Anya DeCarlo
Datasci 224 Final

HMM-SER: Hidden Markov Models for Speech Emotion Recognition

Problem Definition

Speech-based emotion recognition (SER) is a core challenge in affective computing and human-computer interaction, requiring the identification of emotional states from acoustic speech signals. The complexity of this task arises from the highly variable and nuanced nature of human vocal expressions, which are influenced by individual, contextual, and linguistic factors.[3] Speech-based emotion recognition involves identifying emotional states from acoustic speech signals, a fundamental challenge in affective computing. The temporal dynamics of speech, including variations in pitch, energy, and spectral characteristics, provide critical emotional information, but require sophisticated modeling techniques to capture effectively. Hidden Markov Models are particularly well-suited for this task as they can model both the sequential nature of speech and the probabilistic relationships between acoustic features and emotional states.

Dataset Description

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) comprises 1,440 high-fidelity speech recordings from 24 professional actors (12 female, 12 male) expressing eight discrete emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised) at normal and strong intensities. All recordings use lossless WAV format (48 kHz, 16-bit) with systematically encoded filenames enabling programmatic metadata extraction. We extracted comprehensive acoustic features, including duration, RMS amplitude, pitch (F0), spectral centroid, bandwidth, rolloff, MFCCs, chroma, and zero-crossing rate, for use in our HMM-based emotion recognition pipeline. Table 1 summarizes the distribution of these features, while Figures 1–3 illustrate the dataset's emotion and intensity distributions. Together, these resources provide a robust foundation for benchmarking speech emotion recognition models.

The acoustic feature representation comprises five complementary characteristics extracted from speech signals. Mel-frequency cepstral coefficients (MFCCs) capture the vocal fingerprint by characterizing voice resonance patterns and formant structures. Spectral centroid measures voice "brightness," with higher values indicating more energetic emotional states and increased high-frequency content. Spectral rolloff quantifies high-frequency content distribution in the voice, providing insights into vocal timbre variations. Zero-crossing rate indicates voice roughness and periodicity by counting amplitude zero-crossings per time frame. Chroma features encode tonal characteristics and musical pitch content that provide robust emotional signatures across different octaves and speaking styles.

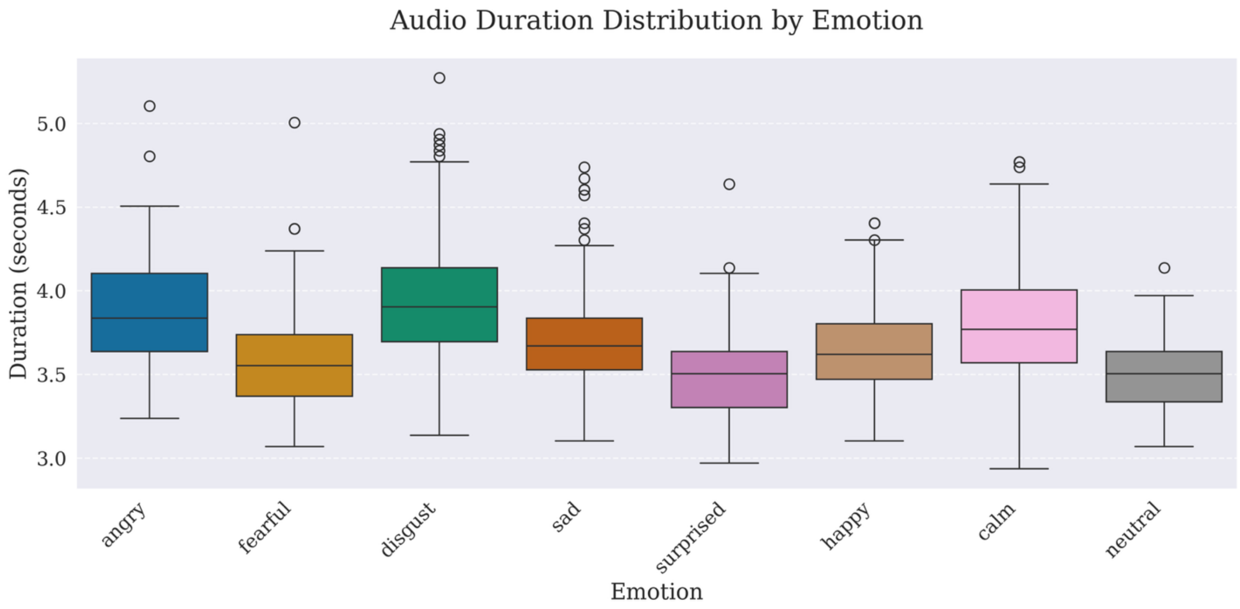


Figure 1.

