



Classifying Emotions in Human- Machine Spoken Dialogs

Chul Min Lee, Shrikanth Narayanan,
and Roberto Pieraccini



Why Recognize Emotions?

- Increasing role of human-computer interaction: need to develop machines that can naturally communicate with humans.
 - Call center applications, Teleconferencing, Tutoring by machines, Entertainment
- Recognizing negative emotions in speech
 - Potential application to call centers
 - Providing feedback to an operator for monitoring purposes.
 - Voice mail messages
 - Sorting messages according to the emotions expressed by senders.
- Theoretical Reason: Studies show that emotion is related to human intelligence.
 - ⇒ Can enlarge the understanding of human cognition, psychology.



Results related to Our Work

- Dellaert, et al ('96): 36 % - 28 % classification error in 5 emotion states (LDA and k-NN)
- Petrushin ('99): ~77 % classification accuracy in two emotions of 'anger' and 'calm' (Neural networks)
 - Call center application
- Batliner, et al ('00): 89 % for actors' speech and 69 % for 'Wizard-of-Oz' cases (assuming the 'naïve' subject and they communication with real computer) with LDA.
 - 2 emotion states : 'emotional' and 'neutral'



Emotion Categories

- Since speech data does not have pre-defined emotion categories (real-world data), the definition would be intuitive and subjective.
- We focused on Two Emotion States
 - Negative: anger, frustration
 - Non-negative: complement of Negative, such as neutral, happiness, etc.
 - Turns out that most non-negative emotional speech are neutral (no emotion at all) in human-machine dialogs



Database and Preparation

- Most past research has used speech recorded from actors (fake emotions).
- A corpus of human-machine dialogs recorded from commercial application deployed by SpeechWorks
 - United Airlines baggage-desk database. (1187 Calls with approximately 7200 utterances)
- First, selecting calls by measuring total number of dialog turns and ASR accuracy.
- Subjective labeling of emotions
 - Listening test by two people according to 2 emotion states.
 - Choosing the utterances of which emotions are in agreement between 2 people.
 - Agreement Percentage: about 65%
- Final data
 - Female: 532 'non-negative' and 133 'negative'
 - Male: 392 'non-negative' and 122 'negative'



Acoustic Features

- This work focused on acoustic information
- There are 10 base features related to pitch (F0) and energy (acoustic features).
 - Utterance-level statistics
 - Pitch: mean, median, standard deviation, maximum, minimum
 - Energy: mean, median, standard deviation, maximum, range (maximum – minimum)
- Other useful features suggested by researchers
 - time/duration related information: speech rate, duration of voiced speech/unvoiced speech
 - Spectral information: energy in certain frequency range
 - voice quality: tense, harsh, and breathy voice



Feature Reduction by PCA

- Principal Component Analysis (PCA): One of popular feature reduction technique in pattern classification.
- To find underlying dimensions of feature space.
- Rationale
 - New or reduced feature set may perform better
 - Eliminate irrelevant features from base feature set.
 - Reduction of dimensionality
- How to compute
 - Compute covariance matrix from the full base feature set (d-dim)
 - Calculating eigenvalues and eigenvectors
 - Sort it according to decreasing eigenvalues
 - Form a matrix with the k eigenvectors corresponding to the k largest eigenvalues.
 - Preprocessing according to:

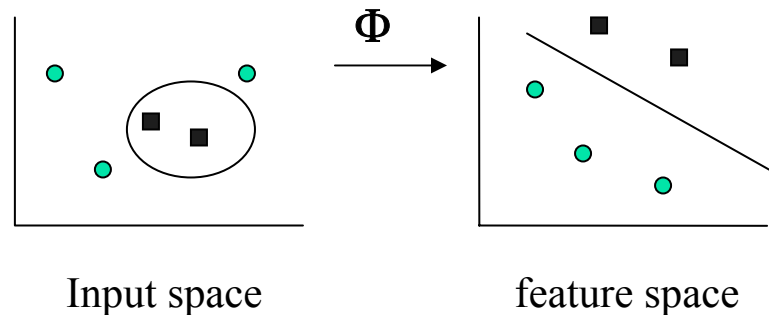
$$\mathbf{x}' = A^T(\mathbf{x} - \boldsymbol{\mu})$$

Support Vector Machines (SVM)

- Mapping the training data nonlinearly into a higher dimensional feature space via Φ , and constructing a separating hyper-plane.

