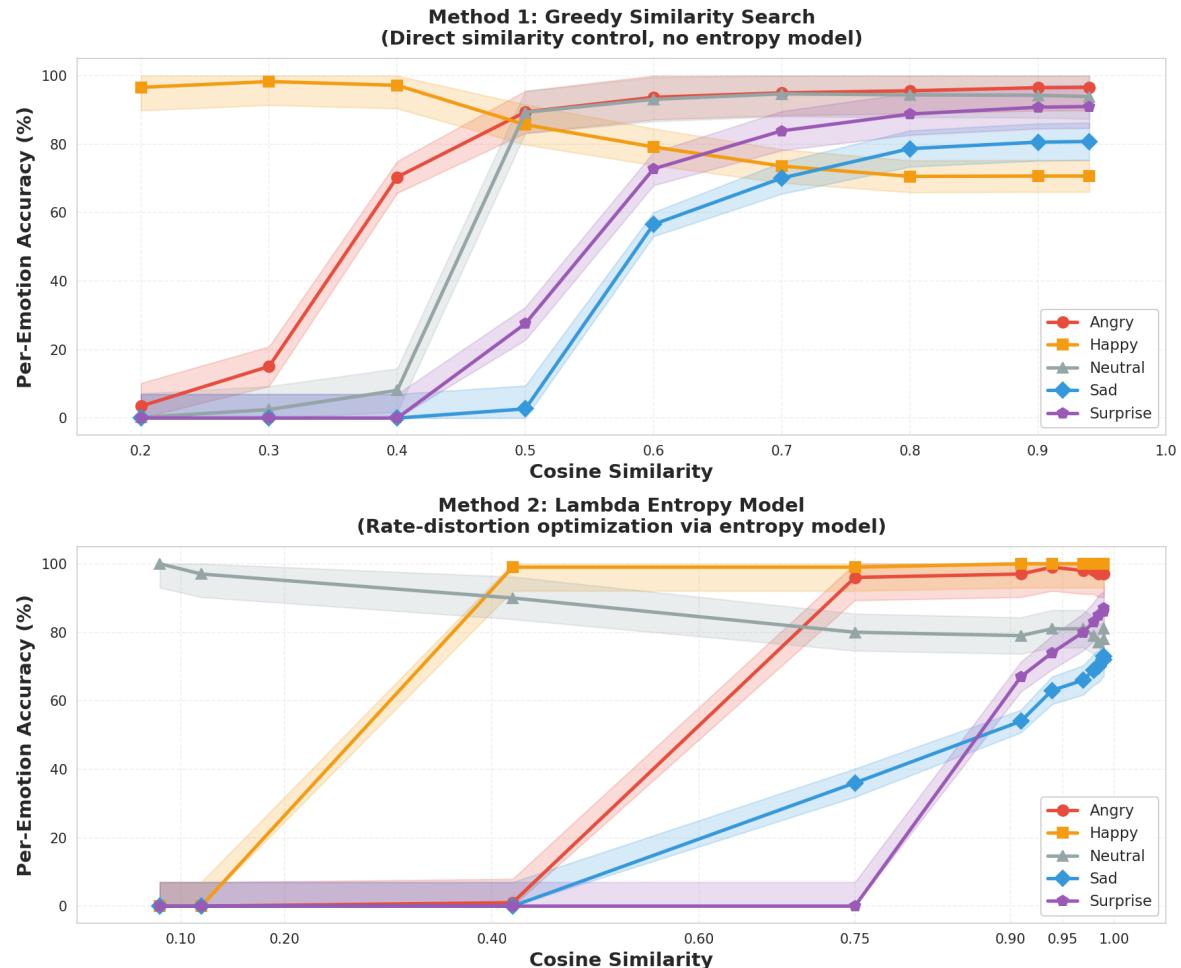


Summary of Work Completed Since Last Meeting

I feel it might be easier to explain directly during the meeting. Given time constraints and my desire to proceed with subsequent experiments quickly, my content may be relatively brief.

1. Systematic Review of Previous Quantization Methods

I systematically reviewed the previous quantization methods and discovered that directly reducing similarity cannot be used for quantization. Direct similarity reduction behaves inconsistently with quantization using GRVQ codebook + entropy model (or RVQ hierarchical) approaches.



