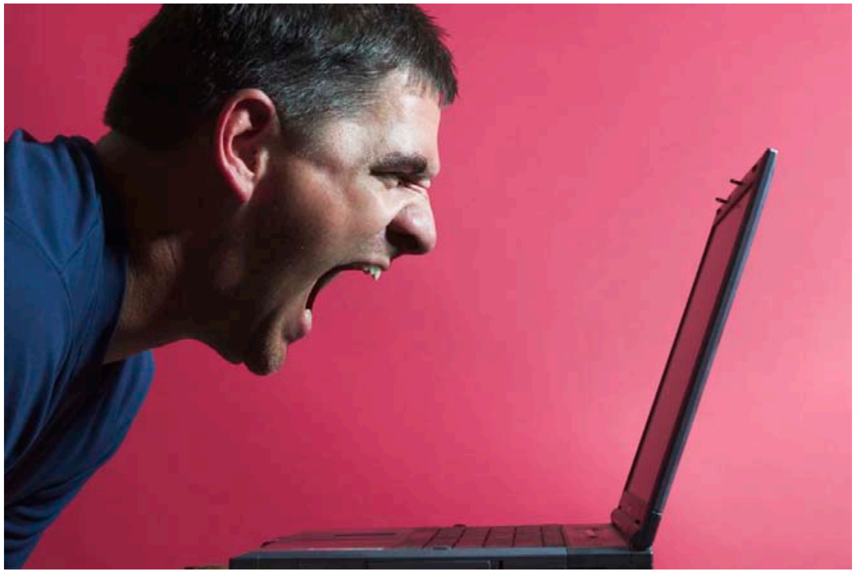
Emotion Recognition

from speech



Carlos Busso
Prof. Shri Narayanan

Some examples..
Lost baggage call center

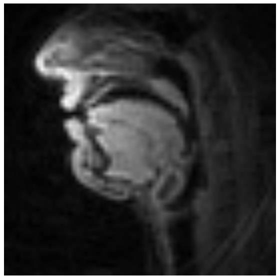
More examples
Child-machine Interactions

CONFIDENT **vs.** UNCERTAIN

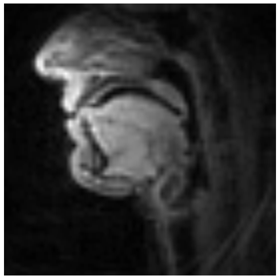
 

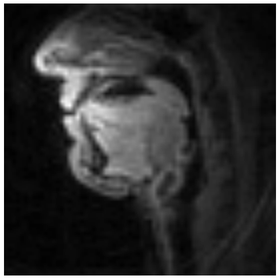
More examples
Visualizing using MRI..



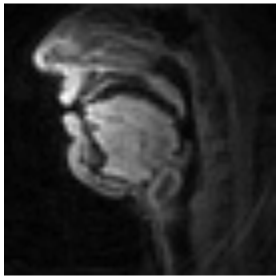
neutral



sad



angry



happy

Human Communication

- Human communication involves a complex orchestration of cognitive, physiological, physical, social processes

- Information resides at multiple time scales, through multiple cues
  - Inherently multimodal: natural communication involves speech, facial/hand gestures, head movement, postures,..

- Spoken language carries crucial information: intent, desires, emotions

Decoding Human Communication Cues is a Multi-level Mapping Problem

Automatic Speech Processing Solutions: mapping speech to words, and beyond

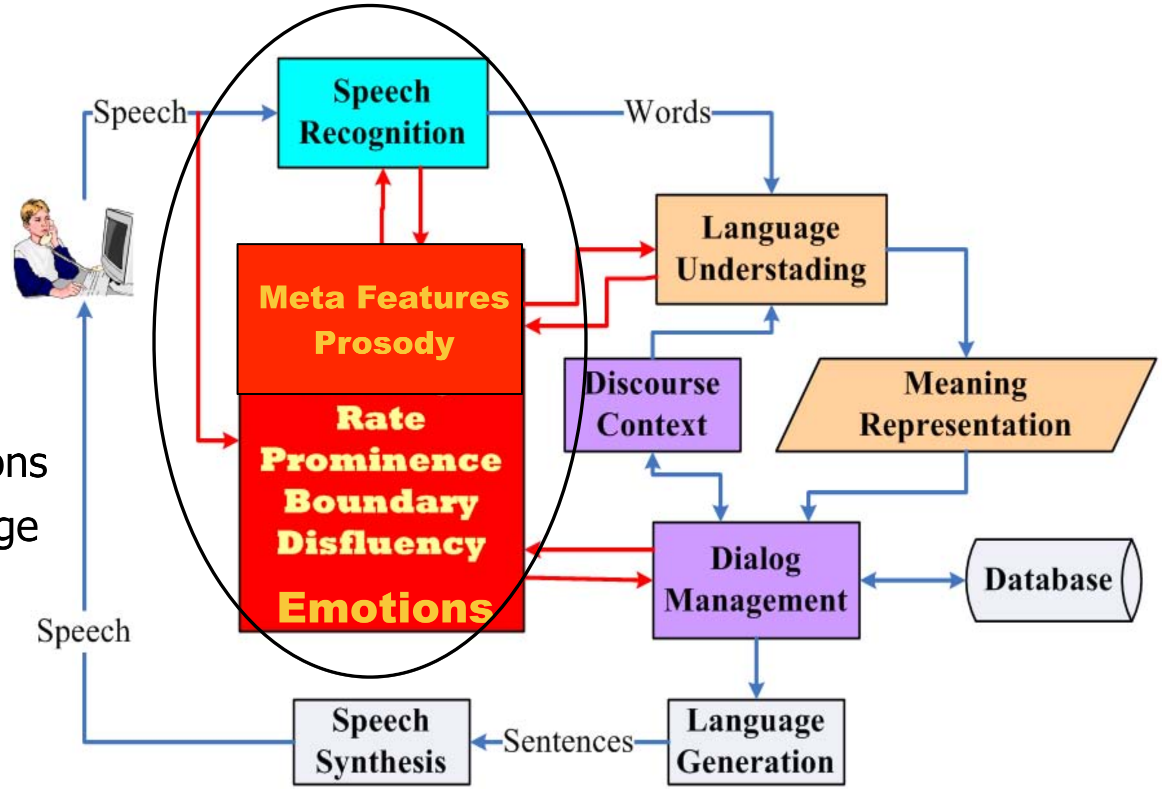
Significance:

Natural spoken language is the primary means of human communication: to negotiate, to seek information, to issue orders and to resolve conflicts

A tightly integrated approach to speech processing: Recognize

- **What:** spoken language content
- **Who:** speaker identity, **and**
- **How:** speaking style and emotions

automatically from spoken language



Why study emotion or attitude?

- Emotions play a crucial role in human interaction
- Knowing the user's emotional state should help to adjust system performance
- User can be more engaged and have a more effective interaction with the system
- Crucial for understanding and modeling both individual and social cognition
  - Emotional (vs. cognitive) reasoning
  - Emotion is reflected in our body
  - Our emotions change the minds of others
  - People rely on emotion for making decisions

Applications

- Call centers
  - Quality of service
  - Coping with frustrated users
- Robots
  - Sense and convey emotions
- Artificial animated agents
  - Sense and convey emotions
- Education
  - Detect frustration
- Games
  - Expressive characters
- Observational practices
  - (e.g. therapy sessions)
  - Diagnosis and coaching

Analysis & perception
Emotional perception
Appraisal theory

Recognition
Emotion recognition

Synthesis
Emotional speech synthesis
Manipulation of body/facial movement
Expressive facial animation

Emotion Research @ SAIL

- SAIL: Signal Analysis & Interpretation Lab.
  - http://sail.usc.edu
- Speech and emotions
  - Analysis, recognition, synthesis
- Speech production
- Multimodal processing

Work in collaboration with USC SAIL members & graduates

- Dr. Shri Narayanan, Dr. Sungbok Lee
- Matt Black, Jeannette Chang, Michael Grimm, Abe Kazemzadeh, Sam Kim, Chi-Chun (Jeremy) Lee , Emily Mower, Angeliki Metallinou, Ilene Rafii, Michelle Dee, Carlos Busso
- SAIL PhD grads/alumni: Murtaza Bulut, Michael Grimm, Dagen Wang, Serdar Yildirim, Chul Min Lee