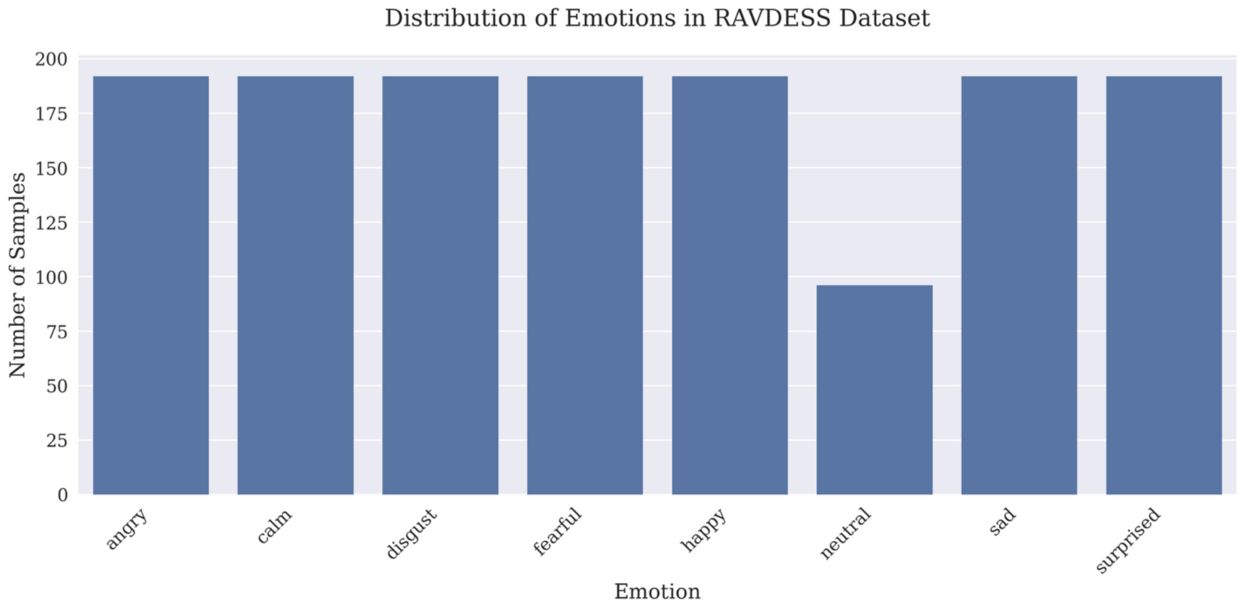


Figure 2.