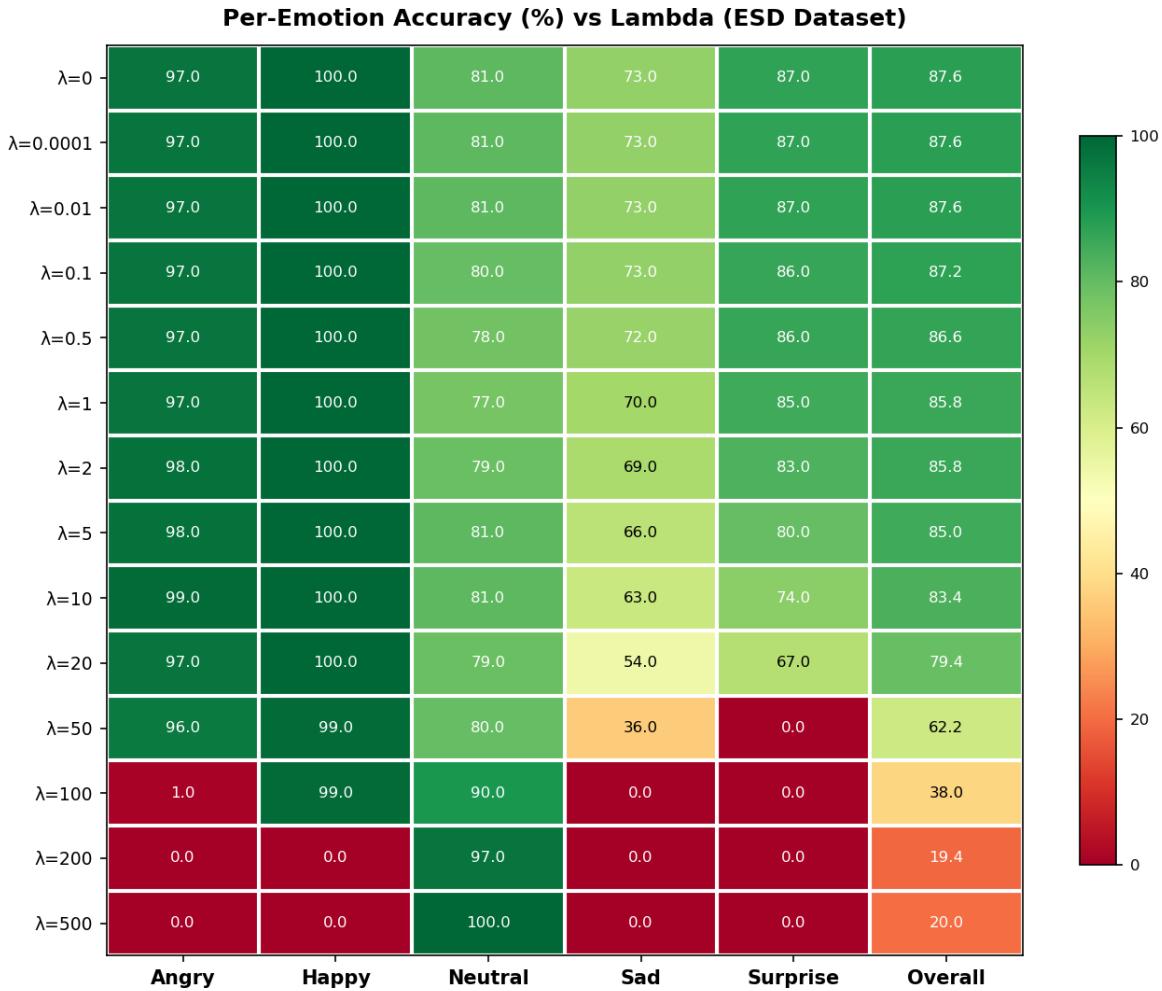
From the images, we can observe that when directly reducing similarity, neutral emotion actually decreases first, followed by happy emotion. This seems to suggest that the quantization direction of directly reducing similarity is inconsistent with our quantization direction. More experiments and insights are currently in progress.

1.2 Entropy Model Bug Fix and Efficiency Discovery

I fixed the similarity bug in the entropy model. I also discovered that after removing specific bitrate control, the entropy model + GRVQ scheme appears to be more efficient than direct binary search greedy search.

1.3 Insight into Neutral Emotion Robustness

I gained insight into why neutral emotion was most robust in previous experiments: it's because neutral emotion's similarity decreases most slowly under our quantization method, while other emotions' similarities drop dramatically. This direction of similarity change is inconsistent with the direction of directly reducing similarity.



This shows the downstream performance loss as quantization intensity increases (λ indicates quantization intensity, non-linear). The subsequent quantization intensity vs. similarity graphs tell us that this result occurs because our quantization method causes representations of non-neutral emotions to rapidly lose similarity under high quantization intensity.

