Reinforcement Learning for Feature Selection in Affective Speech Classification

Eric Lau, Suraj Heereguppe, Chiraag Sumanth

{eclau, hrsuraj, csumanth}@stanford.edu

Stanford | ENGINEERING

I. Summary

- Even with complex state-of-the-art features, affective speech classification accuracies of only 60-70% are reported in the literature
- Previous work has involved hand-engineering features and are impractical and unscalable
- Goal: Apply reinforcement learning (RL) to automatically learn approximately optimal features for affective speech classification
- Built RL procedure with GMM-HMM classifier
- Ensemble classifier trained on computed feature subsets shows 9.3% gain in test accuracy over baseline

II. Data

- RML Emotion Database: 720 audiovisual emotional expression samples with ground-truth emotion class labels
- Samples uniformly distributed over six emotion categories: *Anger, Disgust, Fear, Happiness, Sadness,* and *Surprise*
- Samples span six different languages

III. Baseline Features

- Stripped audio and extracted Mel-Frequency Cepstral Coefficient (MFCC) feature vectors, commonly used in speech recognition
- 13 features in MFCC vector, indexed *0...12*
- MFCC vectors obtained by dividing each audio file with a 25ms sliding frame, with a 10ms frame step size.

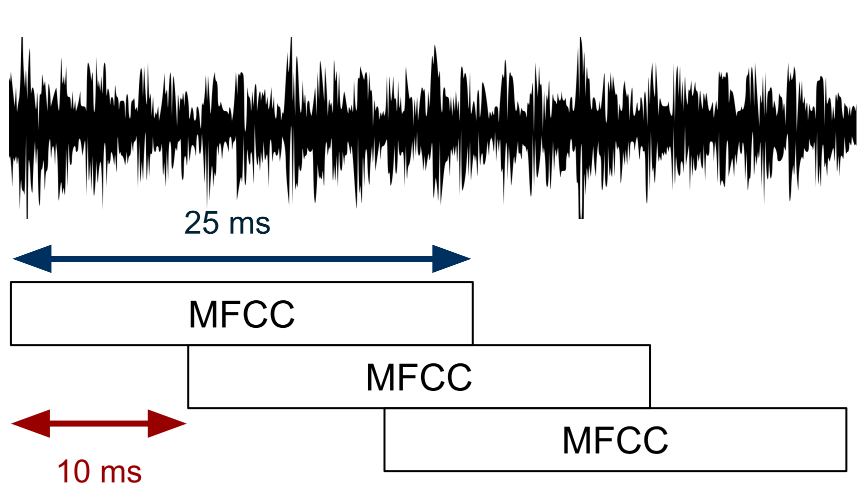


Figure 1: MFCC vector extraction from raw audio.

IV. Models

Choose Approximately Optimal Features: **Reinforcement Learning Procedure**

- Uses modified Q-learning type procedure, parametrized by (S, A, γ, R) and described in Figure 3
- States s in S are k centroids found by k-means clustering of the training set; chose k = 4 by Elbow method (Figure 2)
- Learning rate α_t is inversely proportional to number of visits to state s_t; γ = 0.99

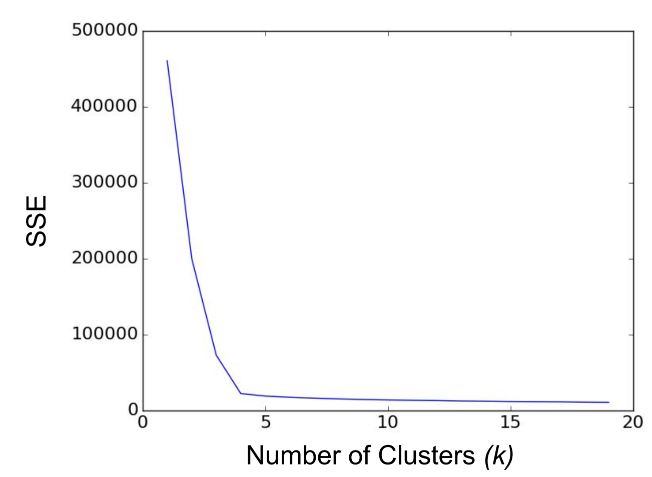


Figure 2. Plot of sum of squared error (SSE) versus number of clusters; chose k = 4.