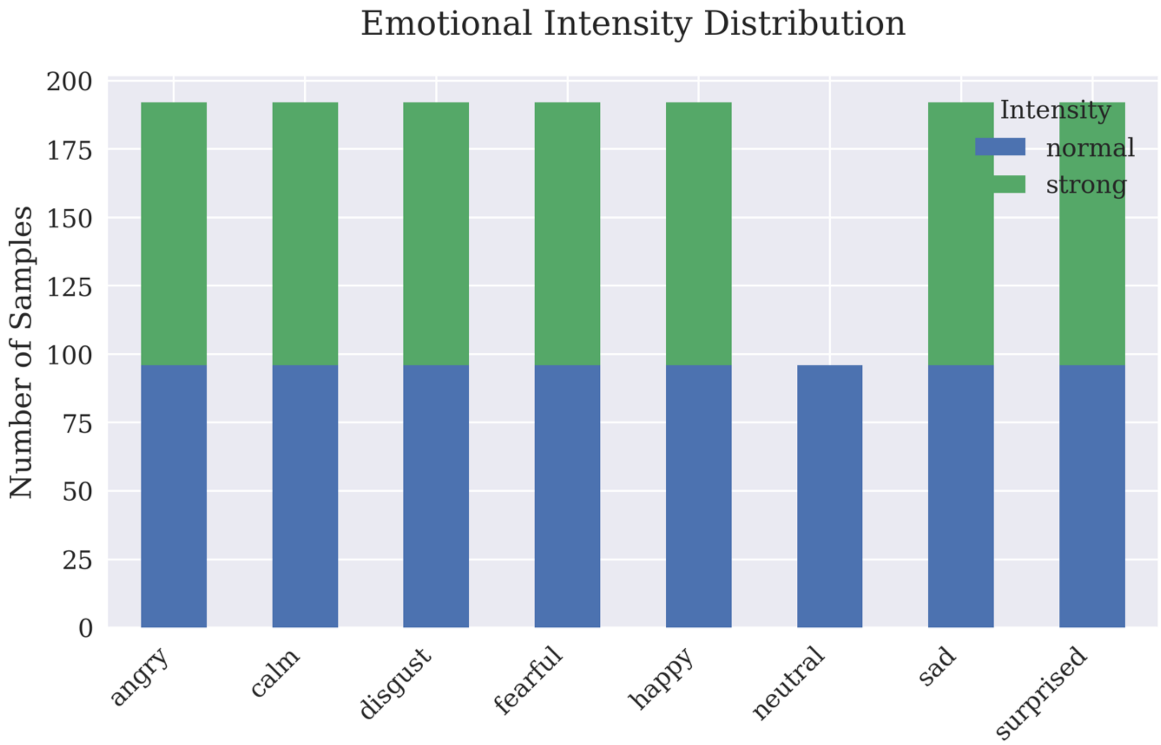
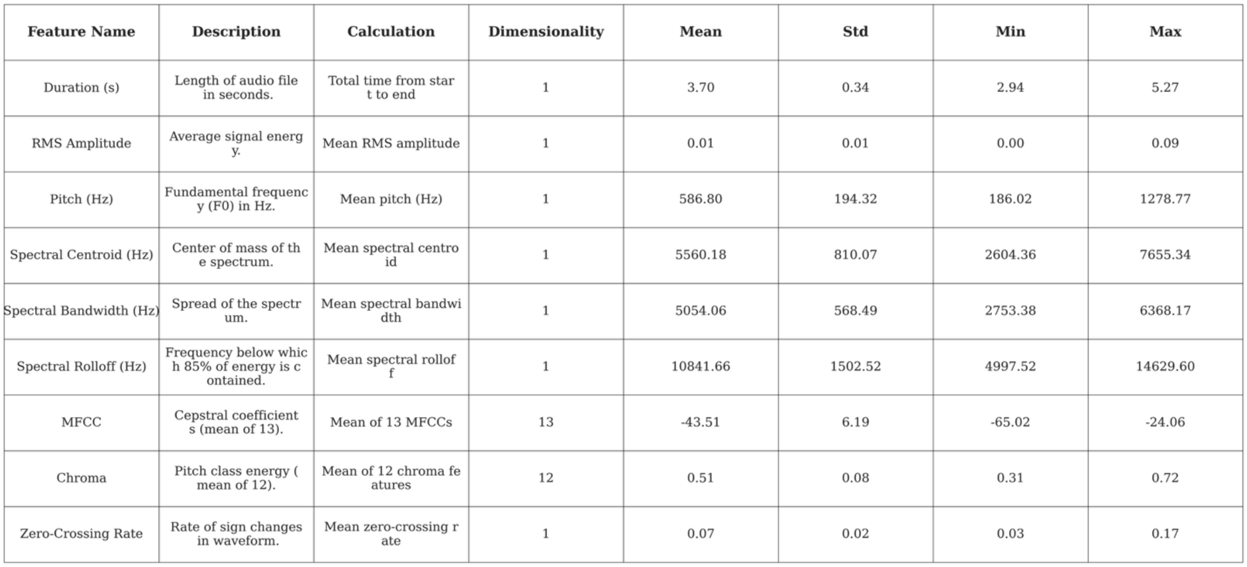


Figure 3.



