR?

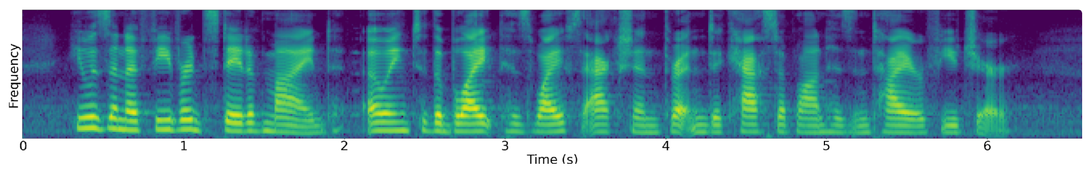
Hypothesis: Masking out most-salient features will lead to a steeper decrease in token-success rate than masking out least-salient or random features



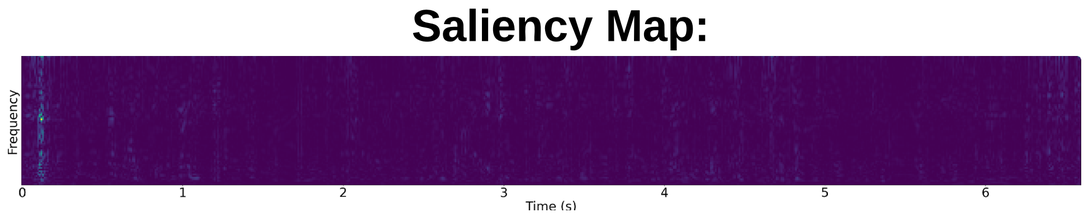
Model: Whisper-large v2
(finetuned for masked spectrograms)
Dataset: Librispeech-ASR

Saliency is the sum of gradients from each token in the transcription over the input spectrogram

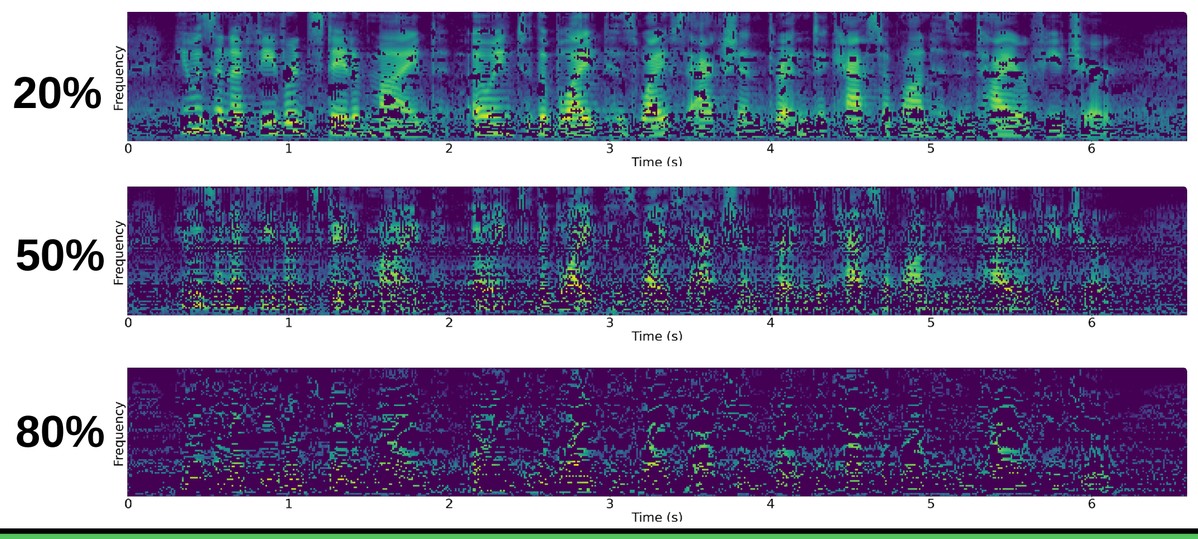
Original Spectrogram:



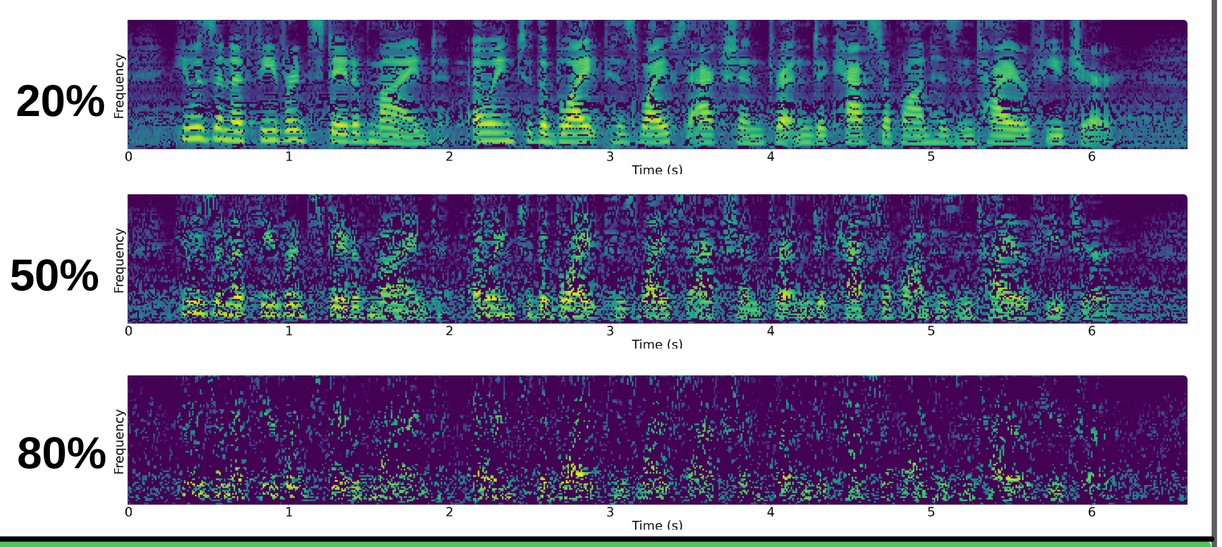
Saliency Map:



Most salient features masked:



Least salient features masked:



Model Results

- We finetune Whisper-v2-large on 3 epochs of 480 masked spectrograms to lessen out-of-domain effects (ROAR)
- We evaluate token success rate with teacher-forcing (i.e. floor is next-token-prediction performance) with respect to model output on the unmasked spectrogram

User study to verify model results

- We seek to ground our model-based evaluation in a human replication of the transcription task
- We elicit transcriptions for ~38 samples per masking condition (r={.2, .5, .8}, {top, random, bottom}), split over 19 participants

Conclusions

- Model scores support our hypothesis
- Suggests that the most salient features contain more (useful) information than randomly selected features
- **We do not conclude that saliency is unfaithful for ASR**
- Gradient-based saliency on spectrograms may be faithful for other audio tasks (translation, etc)
- Highlights align with linguistically-informed intuitions

- Masking most-salient features → lower performance than random or least-salient features (7 out of 9 cases)
- Masking more of the same features → lower performance (6 out of 9 cases)
- User study results are not conclusive (small sample size)
- Often tend in the same direction as the model results
- Tentatively supports our hypothesis