Per-Emotion Cosine Similarity vs Lambda (ESD Dataset)



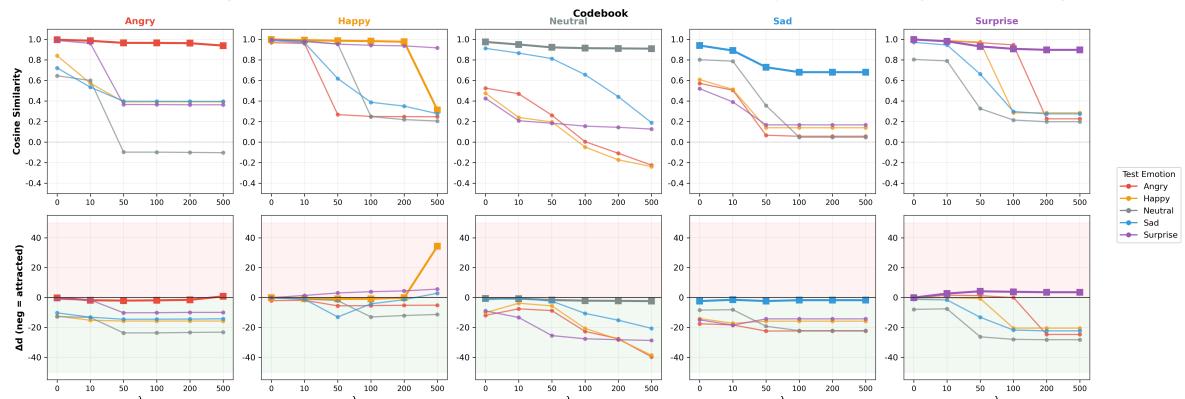
From the images, we can observe that other emotions' similarities rapidly decrease through our quantization method, while neutral emotion consistently maintains high similarity.

1.4 Emotion Similarity Degradation Direction

I discovered that emotion similarity doesn't simply decrease, but rather degrades toward the direction of neutral emotion. In other words, when other emotions' representations move away from themselves, they are simultaneously approaching neutral emotion representations.

1.5 Emotion-Biased Codebook Training Experiments

I attempted to train different emotion-biased codebooks using different emotions. I found that codebooks develop biases toward different emotions based on your training data changes.



I utilized the ESD dataset and separately trained 5 independent codebooks and corresponding entropy models based on different emotion categories. In other words, each codebook was trained using only a single emotion. The first row of graphs shows how similarity changes as quantization intensity increases. We can observe that codebooks trained with corresponding emotion data maintain high similarity for that emotion even under high quantization intensity.

I also conducted similarity approach analysis and found that using the corresponding emotion's codebook for quantization causes emotions to approach that direction (the approach trend for the same emotion is not obvious). Notably, this trend exists from the moment you start using this codebook, regardless of whether you increase quantization intensity.

2. Retrained Genner Speech Emotion Embedding Model

I retrained the emotion embedding model for Genner Speech, changing the emotion embedding from LSTM (built into the model) to emotion2vec. Considering indexing limitations and training speed, I only used 20 hours of ESD data for training, while Genner Speech uses 400-500 hours of data. Audio quality is not as good as Genner Speech, but acceptable given the data volume.

2.1 Quantized Reference Speech Emotion Embedding Experiment

I tried quantizing the reference speech's emotion embedding and applied different levels of quantization intensity. I found that at low quantization intensity, emotions didn't change much. However, as quantization intensity gradually increased, emotions seemed to gradually approach neutral.

Please visit the website directly to listen: <https://1355-xcz.github.io>ShowWork1/index.html>

2.2 Extended Experiments with Different Biased Codebooks

I extended the experiments to different biased codebooks. I found that emotions indeed approach the corresponding emotion based on the codebook's bias, although audio quality is poor.

Please visit the website directly to listen: https://1355-xcz.github.io>ShowWork1/emotion_cross.html

3. Ambiguous Experiment Attempt

I attempted the "ambiguous" experiment proposed by the senior student, but I found no significant results and stopped the subsequent experiments. The specific images are in the folder. Since there are many images, I plan to present them during the meeting.

Specific Application and Improvement Directions

1. **Neutral-First Transformation:** Is it a more effective approach to first restore to neutral and then convert to other emotions?
2. **Multi-Reference Emotion Mixing:** Can we utilize a similar codebook quantization mechanism to break the original limitation of using only a single reference audio or emotion

reference, and introduce multiple emotion references through a codebook quantization-like approach?

3. **Mixed Emotion Audio Creation:** For creating bittersweet (sad-happy mixed) audio, can we train a codebook with both sad and happy audio for representation quantization, thereby achieving emotion mixing? It doesn't have to be a codebook; similar concepts could work as well.