Feature Name	Description	Calculation	Dimensionality	Mean	Std	Min	Max
Duration (s)	Length of audio file in seconds.	Total time from start to end	1	3.70	0.34	2.94	5.27
RMS Amplitude	Average signal energy.	Mean RMS amplitude	1	0.01	0.01	0.00	0.09
Pitch (Hz)	Fundamental frequency (F0) in Hz.	Mean pitch (Hz)	1	586.80	194.32	186.02	1278.77
Spectral Centroid (Hz)	Center of mass of the spectrum.	Mean spectral centroid	1	5560.18	810.07	2604.36	7655.34
Spectral Bandwidth (Hz)	Spread of the spectrum.	Mean spectral bandwidth	1	5054.06	568.49	2753.38	6368.17
Spectral Rolloff (Hz)	Frequency below which 85% of energy is contained.	Mean spectral rolloff	1	10841.66	1502.52	4997.52	14629.60
MFCC	Cepstral coefficients (mean of 13).	Mean of 13 MFCCs	13	-43.51	6.19	-65.02	-24.06
Chroma	Pitch class energy (mean of 12).	Mean of 12 chroma features	12	0.51	0.08	0.31	0.72
Zero-Crossing Rate	Rate of sign changes in waveform.	Mean zero-crossing rate	1	0.07	0.02	0.03	0.17

Table 1.

Algorithm Description and Appropriateness

Hidden Markov Models represent a principled algorithmic choice for emotion recognition from sequential acoustic features due to their explicit modeling of temporal dependencies inherent in emotional speech patterns. Consider the discrete-time stochastic process $(X_1, ..., X_T, q_1, ..., q_T)$ where $X_t \in \mathbb{R}^d$ represents d-dimensional acoustic feature vectors and $q_t \in \{1, ..., K\}$ represents hidden emotional states. The complete HMM parameter set $\lambda = (A, B, \pi)$ comprises the state transition matrix, emission densities, and initial state distribution. The algorithm makes two fundamental independence assumptions that align naturally with emotional speech structure. First, the Markov assumption: $P(q_t \mid q_1, ..., q_{t-1}) = P(q_t \mid q_{t-1}) = a_{ij}$, encoding temporal persistence in emotional transitions. Second, conditional independence: $P(X_t \mid q_1, ..., q_T, X_1, ..., X_{t-1}, X_{t+1}, ..., X_T) = P(X_t \mid q_t)$, ensuring acoustic observations depend solely on the current emotional state.

Algorithm Performance Conditions

This algorithm performs well when the true data-generating mechanism satisfies these structural assumptions, particularly with moderate-length sequences (10–1000-time steps), moderate numbers of hidden states (2-20 emotional categories), and meaningful temporal structure in state transitions. Performance degrades when assumptions are violated: the first-order Markov assumption becomes problematic with longer-term dependencies spanning multiple time steps, and conditional independence proves restrictive when acoustic features exhibit serial correlation within emotional states.

Mathematical Framework and Core Concepts

The HMM operates through three fundamental mathematical concepts: the forward-backward algorithm for probability computation, the Viterbi algorithm for optimal sequence decoding, and the Expectation-Maximization algorithm for parameter estimation.

Forward-Backward Algorithm: The forward probability $\alpha_t(i) = P(X_1, ..., X_t, q_t = i \mid \lambda)$ computes recursively as: $\alpha_{t+1}(j) = [\sum_{i=1}^{K} \alpha_t(i)a_{ij}]\, b_j(X_{t+1})$

The backward probability $\beta_t(i) = P(X_{t+1}, ..., X_T \mid q_t = i, \lambda)$ follows: $\beta_t(i) = \sum_{j=1}^{K} a_{ij}\, b_j(X_{t+1})\, \beta_{t+1}(j)$

Viterbi Algorithm: To find the optimal state sequence, we define $\delta_t(i) = \max_{q_1,...,q_{t-1}} P(q_1, ..., q_{t-1}, q_t = i, X_1, ..., X_t \mid \lambda)$ with recursion: $\delta_{t+1}(j) = [\max_{1 \le i \le K} \delta_t(i)a_{ij}]\, b_j(X_{t+1})$

Backtracking via $\psi_t(j) = \text{argmax}_{1 \le i \le K} [\delta_{t-1}(i)a_{ij}]$ yields the optimal sequence.