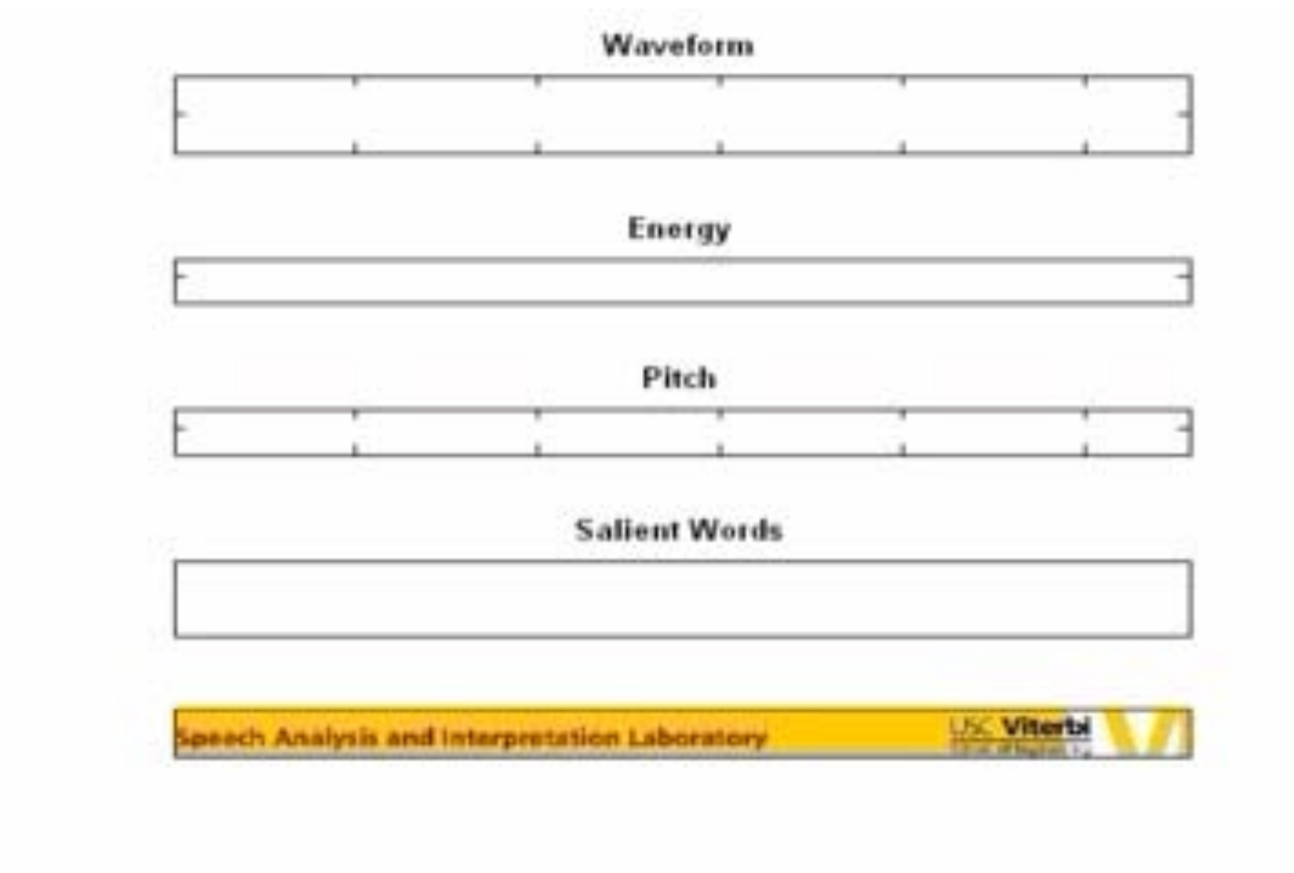
Emotion recognition

focus on speech

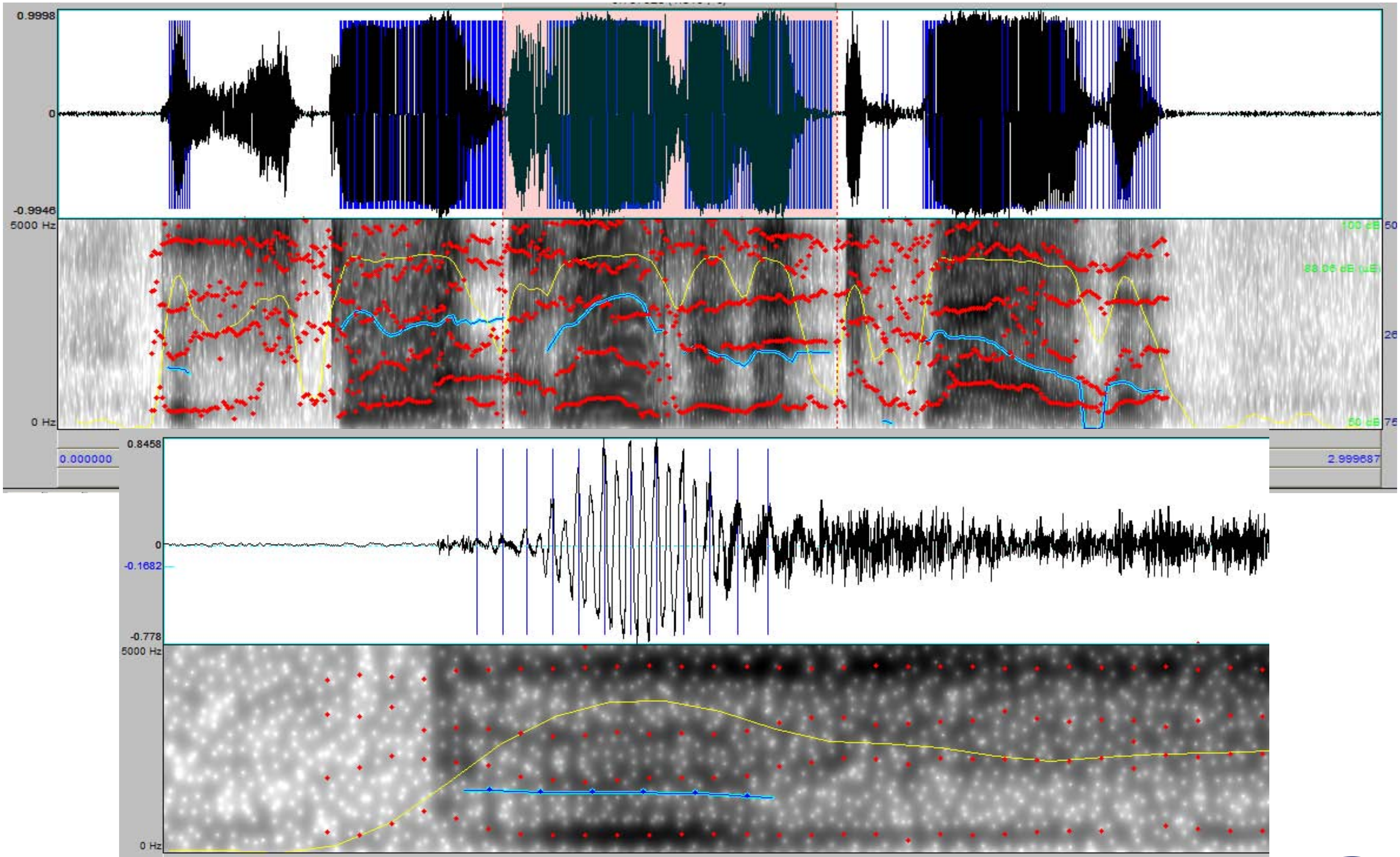
Outline

- Overview
- Challenges in emotion recognition
- Proposed approaches to emotion recognition
- Conclusions

Automatic emotion recognition from speech



Speech: a multimodal signal



Emotion recognition in the lab

- Databases
  - Acted data
  - Categorical representation of emotions
  - Few speakers
- Limited data
- Features
  - Many features are selected
  - Feature set is reduced (pca, fisher linear discriminant, sequential forward feature selection, etc…)
- Results
  - From 50% - 85% depending on the task [Pantic_2003, Cowie_2001]

Emotion recognition in real applications

- Too much variability
  - Speaker dependency
  - Emotional descriptors
  - Acoustic confusion between categories
  - Differences in acoustic environments
- Results are strongly dependent on the recording condition
- Models are not easily generalized to other databases or on-line recognition task

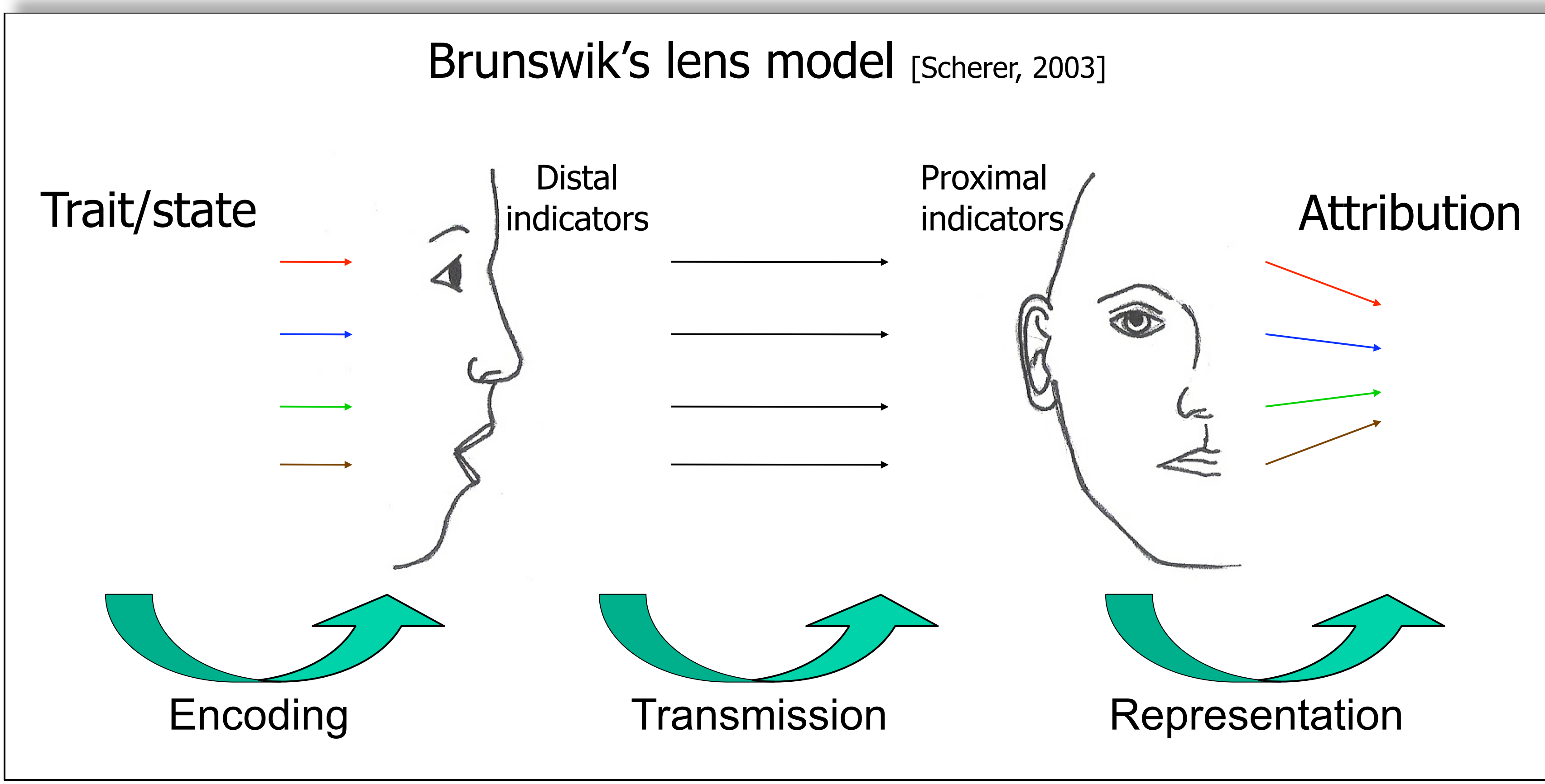
Emotional models do not generalize!!!

Outline

- Overview
- Challenges in emotion recognition
  - Representation
  - Databases
  - Speech normalization
  - Features
  - Models
- Proposed approaches to emotion recognition
- Conclusions

How to describe emotions? (1/4)

- Expression and perception of emotion is a complex process
  - Intended emotion ≠ perceived emotion
  - Representation depends on the listeners



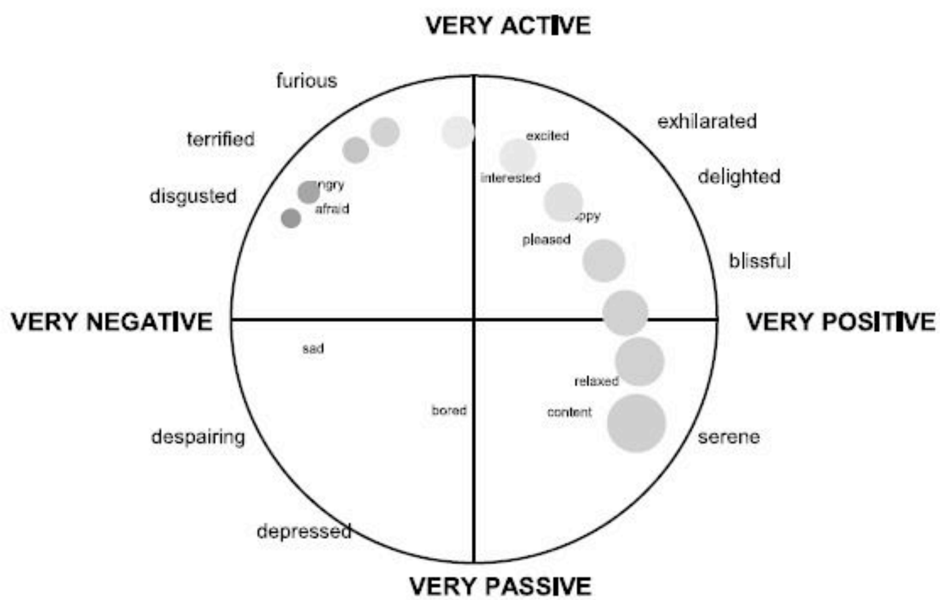
How to describe emotions? (2/4)

- One Pragmatic Approach: Categorical Emotional States
  - Six basic emotions (happiness, sadness, fear, anger, surprise, disgust)
  - Mixed emotion, in the order of hundreds (e.g., content, amused, etc…)
  - Tradeoff between inter-evaluator agreement and description accuracy

- Define domain/application-dependent emotional states:
  - Negative and non-negative in Call Center data
  - Frustration, politeness, attention for child-machine interaction systems
  - Cooperation in negotiation tasks, Like/dislike in opinion polling
  - Hot spots, engagement in meetings

How to describe emotions? (3/4)

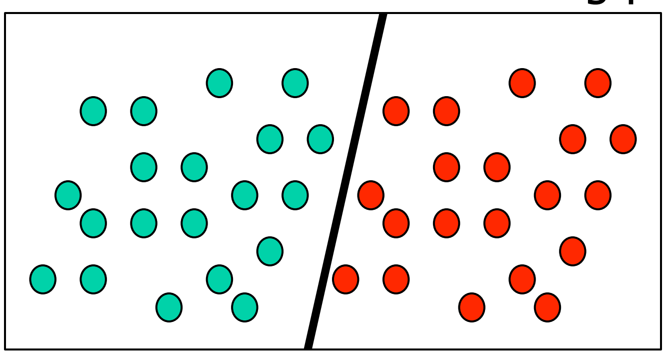
- Another approach: “Primitives based”
- Dimensional attributes
  - Valence, activation/arousal, dominance/control
  - Better inter-evaluator agreement
  - Can help track dynamic variation
  - Not very useful for certain applications



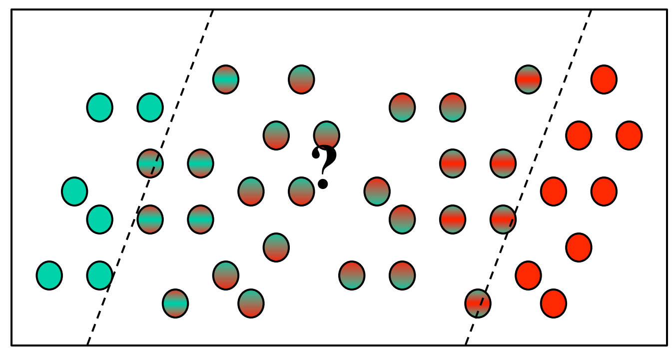
How to describe emotions? (4/4)

- Real emotional label or attribute values are unknown
- Need to use human evaluators
  - Who, Where, How many
- It should be view as an approximation
  - Perceived emotion may differ from intended emotion [Busso, 2008]

Conventional machine learning problem

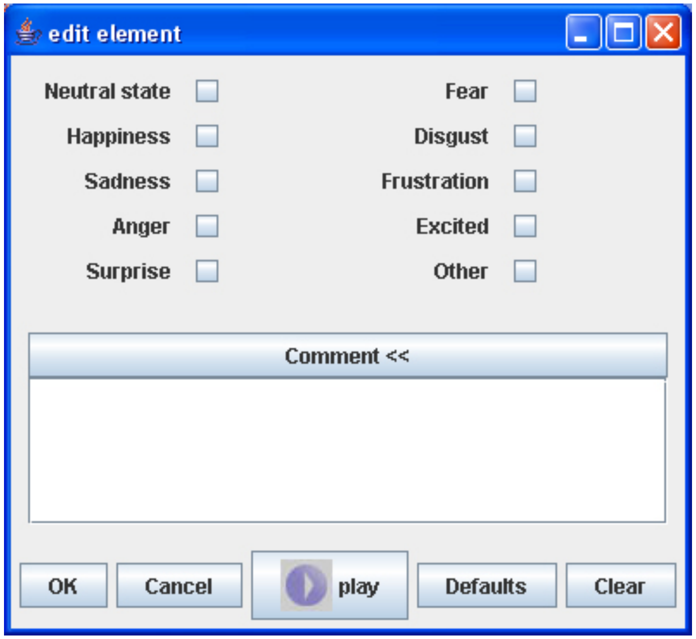


Emotion recognition



Boundaries are blurred!

