



COMBINING ACOUSTIC AND LANGUAGE INFORMATION FOR EMOTION RECOGNITION

Chul Min Lee, Shrikanth Narayanan

University of Southern California

and

Roberto Pieraccini

SpeechWorks International



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.



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

- 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)



Data Preparation

- 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



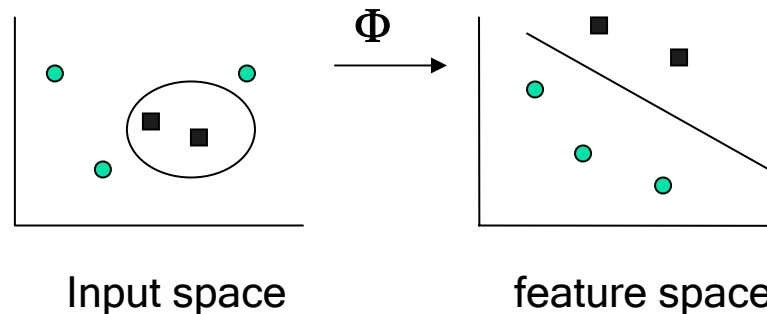
Feature Reduction by PCA

- 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
 - 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
 - $(w \cdot x) + b = 0$, $w \in R^N$, $b \in R$

Combination of acoustic and language information

- Need for combining acoustic and language information to improve the recognition of emotions
 - Previous work by Batliner, et al. ('96) : used topic repetition as 'language' information to combine with acoustic information.
- The emotion information conveyed by words were combined with that from acoustic features.
 - Obtained the emotional 'keywords' by calculating 'emotional salience, defined as mutual information between a specific word and emotion category.

$$sal(w_n) = I(E; W = w_n) = \sum_{j=1}^k P(e_k | w_n) \log_2 \frac{P(e_k | w_n)}{P(e_k)}$$



Salient Words

Word	Salience	Emotion
You	0.73	Neg.
No	0.56	Neg.
Computer	0.47	Neg.
Damn	0.47	Neg.
Baggage	0.25	Non-Neg.
Right	0.01	Non-Neg.

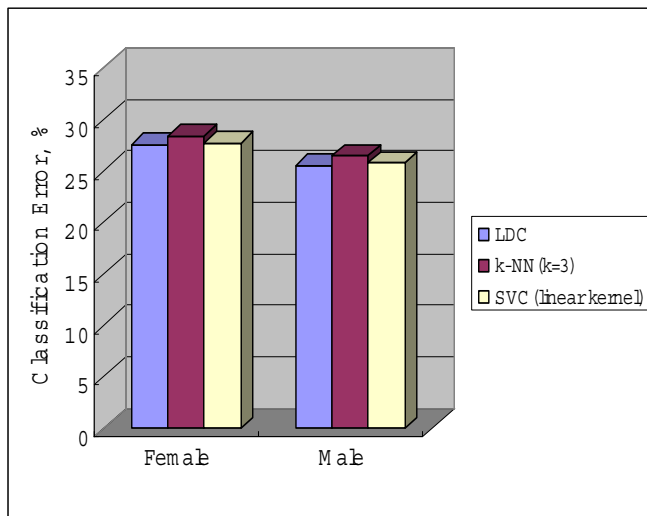
- A partial list of salient words.
- ‘Emotion saliencce’ is a measure of the amount of information that a specific word contains about the emotion class.
- “Emotion” represents maximally correlated emotion class given words.



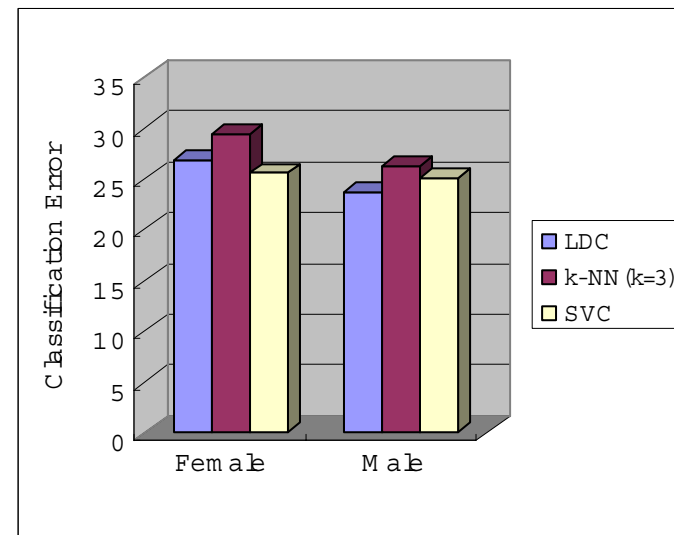
Experiments 1

- Acoustic information only
 - Classification methods
 - Linear discriminant classifier (LDC)
 - K-nearest neighborhood classifier (k-NN)
 - Support vector machine classifier (SVC)
 - Two feature set
 - Base feature set: 10 acoustic features from pitch and energy
 - Feature set by PCA
 - 260 data for female and 240 data for male were selected and classification error rates were computed by 10-fold cross-validation
 - separating training data into 10 disjoint sets of equal size, and error was computed 10 times, each time with a different set held out as a validation set
 - Error rate is the mean of these 10 errors.