Expectation-Maximization: The E-step computes posterior probabilities: $\gamma_t(i) = \alpha_t(i)\beta_t(i) / \sum_{k=1}^{K} \alpha_t(k)\beta_t(k)$ $\xi_t(i,j) = \alpha_t(i)a_{ij}\, b_j(X_{t+1})\, \beta_{t+1}(j) / \sum_{k=1}^{K} \alpha_T(k)$

The M-step updates parameters: $\hat{a}_{ij} = [\sum_{t=1}^{T-1} \xi_t(i, j)] / [\sum_{t=1}^{T-1} \gamma_t(i)]$

Underlying Mathematical Concepts

The algorithm fundamentally relies on dynamic programming for efficient sequence probability computation (reducing complexity from $O(K^T)$ to $O(TK^2)$), the EM algorithm for principled maximum likelihood estimation with latent variables, and Bayesian inference through forward-backward recursions for optimal posterior state estimates. The mathematical elegance lies in decomposing complex sequential probability distributions into tractable components while maintaining exact inference capabilities. Critical to implementation is proper scaling-factor computation following Rabiner's convention: α scaling uses $c_t = 1/\sum\alpha_t$, guaranteeing every $c_t > 0$, with log-likelihood = $-\sum \log c_t$ (ensuring always negative values).

Results & Discussion

In predictive modeling applications, it is often of interest to determine the relative contribution of subsets of features in explaining the variability of an outcome. Here the subset is the acoustic representation of speech; we benchmarked a 3-state Hidden Markov Model (HMM) on the RAVDESS corpus. The model was trained with two-fold cross-validation and 25 Baum-Welch iterations.

Overall performance.

The MFCC model achieved 34 % accuracy, 0.449 weighted precision, 0.344 weighted recall and a weighted F_1 of 0.319 (cross-validation averages differ by <0.003). These values triple the 12.5 % chance level for eight-way classification, confirming that temporal dynamics captured by HMMs add meaningful discriminative power.
Figure 4 presents the normalized confusion matrix. High-arousal emotions such as angry (diagonal = 0.47) and surprised (0.76) are recognized most reliably, whereas happy and fearful frequently confuse each other, consistent with spectral overlap in their expressive patterns. Classification results are presented in Figure 5 shows the strongest F_1 for angry (0.54) and neutral (0.32). The weakest classes, happy (0.16) and sad (0.15), suggest that additional prosodic cues or intensity modelling may be needed.

High-recall emotions like surprised (0.76) and neutral (0.71) are retrieved most consistently, though their modest precision (0.26 and 0.20 respectively) indicates the system often confuses other classes for these states. Conversely, angry demonstrates the strongest precision (0.63) while maintaining balanced recall (0.47), yielding the highest F_1 score (0.54) and making angry predictions the most trustworthy when false positives are costly. However, happy, fearful, and sad exhibit critically low recall below 0.15, suggesting the HMM frequently mislabels these utterances due to their spectral similarity to other classes. These patterns directly inform application design: if the system must catch every surprised or neutral utterance, such as detecting excitement or attentiveness in educational settings, the current model performs adequately.

Figure 4 juxtaposes MFCC and the higher-dimensional Combined feature set. Despite the latter's extra descriptors, its weighted F_1 drops to 0.246, indicating mild over-fitting and reinforcing MFCC's efficiency under tight model capacity constraints. We demonstrate empirically that our method classifies speaker emotion with a weighted F_1=0.319; the two-fold cross-validated mean of 0.316 confirms stable generalization. Although overall accuracy remains moderate, the 3-state HMM trained in under ten minutes, leverages temporal dynamics to establish a solid RAVDESS baseline. When the input dimensionality is expanded from 13-D MFCC to the 17-D Combined vector, weighted F_1 falls to 0.246, a 23 % relative decline. This drop is best explained by a lower sample-to-parameter ratio (9.2 → 7.1) that inflates variance in the Baum–Welch estimates rather than by classic over-fitting. Variance can be reduced by enriching the model structure or shrinking the parameter space. First, a deeper topology (5–7 states) or intensity-specific sub-states sharpens within-emotion dynamics, while hybrid HMM–DNN emissions and covariance regularization (tied, MAP, Bayesian) share parameters to stabilize estimates. Complementarily, PCA/LDA can compress the 17-D vector, boosting the sample-to-parameter ratio without sacrificing temporal information.