Representation



Examples

[fru; ()] [ang; ()] [neu; ()]

[fru; ()] [oth; (exasperated)] [neu; ()]

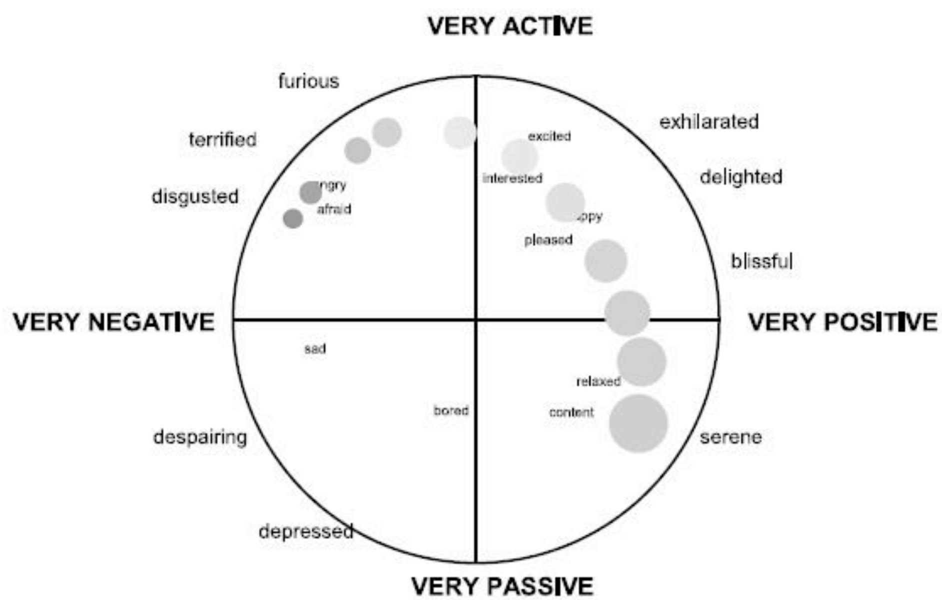
[fru; ()] [ang; ()] [oth; (annoyed)]

[ang; ()] [ang; ()] [ang; ()]

Time scale

- The best time scale to evaluate/analyze the emotion is not clear
  - Emotional content may change within a turn
  - Sentence, chunk, words

- Continuous evaluation of emotion
  - FEELTRACE [Cowie, 2000]



Emotional databases

- Availability of appropriate emotional databases is a major limitation for scientific research and technology development
  - Genuine realizations
  - Integrated information from relevant modalities
  - Models that generalize across domains/applications

- Acted databases versus spontaneous (natural) databases
  - Tradeoff versus Naturalness and control

Natural databases

- A variety of sources for spontaneously elicited material
  - Broadcasted television programs (VAM [Grimm, 2007], EmoTV [Abrilian, 2005], Belfast [Douglas-Cowie, 2003])
  - Recording in Situ (Lost luggage [Scherer, 1997])
  - Recalling emotions ([Amir, 2000])
  - Wizard of Oz (SmartKom [Schiel, 2002])
  - Games (FAU AIBO [Steidl,2009], EmoTaboo [Zara, 2007])
  - Carefully design human-machine interaction (SAL [Cowie, 2005])
- Core limitations
  - Ethical issues (i.e., inducing emotions)
  - Copyright problems
  - Constrained to specific domains
  - Lack of control over the microphones and camera locations
  - Noisy visual and/or acoustic background
  - Incomplete information from modalities

Acted databases

- Acting and actors have played a key role in the study of emotions
- Current techniques to record databases from actors have limitations
  - Use of naïve or inexperienced subjects
  - Lack of contextualization
  - Emotional descriptors ("read this sentence portraying anger")
  - Unfamiliar tasks to the actors
- Can specific acting methods be used to mitigate the limitations of recording emotional data from actors?
- Acting provide opportunities to tackle the problem in a systematic and controlled fashion [Enos, 2006]
- How?
  - Using better elicitation techniques
  - Make use of acting techniques
  - Make connection with real-life scenarios
  - Create suitable social settings in the recording

Databases used in our studies (1/2)

Simulated database: To aid feature analysis

- LDC Database
  - Linguistic Data Consortium Emotional Prosody and Transcripts Database
    - Recorded from actors (4 female, 3 male)
    - Utterances contain standardized contents (date, number)
    - 15 emotional states: neutral, happy, angry, sad, disgust, boredom, etc.
- Actors Database
  - Recorded from actors in our lab: speech, facial expressions, head, body postures, motion capture, articulatory data..
    - 4 emotions: anger, happiness, sadness, and neutral
- German emotional speech [Burkhardt, 2005]
- Interactive emotional dyadic motion capture database (IEMOCAP)

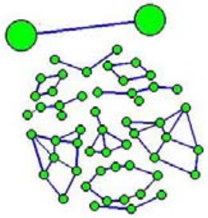
Databases used in our studies (2/2)

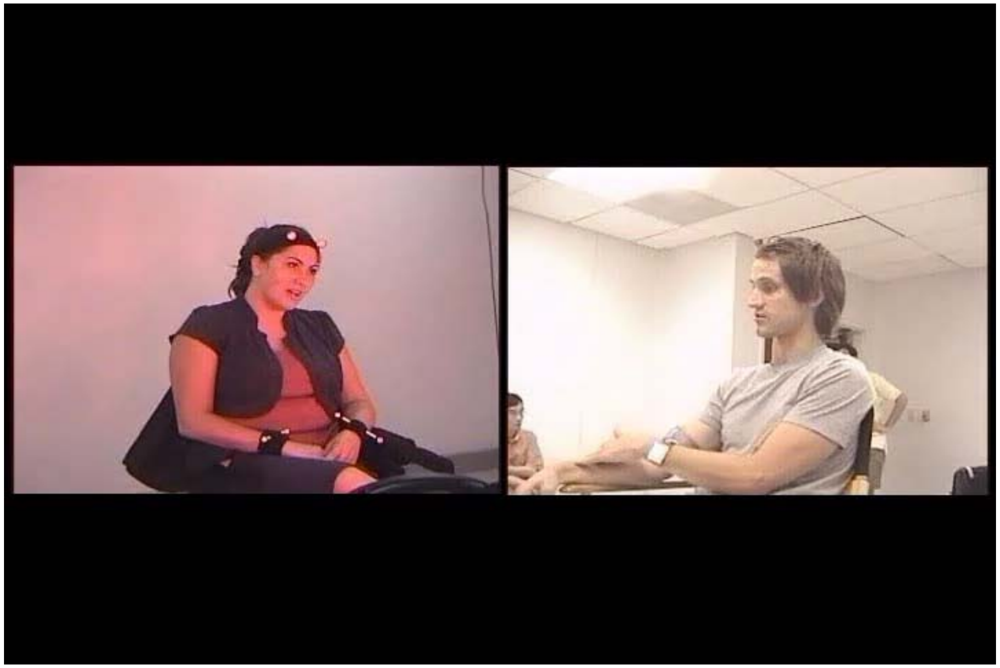
Natural Database: To aid recognition experiments

- Call Database
  - A corpus of human-machine dialogs recorded from commercial application (1187 Calls with approximately 7200 utterances)
  - Useful in the design of real-world applications
  - We defined 2 emotions (negative vs. non-negative)
- Broadcast television show (VAM)
- Movie Database
- Children's Games Database: 260 dialogs from 150 children 8-14 years playing voice activated computer game (Where is Carmen SanDiego)
- Meetings & small group discussions
- FAU AIBO, children interacting with robots

Interactive emotional dyadic motion capture database (IEMOCAP)

- Features
  - Dyad interactions
  - 5 sessions 2 actors each
  - ~12 hour of data (read, scripted and spontaneous)
  - Emotions were elicited in context
  - Markers on the face, head and hands
  - Happiness, sadness, anger, frustration and neutral state





* C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "*IEMOCAP: Interactive emotional dyadic motion capture database*," Journal of Language Resources and Evaluation, vol. In press, 2008.

Benchmark databases

- FAU AIBO
  - The INTERSPEECH 2009 Emotion Challenge [Schuller, 2009]
- Recent trend in the community is to share new databases

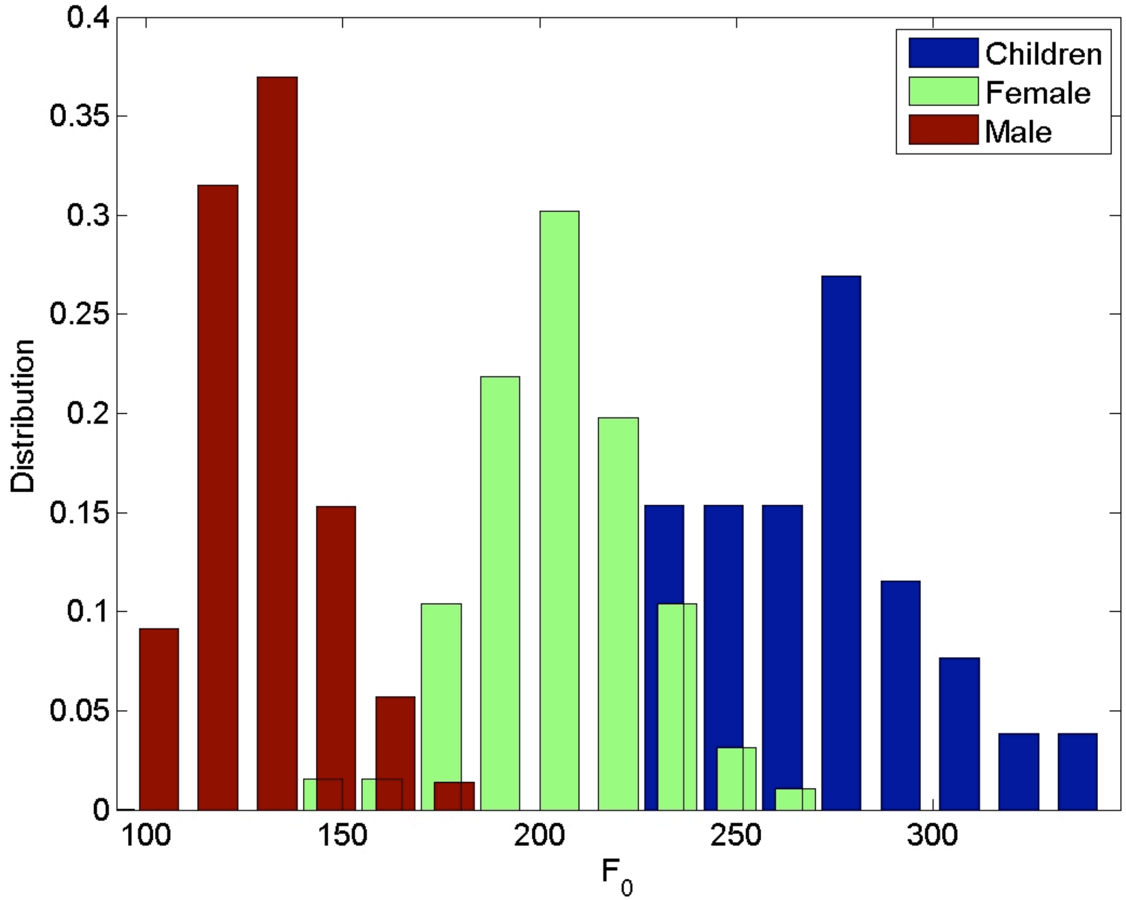
Speaker normalization

- Normalization is needed to reduce intrinsic variability
- Goals
  - Remove speaker and environment variability
  - Preserve emotional discrimination