Base vs. PCA Features



(a)



(b)

Classification error for (a) Base feature set and (b) Feature set by PCA (dimension = 6)

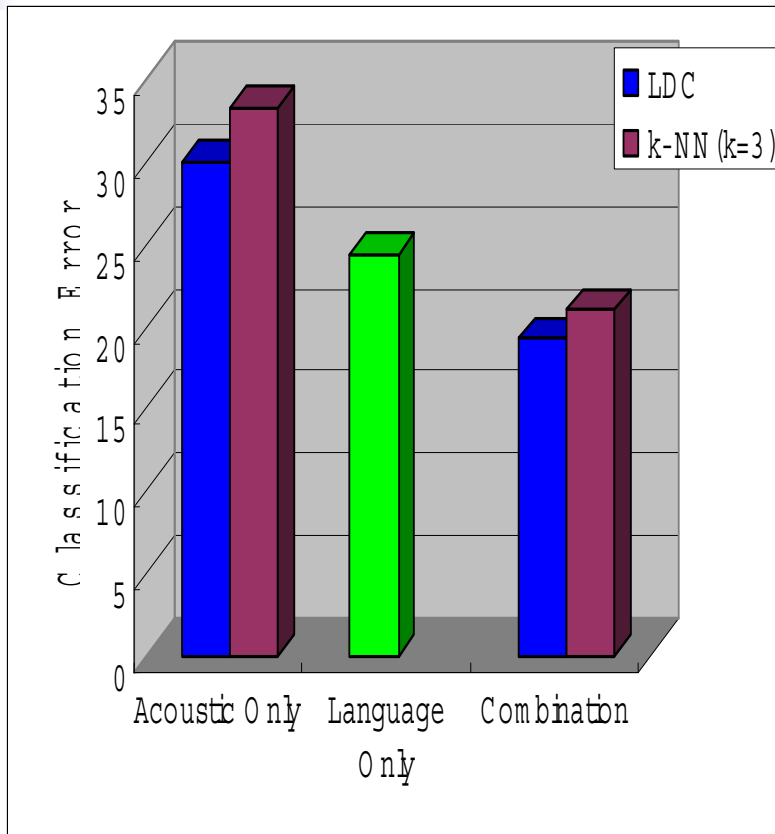
- Comparable performance by PCA even in the reduced dimensionality.



Experiments 2

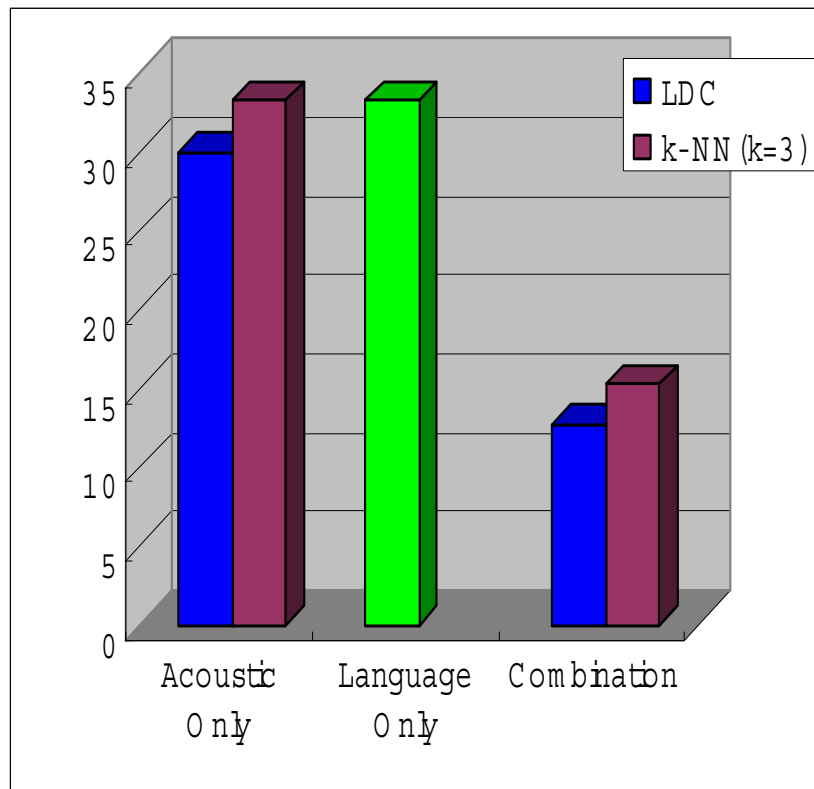
- Combination of acoustic and language information
 - Acoustic information
 - Two classification methods : LDC and k-NN
 - 10 base feature set was used.
 - 200 training data 40 test data each from female and male were selected 10 time randomly and the estimated error rate was obtained by averaging those results.
 - Language information
 - Two different training sets were used in the estimation of emotional salient words and $P(E|W)$.
 1. All the data was used as training set (1179 utterances)
 2. The same training data was used as acoustic information.
 - Same test data was used in both acoustic and language information.

Combined information (1)



- Classification error in female data for the same training data
- Improvement in combined information
35.8% in LD
36.8% in k-NN

Combined information (2)



- All the data was used for training language information.
- Improvement in combined information
57.5% in LDC
54.1% in k-NN

Conclusions and Future Works

- 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
- By combining language information, we can improve the recognition of emotions
- Future works
 - Combination methods for acoustic and language information
 - Applying other classification methods since the definition and categories of emotions are fuzzy and uncertain.