Input: Training samples, feed in randomly; initialize Q-table

↓

For each training sample:

1. Get state s in S: nearest centroid to sample
2. Choose action a in A: randomly pick single feature and feed to classifier trained on single feature
3. Reward R: r = 1000 if correctly classified compared to ground truth, -1000 otherwise
4. Update Q-Table entry of (s_t,a_t) at this step t as follows:

$$Q(s_t, a_t) \leftarrow (1-\alpha_t)Q(s_t, a_t) + \alpha_t(r + \gamma \max_a Q(s_t, a))$$

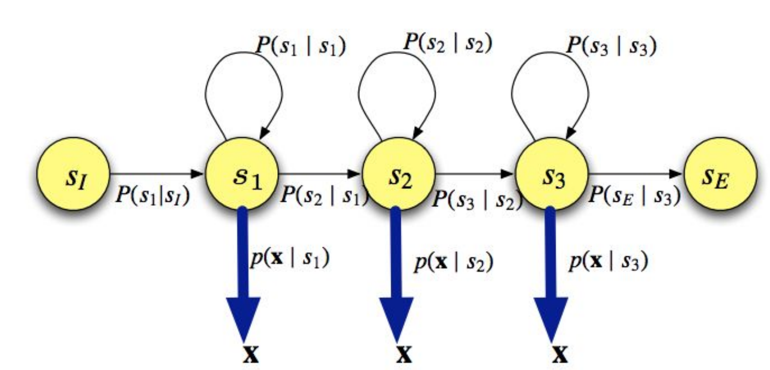
↓

Output: Generate ranked list of features from populated Q-table; choose top-N subset

Figure 3. Reinforcement learning procedure.

Benchmark Model: **GMM-HMM Classifier**

- Used 5-state Hidden Markov Model (HMM)
- Each state represented by a 5-mixture Gaussian Mixture Model (GMM), M = 5
- Output prediction score $b_j(x)$ for each emotion class



Single multivariate Gaussian with mean μ^j, covariance matrix Σ^j:

$$b_j(\mathbf{x}) = p(\mathbf{x} \mid s_j) = \mathcal{N}(\mathbf{x}; \mu^j, \Sigma^j)$$

M-component Gaussian mixture model:

$$b_j(\mathbf{x}) = p(\mathbf{x} \mid s_j) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(\mathbf{x}; \mu^{jm}, \Sigma^{jm})$$

Figure 4. Diagram of example GMM-HMM.

Training and Evaluation

- Randomly divide dataset: 70% for training set; 15% for validation set; and 15% for test set, with uniform distribution of emotion class labels in each; RL procedure tuned on training and validation sets only
- Ran RL procedure and chose Top-3, Top-6 subsets of features (N = 3, 6); train model on each subset
- Ensemble classifier averages predictions of Top-3, Top-6, and baseline classifiers, weighted by score
- Classification is correct if ground truth is contained in the top 2 scores outputted by the classifier

V. Results

- Measured per-class and overall classification accuracy (shown below), precision, and recall for each classifier trained on each feature subset
- Performed on both training (504 samples) and test (108 samples) sets.

Classifier *[features]*	Train	Test
Baseline *[0...12]*	0.649	0.520
Top-3 *[12,11,2]*	0.518	0.520
Top-6 *[12,11,2,6,9,10]*	0.644	0.533
Ensemble: Top-3, Top-6, Baseline	**0.709**	**0.613**

Table 1. Overall train and test classification accuracies for classifiers trained on specified feature subsets.

VI. Discussion and Future Work

- RL-based feature selection returns feature subsets (top-3, top-6) that better discriminate certain emotions than baseline features
- Ensemble achieves better overall and improved/commensurate per-class accuracy
- Results are expected, as we also take into consideration the “structure” of our dataset when approximating the optimal features
- RL-procedure removes the need to greedily pick features one-by-one (via forward/backward search, etc.)
- Next steps: collect more training/test data and measure effect on classification performance
- Future model: explore using LSTM-based RNNs to capture temporal variations in speech audio

Selected References

[1] R. Picard. *Affective computing.* Cambridge: MIT Press, 1997.

[2] Y. Wang, et. al. “Recognizing human emotion from audiovisual signals”, in *IEEE Transactions on Multimedia*, 2008.

[3] M. Pinol et. al. “Feature selection based on reinforcement learning for object recognition”, in *Adaptive Learning Agent Workshop*, p. 4-8, 2012.