- Recording condition
  - Compensate different recording setting (e.g., energy gain)
  - Telephone, mobile devise, far-field speech, close-talking microphones
- Inter-speaker variability
  - Gender differences
  - Inter-speaker differences (e.g., larynx)

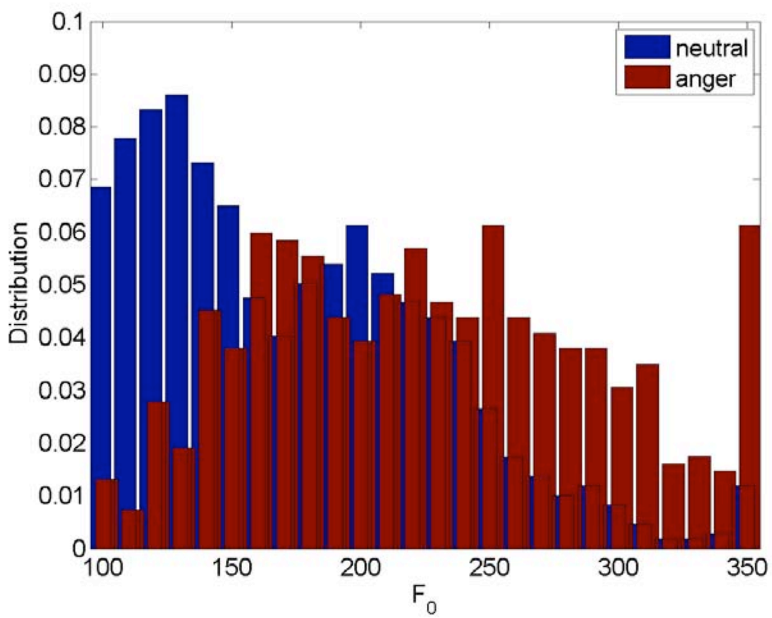
Fundamental frequency mean (1/2)

- Neutral speech for men, women and children (TIMIT, FAU AIBO)

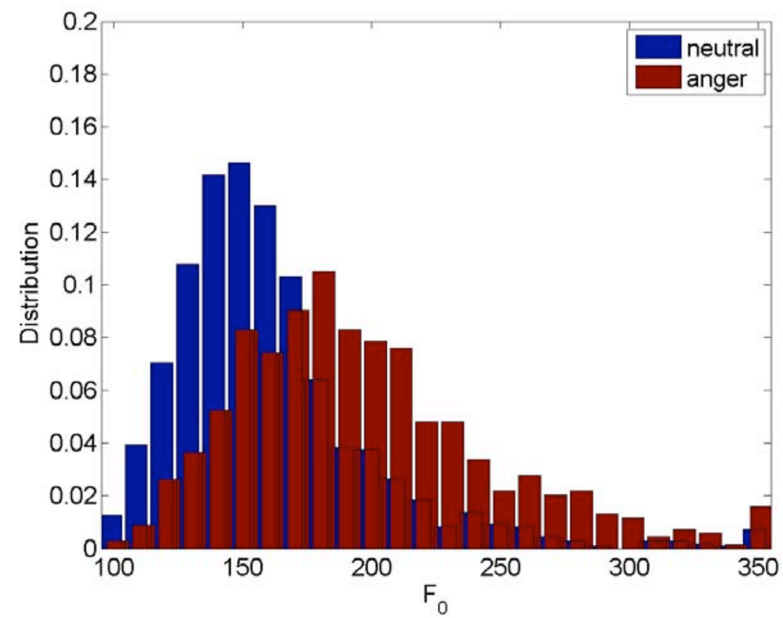


Fundamental frequency mean (2/2)

- Angry versus neutral speech (IEMOCAP)



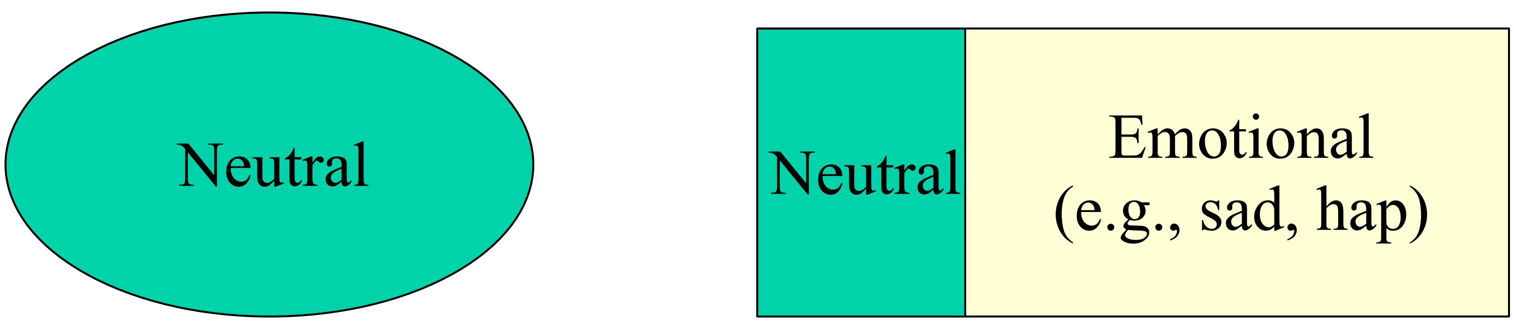
Without normalization



With normalization

Speaker dependent normalization

- Energy
  - Energy of neutral speech of the emotional data match the energy of the reference
- Pitch
  - Mean pitch of the neutral speech of the emotional data match the mean pitch of the reference

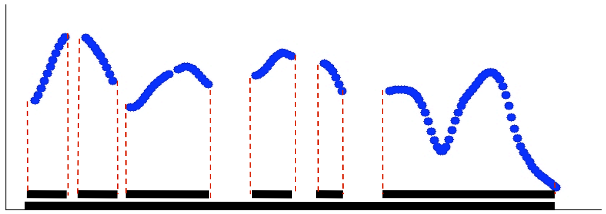


Reference database

Speech from one subject

Features emotion recognition

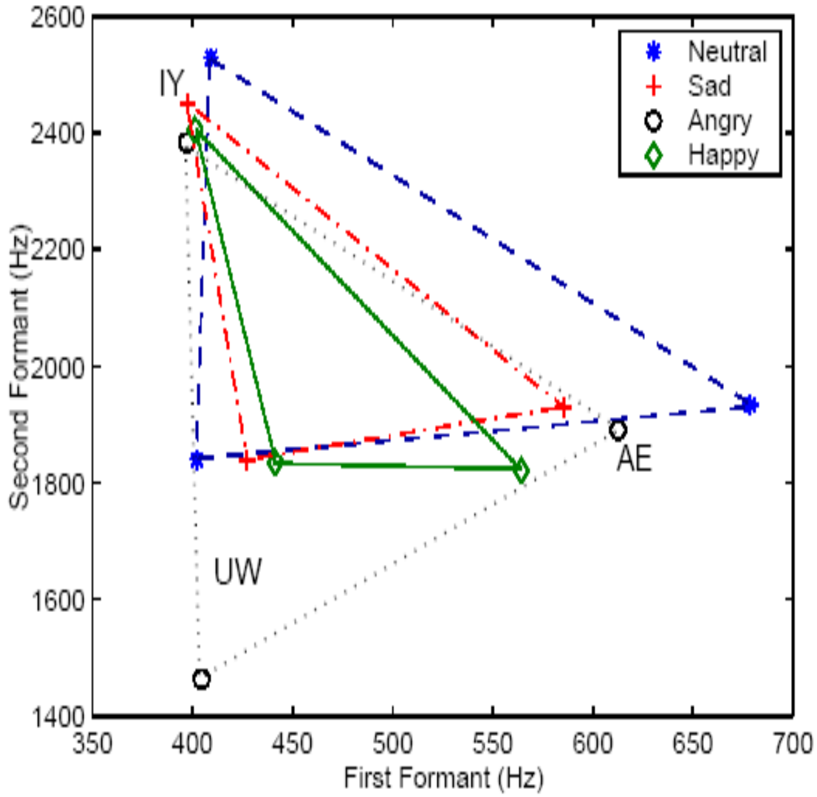
- Supra-segmental acoustic features
  - Related to prosody (Pitch, energy, and duration)
  - Most investigated features in association with emotion
    - Pitch: mean, median, standard deviation, maximum, minimum, range (max-min), linear regression coefficient
    - Energy: mean, median, standard deviation, maximum, minimum, range, linear regression coefficient
    - Duration: speech-rate, ration of duration of voiced and unvoiced region, duration of longest voiced region
    - Formant: F1, F2 and their bandwidth BW1, BW2.



Features emotion recognition

- Segmental acoustic features
  - Related to short-term spectrum of speech
  - Variations across emotions in spectral features, especially vowel sounds (Lea Leinonen and Tapio Hiltunen '97)
  - Mel-frequency cepstral coefficients (MFCC)
  - Mel filter bank (MFB) [Busso, 2007]
- Voice quality features
  - e.g. harsh voice, tense voice, modal voice, breathy voice, whispery voice, creaky voice and lax–creaky voice
  - Jitter, HNR, shimmer

Emotion Dependencies at the Segment level: Vowel Triangle Example



- First and second formant frequencies for the three vowels, /IY/, /AE/, /UW/ for various emotions
- We can observe that distinct constellation for different emotions
  - Emotions have different effects on different phonemes
- Notice that the lower vowels /AE/ and /UW/ more affected by emotions than high vowel /IY/.

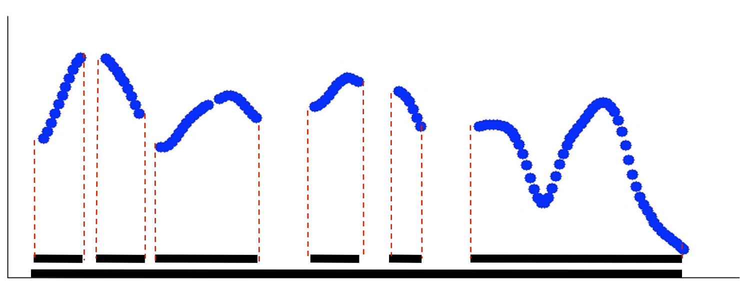
Practical consideration about features

- How many features to use?
  - People have used more than 4K features to recognize emotions
  - Risk of overfitting
- Feature selection
  - Forward or backward feature selection
  - Generic algorithms
  - Evolutionary algorithms
  - Linear discriminant analysis
  - Principal component analysis
- Do we really want to find the best features for **this** database?

F0 features

- Several databases:
  - Neutral reference (WSJ1) [Paul, 1992]
  - EMA (680 sentences, 3 speakers, neu, sad, hap, ang) [Lee,2005]
  - EPSAT (4738 sentences, 8 speakers, neu, sad, hap, bor, dis, fea, pan, cold ang, hot ang, des, ela, int, sha, pri) [Liberman, 2002]
  - GES (535 sentences, 10 speakers, neu, sad, hap, ang, bor, dis, fea) [Burkhardt, 2005]

- Which F0 feature is better?