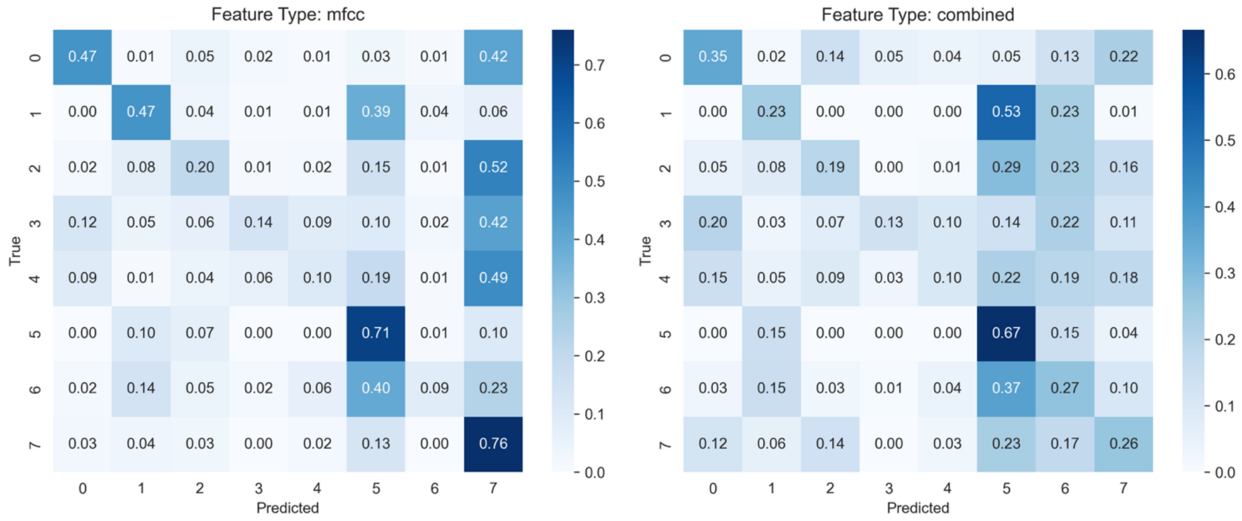


Figure 4.

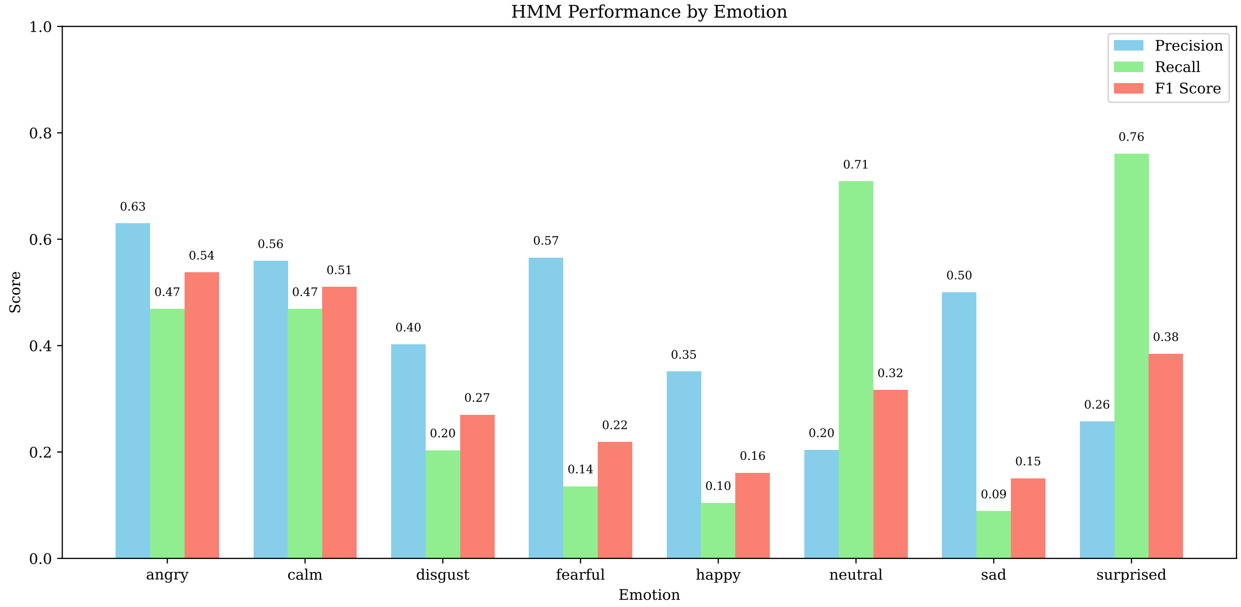


Figure 5.