- $\Phi : R^N \rightarrow F$

- **Linear algorithm is performed in F which has usually higher dimension than input data**



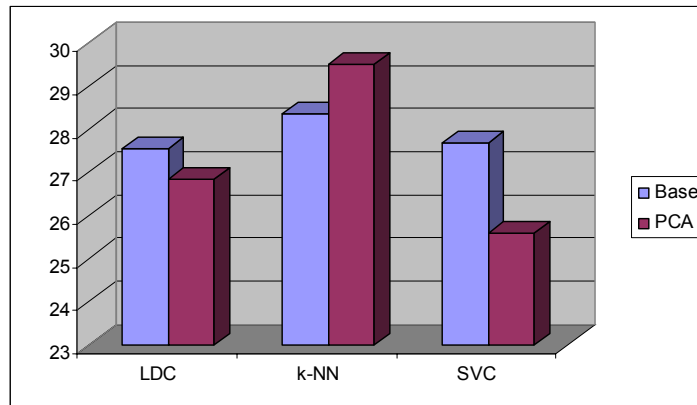
- SV classifier are based on the class of hyper-plane
- $(w \cdot x) + b = 0$, $w \in R^N$, $b \in R$



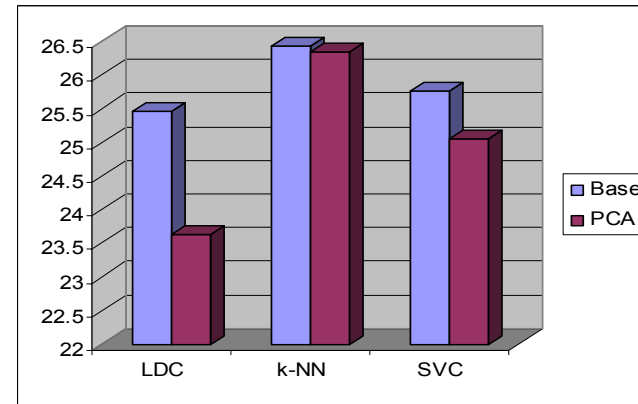
Experiments

- Pattern classification problem
 - Linear discrimination (LDC) with Gaussian class-conditional probability distribution
 - K-nearest neighborhood (k-NN)
 - Support vector machine classifier (SVC)
- Performance Criterion : Classification error
 - 10-fold cross validation
 - For the maximal use of available data (Problem of data sparsity)
 - Training data is randomly divided into 10 disjoint sets of equal size, and classifiers are trained 10 times, each time with a different set held out as a validation set.
 - The estimated error is the mean of these 10 errors.

Classification Error Results



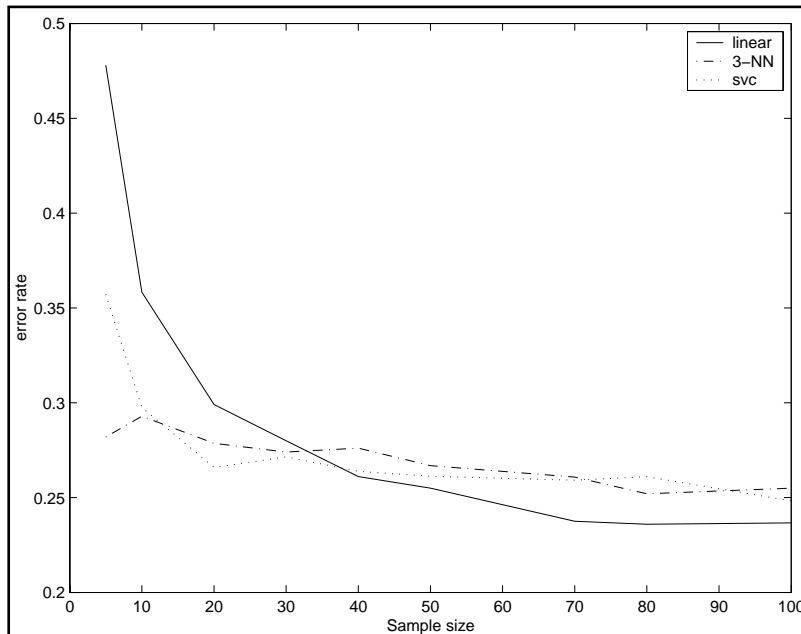
(a)



(b)

- Classification error rates for (a) female and (b) male data
- The numbers of neighborhood k in k -NN are set to be 3 for female and 7 for male data and for SVC, polynomial kernels with degree of 1 were used (all the parameters are set by 10-fold cross validation)
- The reduced dimension for PCA was 6 for both female and male data.
- Note that PCA feature sets showed comparable performance even in reduced feature dimensions.

Learning Curves



- Learning curves for 3 classifiers.
- Training data was chosen 10 times in random manner and then averaged.
- The generalization of SVC is high even in small number of training data due to the fact that decision by SVC is made by usually by small number of support vectors.



Conclusions and Future Research

- Emotion recognition as a pattern recognition problem. => we can apply many techniques in pattern classification to the problem, e.g. SVC.
- Comparable performances by PCA may show a new way to search for useful feature set
- Combining with linguistic information or other knowledge sources: multi stream of data
 - Swear words, recognition of repetition, ASR accuracy, etc.
 - How to combine multi stream data/decision.
- Applying other classification methods since the definition and categories of emotions are fuzzy and uncertain (no clear-cut boundaries between emotion categories).