Feature selection

- Finding the most emotionally prominent features from pitch
  - Logistic regressions

$$E(Y \mid x_1,\dots,x_n) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\dots\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\dots\beta_n x_n}}$$

$$g(x) = \ln\left[\frac{\pi(x)}{1+\pi(x)}\right] = \beta_0 + \beta_1 x_1 + .. + \beta_n x_n$$

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{-y_i}$$

$$G = -2\ln\left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right] \sim \chi^2$$

  - Comparing two nested models

$$H_0 : \beta_n = 0 \qquad g_0(x) = \beta_0 + \beta_1 x_1 + .. + \beta_{n-1} x_{n-1}$$

$$H_1 : \beta_n \neq 0 \qquad g_1(x) = \beta_0 + \beta_1 x_1 + .. + \beta_{n-1} x_{n-1} + \beta_n x_n$$

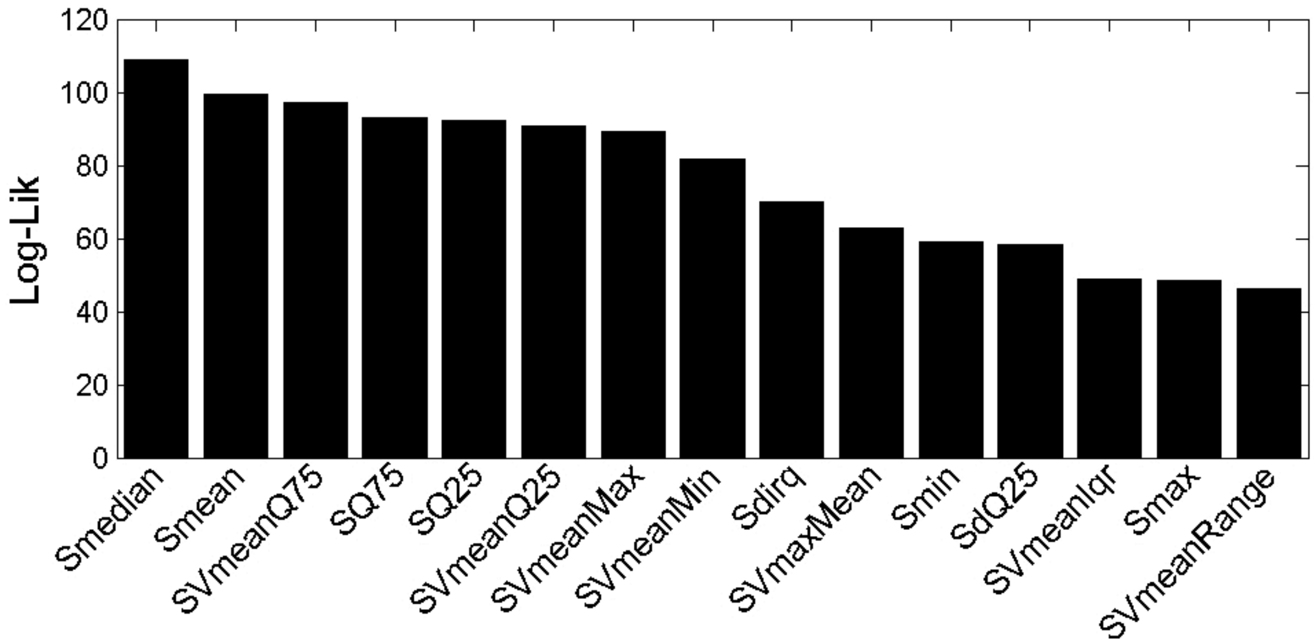
Emotionally salient F0 statistics (1/2)

- Run logistic regression with only one feature at a time (neutral vs. anger)

$$H_0 : \beta_1 = 0 \qquad g_0(x) = \beta_0$$

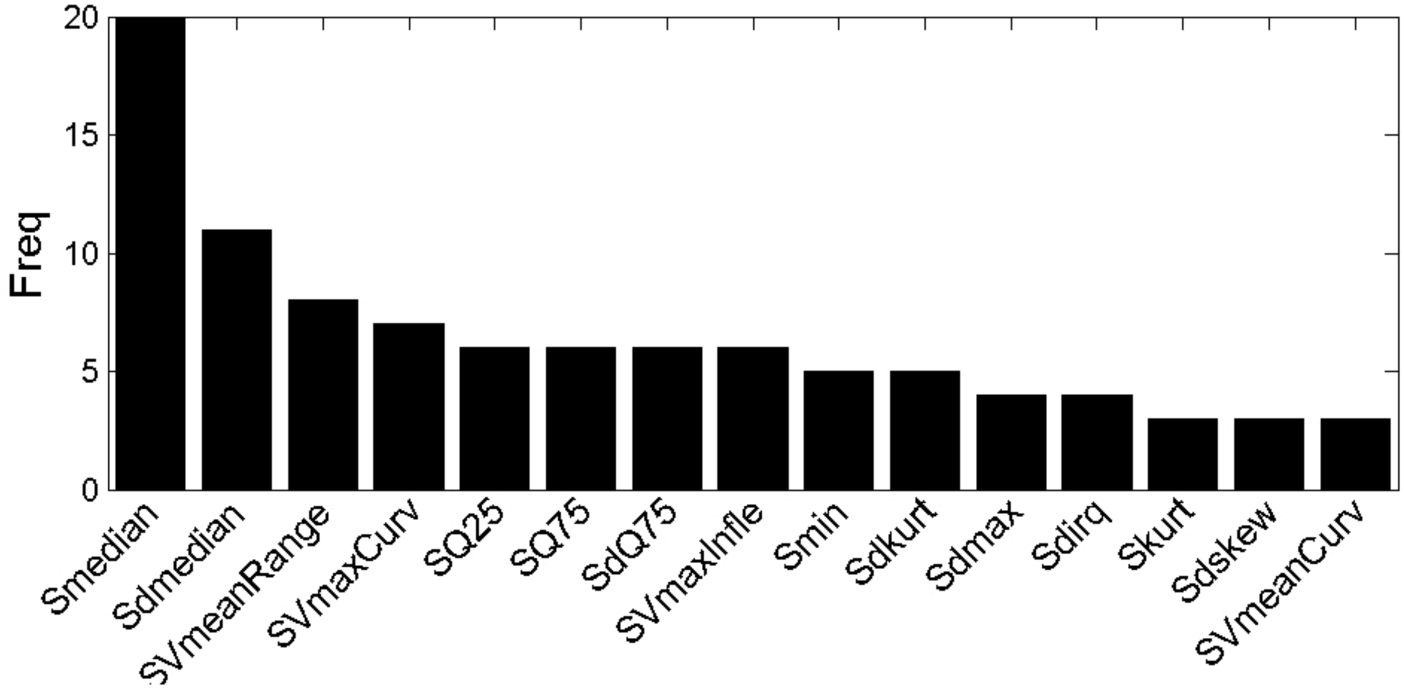
$$H_1 : \beta_1 \neq 0 \qquad g_1(x) = \beta_0 + \beta_1 x_1$$

- Measure the improvement in the model in terms of the Log-Likelihood ratio test



Emotionally salient F0 statistics (2/2)

- Experiment 2:
  - What about overlapping information between features?
  - Run logistic regression with FFS (neutral vs. anger)
  - Count the number of time that each feature was selected



Models

- Dynamic modeling
  - Short frame by frame
- Static modeling
  - Sentence level features
- Machine learning technique used for emotion recognition
  - Linear discriminators
  - Gaussian Mixture Models (GMM)
  - Hidden Markov models (HMM)
  - Neural network (NN)
  - Bayes classifiers
  - Fuzzy classifiers
  - Support vector machines (SVMs)

Performance of emotion recognition system

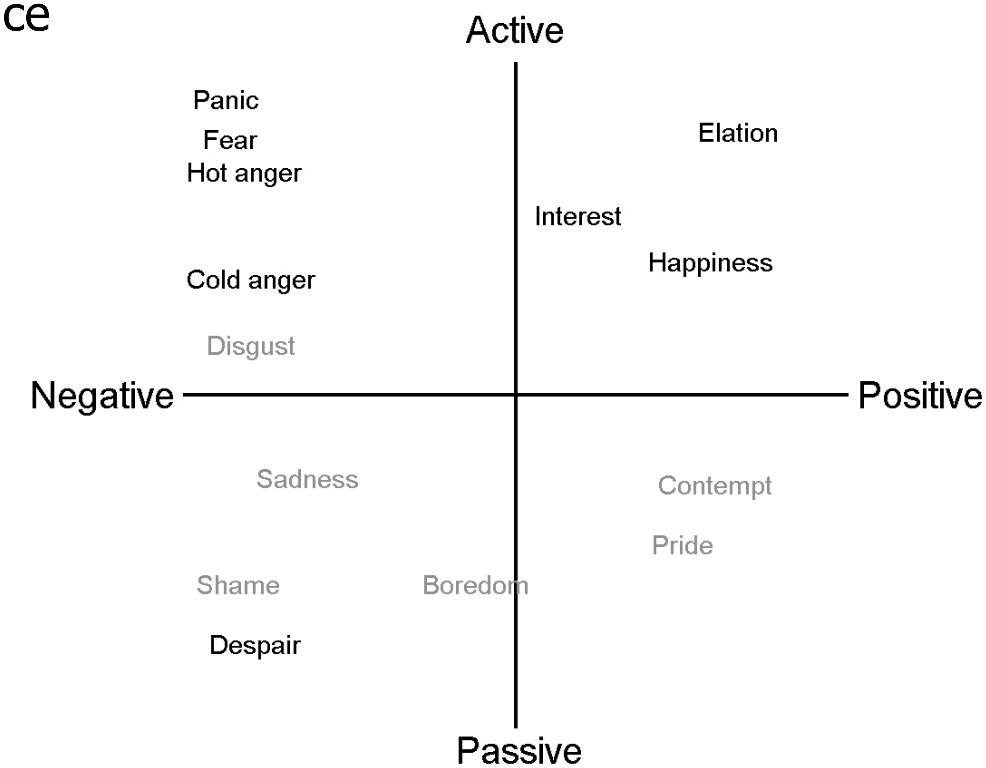
- From 50% - 85% depending on the task [Pantic_2003, Cowie_2001]
- Upper bound: subjective human evaluation

	Ang	Hap	Neu	Sad	Other
Ang	82	2	3	1	12
Hap	12	56	7	6	19
Neu	8	1	74	14	3
Sad	5	1	20	61	13

- EMA database (acted)
  – 68.3% accuracy by humans (4 subjects)

Being aware of limitations

- Speech is just one modality
  - ✓ Activation
  - x Valence



What is next?

- Dynamic analysis of emotion in dialog (context)
- Tracking state shift
  - Changing in the emotion rather than emotion itself
- Join model of emotion in multi-person meetings
  - How my emotions change your emotions

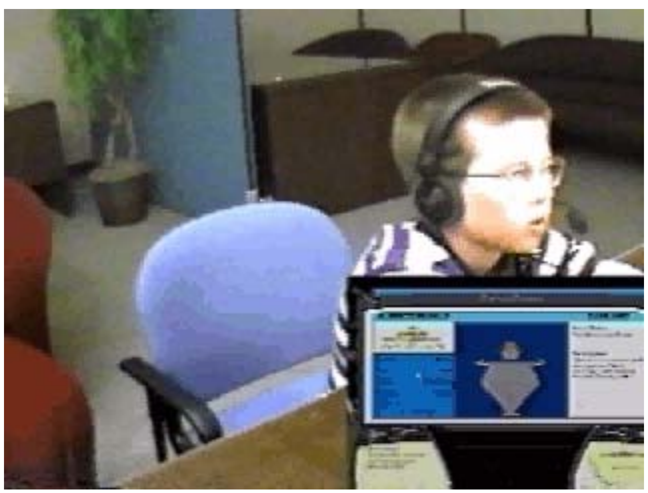
Outline

- Overview
- Challenges in emotion recognition
- Proposed approaches to emotion recognition
- Conclusions

Chimp data

Recognizing Politeness and Frustration in Child-machine Interactions (Eurospeech 05)

- Child-machine interactions in a game setting.
- The task is to play "Where in the USA is Carmen Sandiego?", an interactive computer game using speech.
- The goal of the game was to identify and arrest a cartoon criminal.
- WoZ spoken dialog interactions from 160 boys and girls, 7 to 14 years



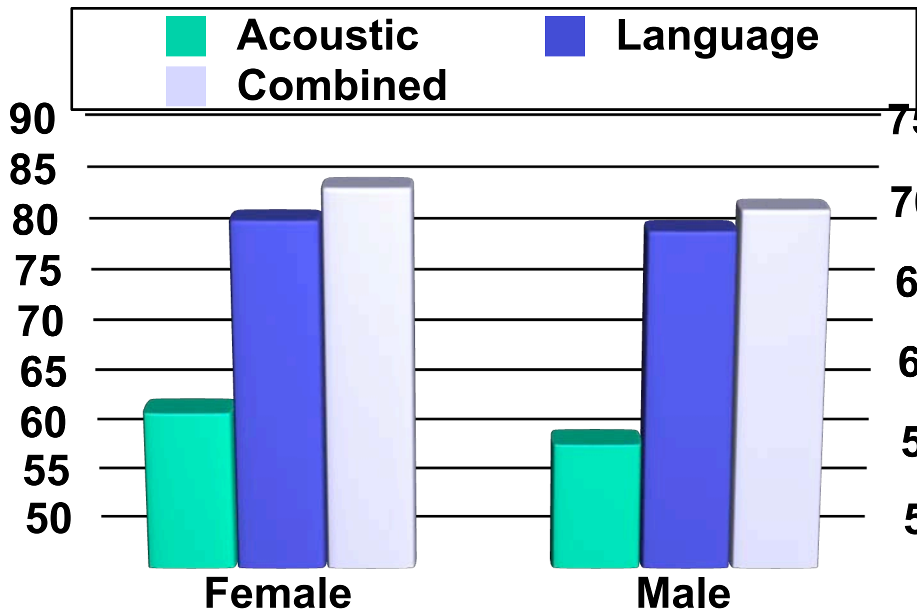
Frustration

Politeness

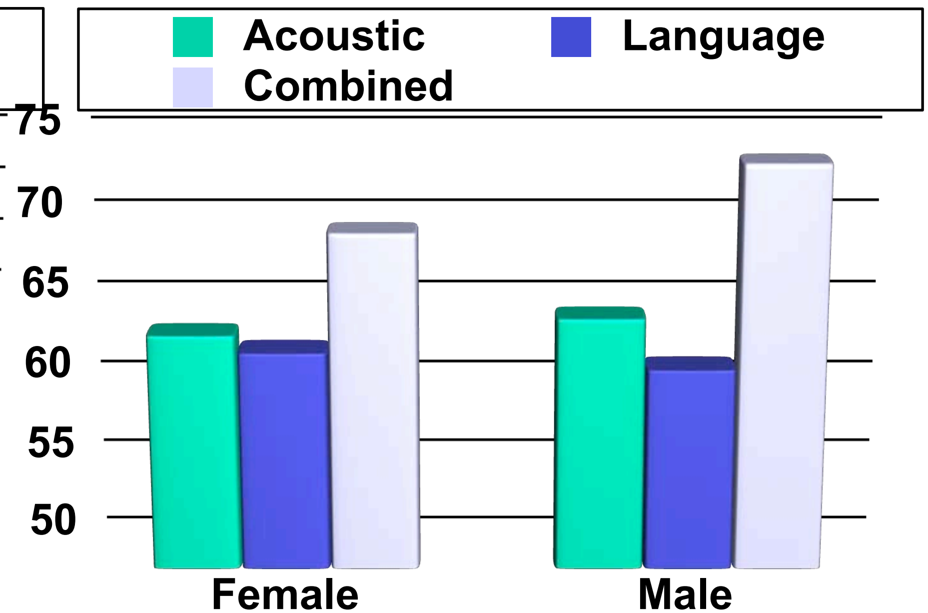
Chimp data

Results on "Carmen Sandiego" Game Task

Politeness Detection



Frustration Detection



Acoustic cues were more informative than language information for detecting frustration whereas the trend was opposite for politeness.