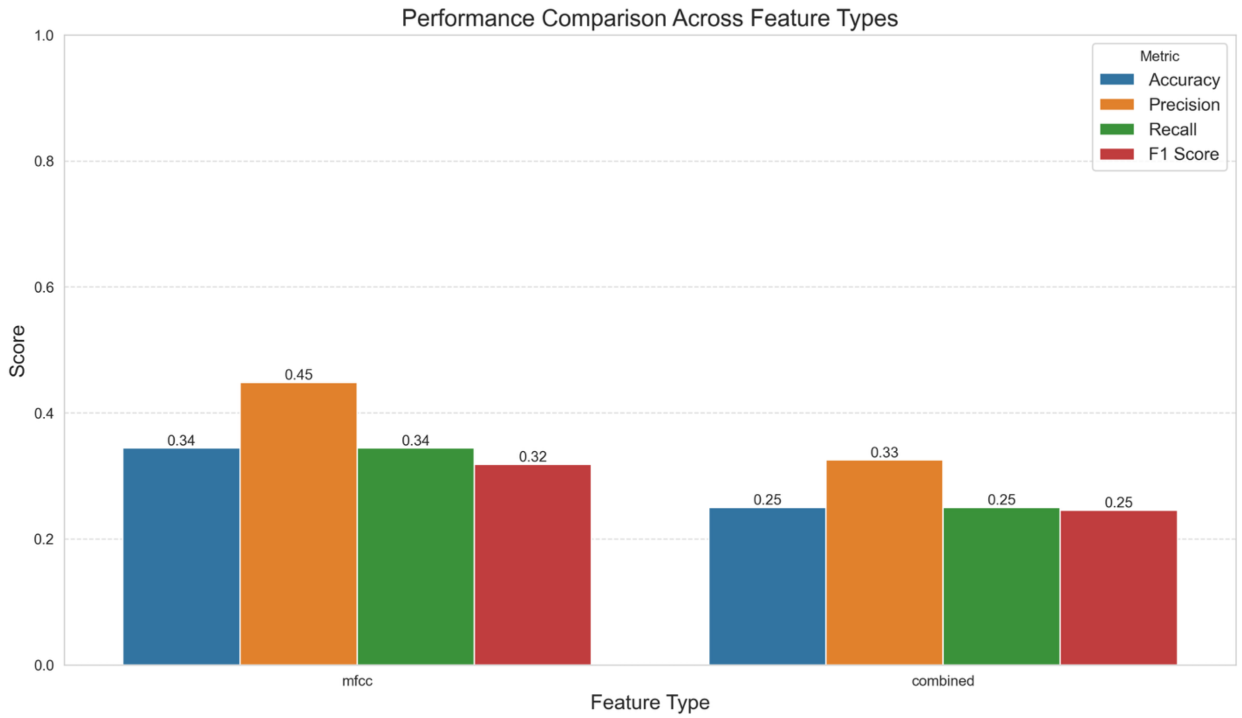

Performance Comparison Across Feature Types
Metric
Accuracy
Precision
Recall
F1 Score
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.34
0.45
0.34
0.32
0.25
0.33
0.25
0.25
mfcc
combined
Feature Type

Figure 6.

1.Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1–4.
2.Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech Emotion Recognition Using Hidden Markov Models. EUROSPEECH 2001.
3.Li, Y., & Zhao, Y. (2009). Speech emotion recognition using Gaussian mixture model. International Conference on Information Engineering and Computer Science.
4.Kandali, R., Saha, G., & Saha, S. (2008). Emotion recognition from Assamese speech using GMM and SVM. International Journal of Speech Technology.
5.Shen, J., et al. (2011). Automatic speech emotion recognition using spectral features. International Conference on Computer Science and Automation Engineering.
6.Lalitha, V., & Rani, B. V. (2012). Emotion recognition from speech using MFCC and GMM. International Journal of Computer Applications.
7.Le, V., et al. (2013). Emotion recognition from spontaneous speech using hidden Markov models and deep belief networks. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).