EMA data

Experimental Results using Segmental and Spectral Envelope Features (ICSLP'04)

Classification Method		Accuracy (%)
SVC with prosodic features		55.68
generic "emotional" HMM		64.77
Phoneme-class dependent HMM	every phoneme class	75.57
	vowel only	72.16
	glide only	54.86
	nasal only	47.43
	stop only	44.89
	fricative only	55.11
Combination of prosody and phoneme-class classifier		76.12

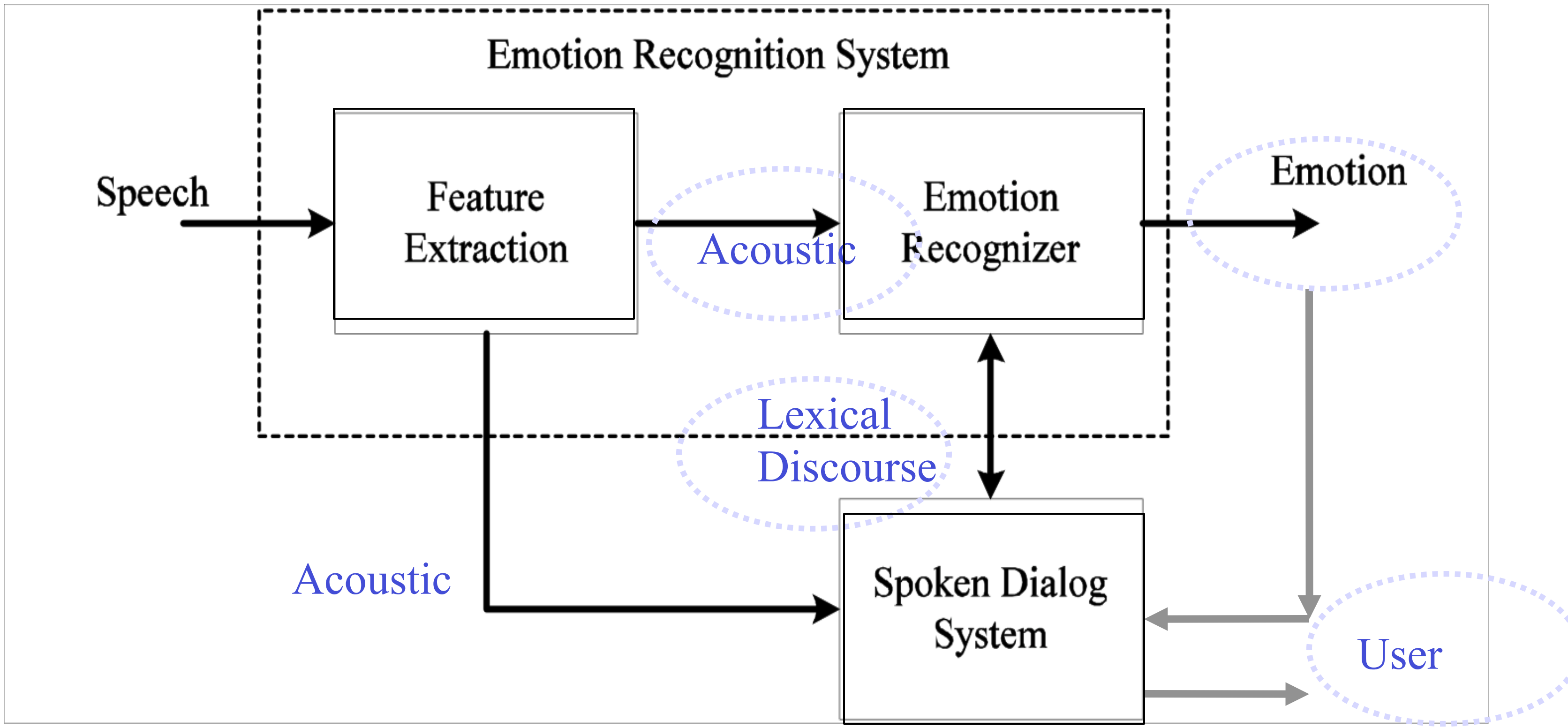
- Prosodic features
  - F0: mean, max, min
  - Mean and max of F0 slope
  - Speech rate
  - Classification: SVC
- Spectral features
  - MFCC
  - HMM classifier
- Assumption:
  - different emotional categories affect different phonemes in distinct ways
  - automatic emotion classification has to incorporate phoneme dependencies
- 5 different phoneme-classes
  - vowel, glide, nasal, stop, and fricative
  - Vowel productions, characterized by open vocal tracts and the less constrained articulation, not surprisingly show the greatest effects of emotion coloring
  - Non-continuant stop sounds seem to carry the least emotional information

Case Study: Call Center Data

- Emotion Classes
  - Two emotions were defined: Negative vs. Non-negative
  - The choice is practical: many applications need detect the users' frustration
- Acoustic information
  - Prosody (supra-segmental) features were used
    - 21 base features: Utterance-level statistics were calculated
    - Reduced set using forward selection (FS)
- Lexical Information
  - Features were calculated using emotional salience
- Discourse Information
  - Users' response (dialog acts) to the automated call system

CCD data

Recognizing Socio-affective state from Spoken Language



IEEE Trans. Speech&Audio, 2005

Lexical Information for emotion recognition

- People tend to use specific words in expressing their emotions
  - Speaker-dependent, but the usage of specific words in the expression of emotions is widely adopted in the given culture and society
- How to automatically extract and associate words to emotional categories
  - "Emotional Salience": information-theoretic quantity
- We used "true" transcription of the utterances for the classification
  - In real-world applications, we have the recognized word strings from a speech recognition system

Emotional Salience

- A measure of the amount of information that a specific word contains about an emotion category
  - A salient word w.r.t. an emotion category is one that appears more often in that category
- Defined as mutual information between emotional class and given specific word

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

Example of Emotionally Salient Words from Call Data

Word	Salience	Emotion
Wrong	0.72	Neg.
Computer	0.72	Neg.
Damn	0.72	Neg.
No	0.45	Neg.
Arrival	0.33	Non-Neg.
Delayed	0.21	Non-Neg.
Baggage	0.20	Non-Neg.

- A partial list of salient words
- "Emotion" represents maximally correlated emotion class given words
- The number of salient words are chosen such that they are greater than the preset threshold

Discourse Information

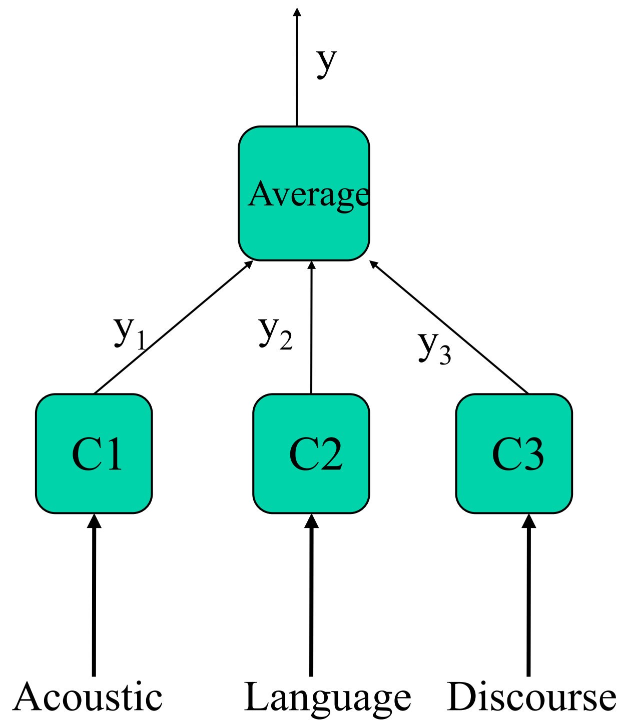
- Studies suggested that discourse information is useful for emotion recognition
  - Batliner, et al. ('96) : used topic repetition as 'language' information to combine with acoustic information
- Discourse information: Situational information in the dialog
- Several recent reports show results from several case studies
  - Five labels : rejection, repeat, rephrase, ask-start over, none of the above
  - These labels are used for discourse features for classification
  - Many utterances in negative emotion are in the rejection (26% for male, and 34% for female), whereas only 2% of the non-negative emotion utterances

Discourse Information in the Call Data

Tag	Male		Female		Total	
	Neg	Non-Neg	Neg	Non-Neg	Neg	Non-Neg
Reject	37	7	72	10	109	17
Repeat	4	35	23	38	27	73
Rephrase	15	34	10	39	25	73
Ask/ startover	29	33	33	44	62	77
Non of the above	57	350	71	448	128	798
Total	142	454	209	579	351	1038

Combining information sources

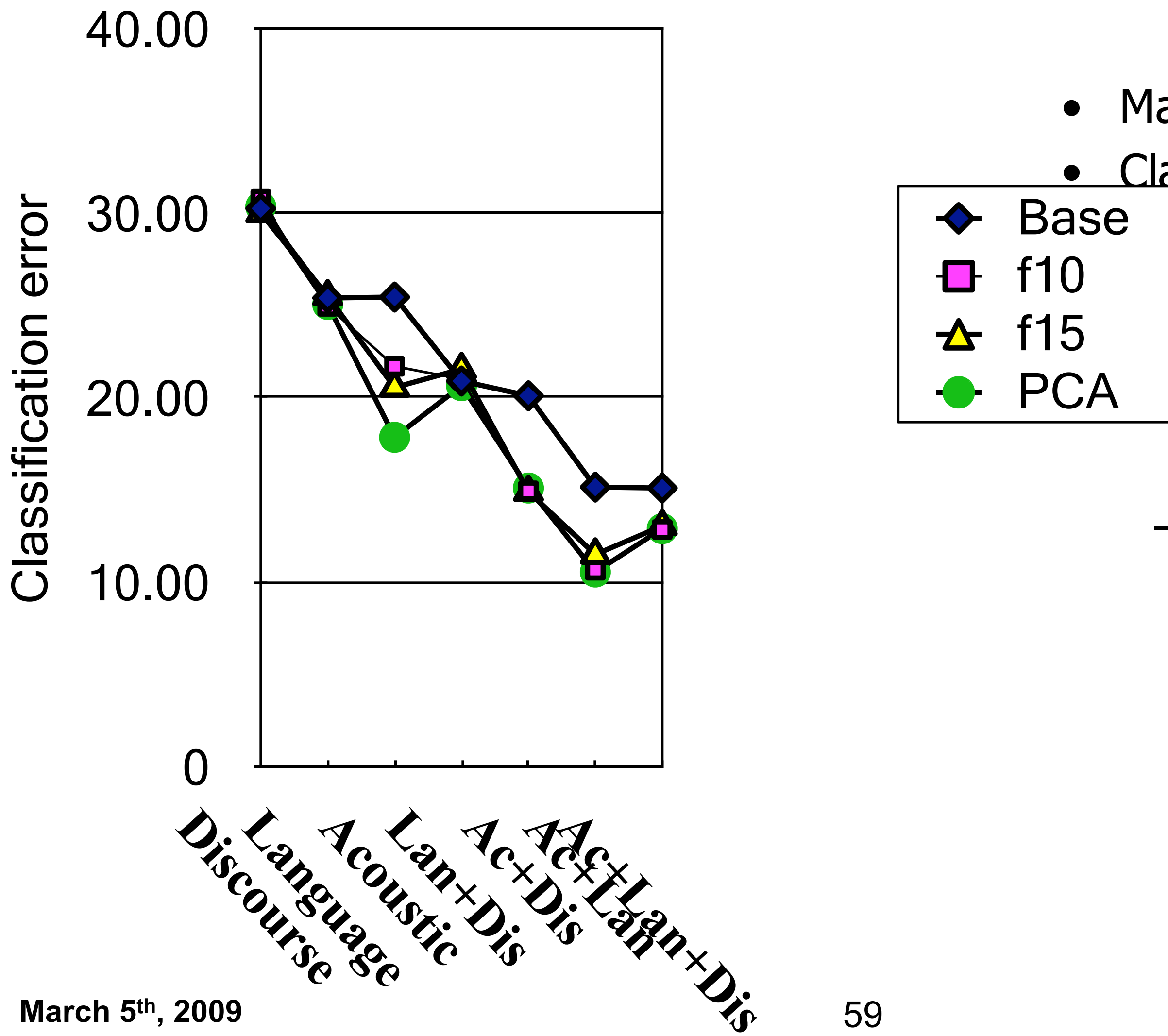
- Decision level combination
- Other scheme
  - Feature level combination:
    - Used by other authors
    - Problem: curse of dimensionality and dominance by acoustic information due to its large dimensionality
- In this study, average of the outputs from each source of information was taken.
  - Probabilistically, each output from the corresponding classifier is posterior probability
  - Averaging methods are less error-sensitive to the estimation of posterior probability



$$y = \frac{1}{N} \sum_{n=1}^{N} y_n(x)$$

Results

(IEEE Trans. Speech & Audio Proc. 2005)



- Male Call data
- Classification method:
  - Linear discriminant classifier for each information
- oustic features:
  - Prosody-related acoustic features
  - 4 feature sets
    - 21 full feature set
    - 10 and 15 best feature set
    - PCA feature set

Discussion

- Best performance when the information of 'acoustic' was combined with 'language' information
  - 'Discourse' information does not seem to provide significant improvement in conjunction with 'acoustic' and 'lexical' information
  - The reason may be due to the high correlation between 'lexical' and 'discourse' information
  - Q-statistic: A pair-wise measure of similarity between classifiers, and defined as

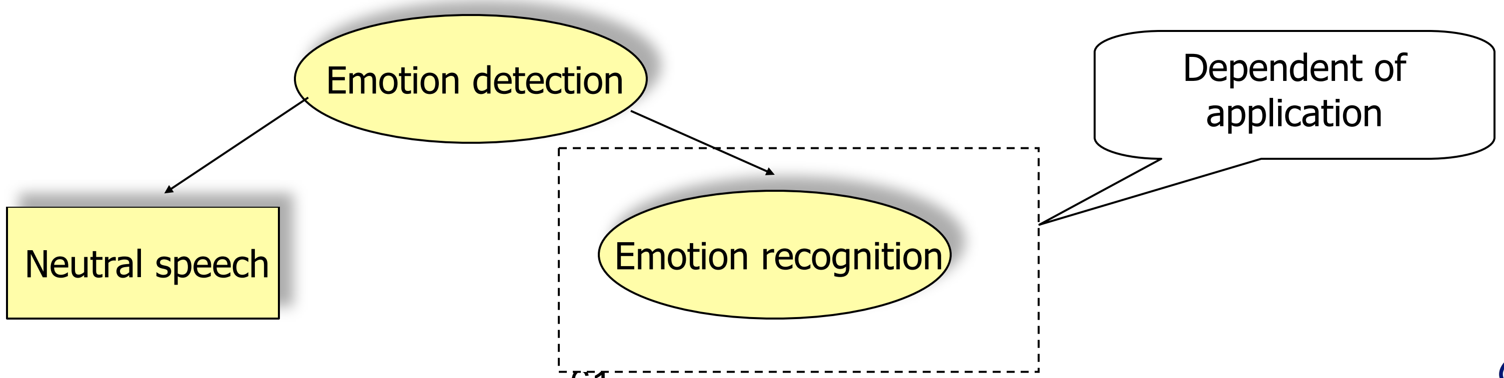
$$Q_{ij} = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}}$$

(where subscript 1: correct classification, 0: incorrect)

	Male	Female
$Q_{a,l}$	0.44	0.03
$Q_{a,d}$	0.28	0.18
$Q_{l,d}$	0.93	0.92

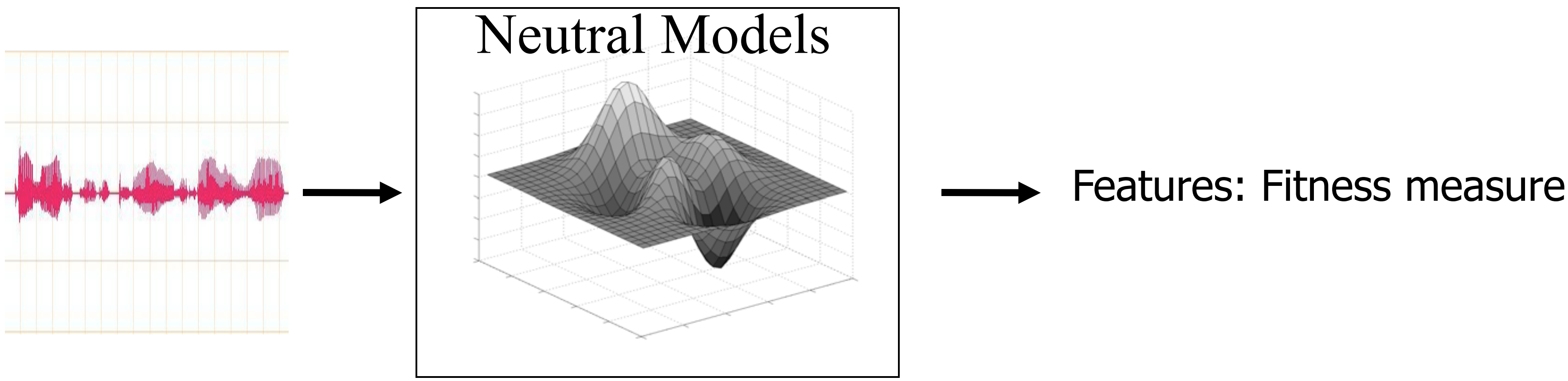
Neutral model approach

- Simplification: Neutral versus emotional speech
  - Emotional speech **detection**
  - Independent of emotional descriptor
  - Independent of applications
- It can be used as a first step in a more sophisticated multi-class emotion recognition system
  - Second level classification to achieve a finer emotional description



Proposed approach (1/2)

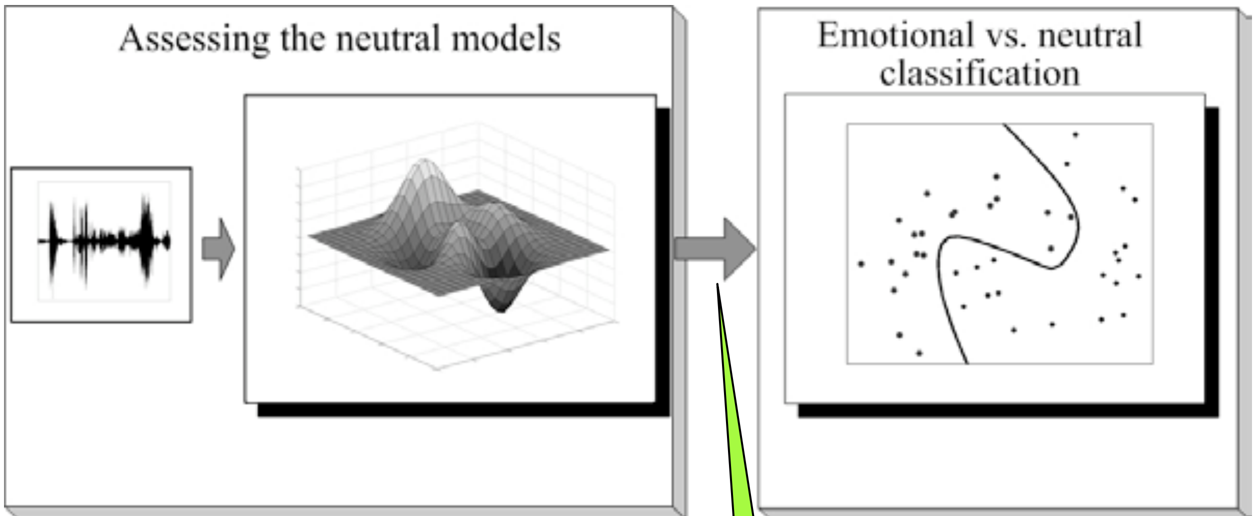
- Discriminate between emotional and neutral speech
- Acoustic reference models are used for emotion evaluation



- Emotional speech differs from neutral speech
- Many emotionally-neutral databases
  - Robust models

Proposed method (2/2)

- Two-step method:



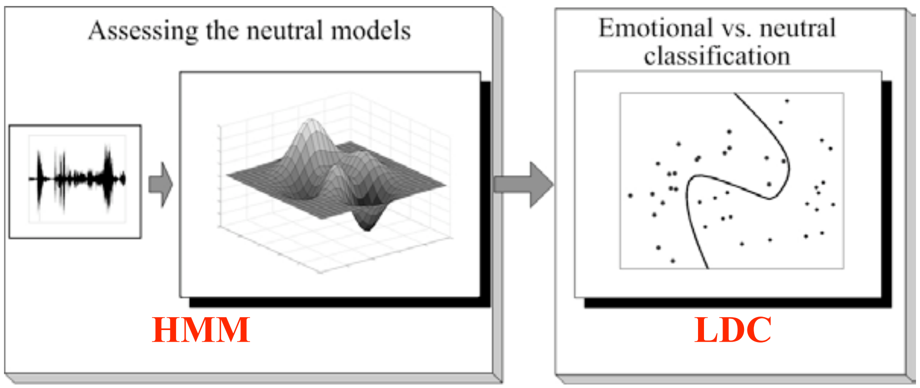
- Independent of
  - ✓Emotional classes
  - ✓Speaker
  - ✓Emotional databases

- Dependent on
  - ✓Emotional classes
  - ✓Speaker
  - ✓Emotional databases

Fitness measure

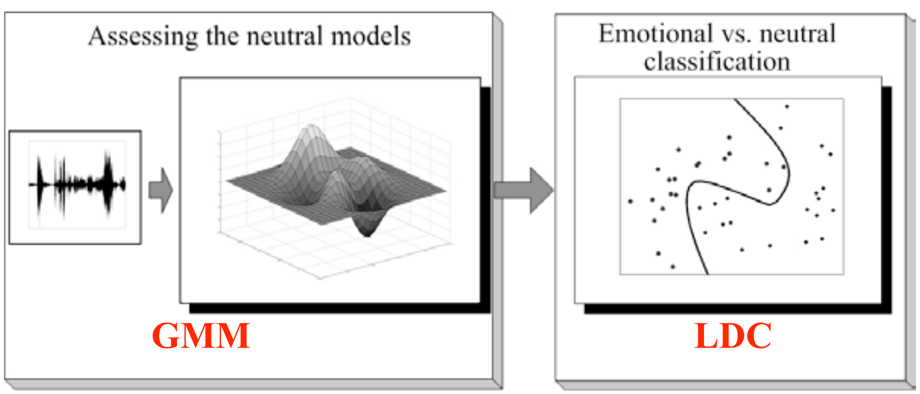
Implementation

MFB features (spectral features)



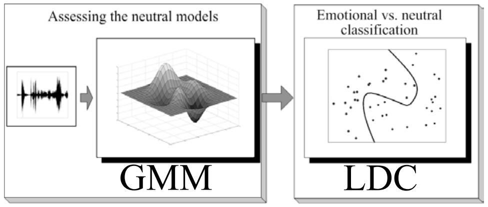
- Conventional HMMs are used to trained broad phonetic classes
- Fitness measurement: Normalized likelihood score
- Reference corpus: TIMIT
  - 460 speakers, 6300 sentences

F0 features (prosodic features)



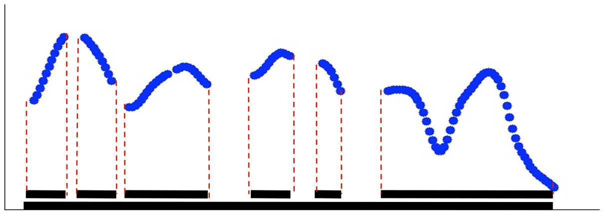
- Selection of emotional salient statistic from F0
- GMM for each selected feature
- Reference corpus: WSJ1
  - 50 speakers, 8104 sentences

F0 features



- Several databases:
  - Neutral reference (WSJ1) [Paul, 1992]
  - EMA (680 sentences, 3 speakers, neu, sad, hap, ang) [Lee,2005]
  - EPSAT (4738 sentences, 8 speakers, neu, sad, hap, bor, dis, fea, pan, cold ang, hot ang, des, ela, int, sha, pri) [Liberman, 2002]
  - GES (535 sentences, 10 speakers, neu, sad, hap, ang, bor, dis, fea) [Burkhardt, 2005]
- All databases together (reduce database dependency)
- Select equal number of samples for each emotional class (baseline 0.5)
- Classification is done 400 times

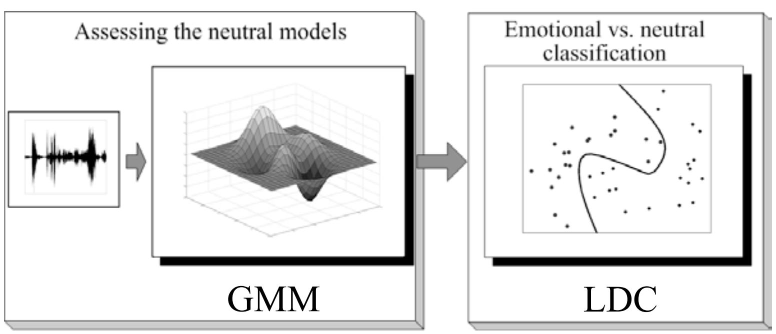
- Which features to use?



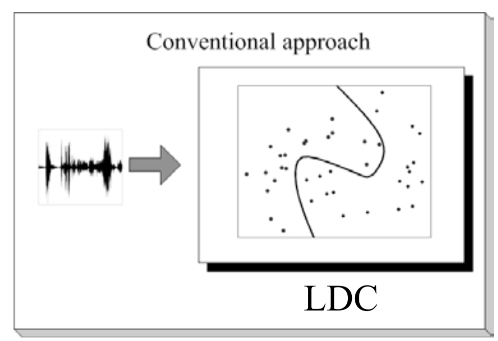
Results (1/3)

- Selected features from F0:
  - Sdiqr, Smedian, SQ75, SQ25, Sdmedian, SVmeanRange, SVmaxCurv

Neutral model approach (77.3%)



Conventional approach (74.7%)



Results (2/3)

- Mismatch between testing and training condition

Neutral model approach

Databases		Neutral model				
Training	Testing	Acc	Pre	Rec	F	dAcc
English (EPSAT,EMA)	German (GES)	0.802	0.778	0.818	0.798	4.1%
German (GES)	English (EPSAT,EMA)	0.751	0.732	0.762	0.746	4.6%
English (EPSAT,EMA)	Spanish (SES)	0.782	0.739	0.809	0.772	17.9%
German (GES)	Spanish (SES)	0.792	0.708	0.851	0.773	10.6%
English, German (EPSAT,EMA,GES)	Spanish (SES)	0.794	0.729	0.838	0.780	14.5%

Conventional approach

Databases		LDC Classifier				
Training	Testing	Acc	Pre	Rec	F	dAcc
English (EPSAT,EMA)	German (GES)	0.761	0.620	0.864	0.722	4.1%
German (GES)	English (EPSAT,EMA)	0.705	0.509	0.837	0.633	4.6%
English (EPSAT,EMA)	Spanish (SES)	0.604	0.412	0.668	0.510	17.9%
German (GES)	Spanish (SES)	0.686	0.445	0.857	0.586	10.6%
English, German (EPSAT,EMA,GES)	Spanish (SES)	0.649	0.420	0.775	0.545	14.5%

Results (3/3)

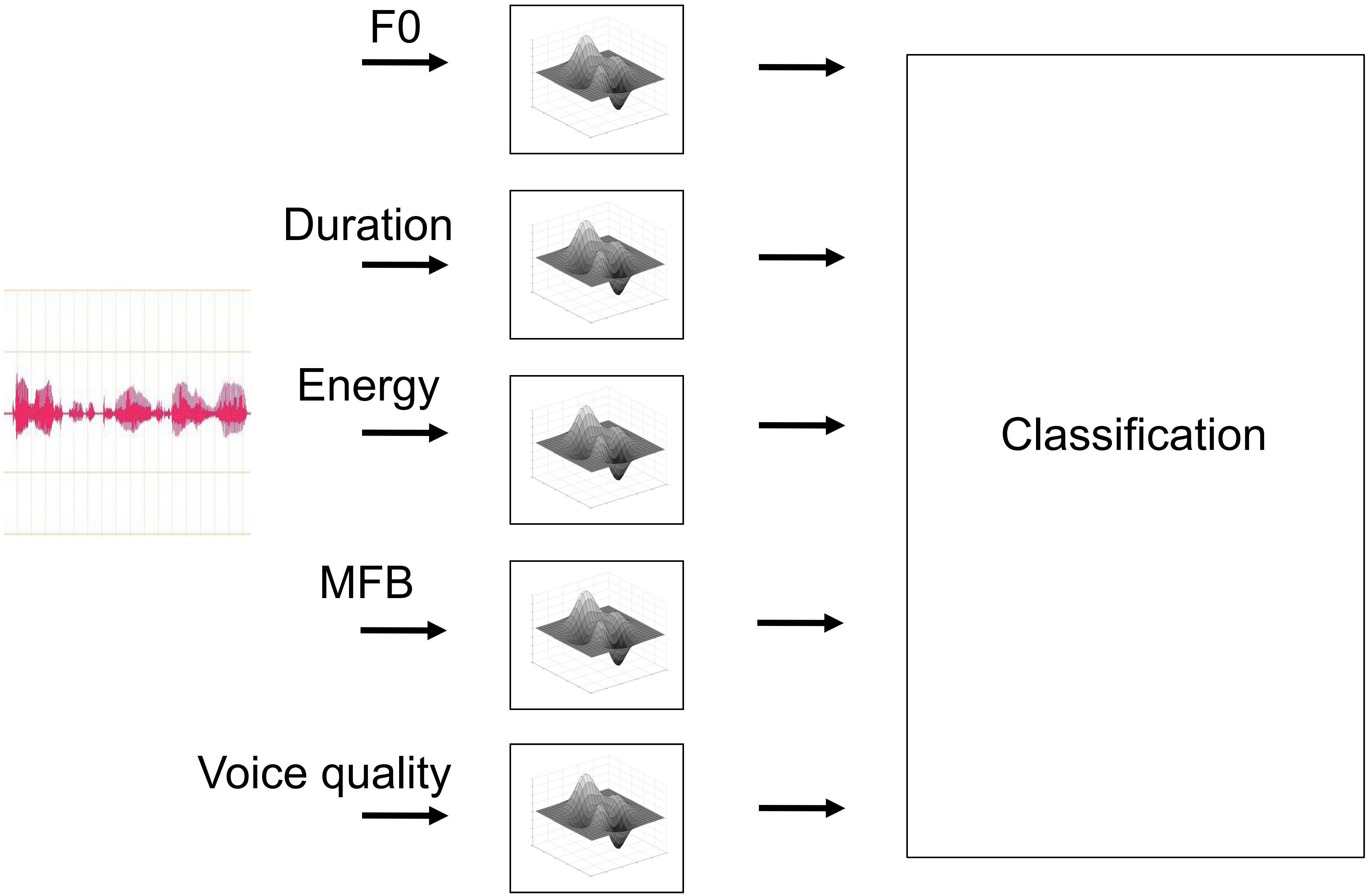
Without normalization

	Acc	Pre	Rec
EMA	0.7318	0.6968	0.5928
GES	0.7187	0.7939	0.6769
EPSAT	0.6555	0.6527	0.7146
Total	0.6787	0.6754	0.6896

Speaker dependent normalization

	Acc	Pre	Rec
EMA	0.8656	0.9227	0.7273
GES	0.8103	0.8671	0.7801
EPSAT	0.7416	0.7587	0.7313
Total	0.7749	0.7959	0.7376

Next directions



Outline

- Overview
- Challenges in emotion recognition
- Proposed approaches to emotion recognition
- Conclusions

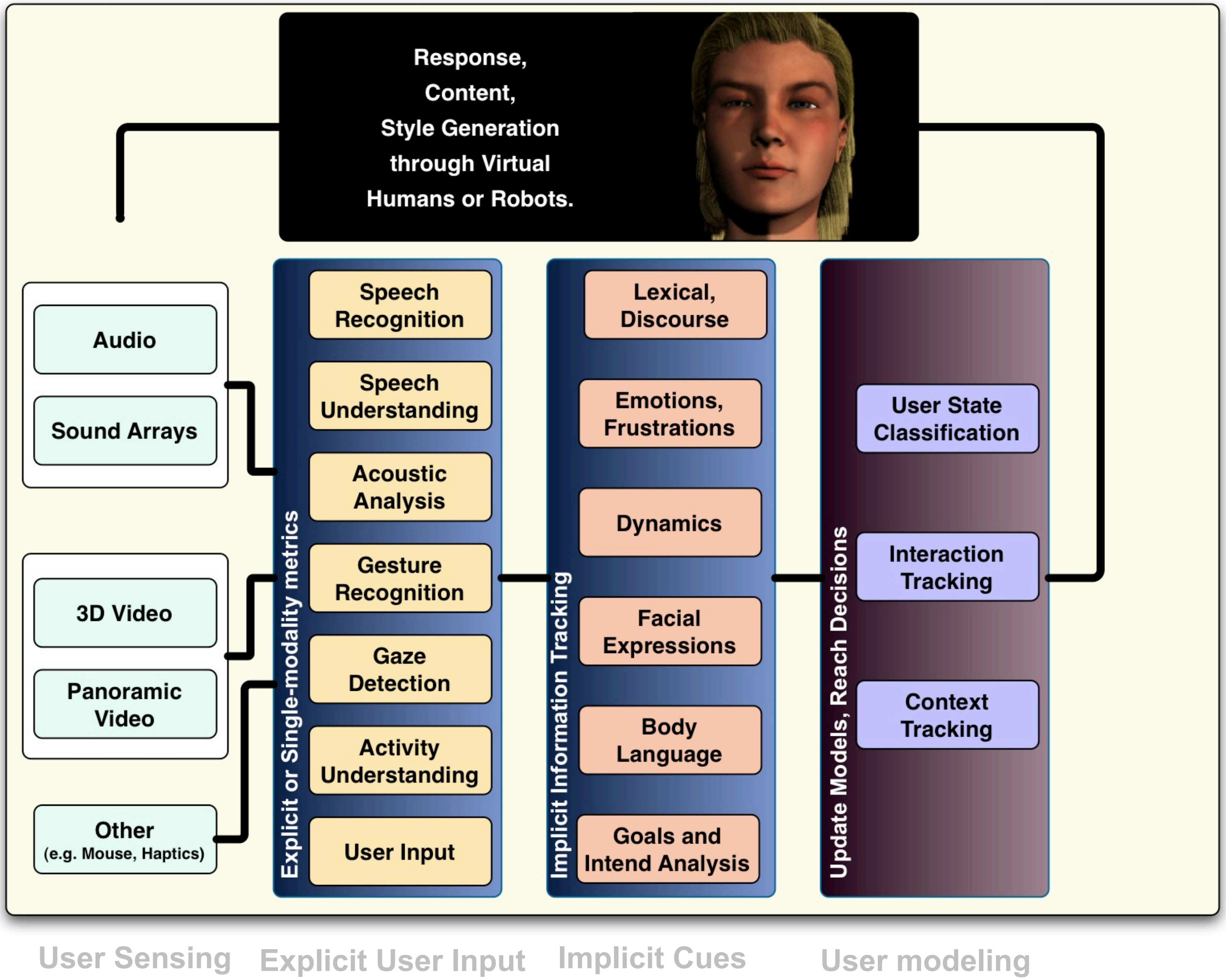
Conclusions

- Humans use multiple cues for emotion display/detection
  - From spoken language: 'lexical' and 'discourse' information with 'acoustic information' in the detection of emotions of other people
  - Gestures: various parts of the face, head and hand movements, body posture
- Contributions:
  - A comprehensive study of prosodic and segmental acoustic features
  - The use of fuzzy inference for emotion recognition
  - Information-theoretic concept of 'emotional salience' was adopted in obtaining 'lexical' information
  - Combination of information sources can improve the performance
  - Comparison of different representations
  - Real applications

Much remains to be done..

- More detailed incorporation of dynamic cues, including rate, boundary information
- Explore human emotional perception
  - Different combination of modalities may create different emotion percept
- Idiosyncratic influence in expressive human communication
  - How speaker-dependent are the results presented here
- Effect of "others" (listeners) on the expressive communication
  - Dyad and small group interaction
- Multimodal integration: visual gestures, physiological cues,..

A Multimodal Interaction Framework



Thank you…

Work Supported by: ONR, US Army, DARPA, NSF and NIH