

# CS 224S LING 281 Speech Recognition and Synthesis

Lecture 17: Emotion Detection and Synthesis  
Dan Jurafsky

Slides borrowed from Li, Chikberg, Jackson-Liacombe

## In the last 20 years

- A huge body of research on cognition and emotion
- Just one quick pointer: Ekman: basic emotions:



## Why Emotion Detection from Speech?

- Detecting frustration of callers to a help line
- Detecting stress in drivers or pilots
- Detecting "interest", "certainty", "confusion" in on-line tutors
  - Pacing/Positive feedback
- Synthesizing emotion for text-to-speech
  - On-line literacy tutors in the children's storybook domain
- Lie detection

## Some systems

- Detecting acted emotions, holding words constant
  - Hirschberg EPSaT corpus
- Detecting frustration of callers to appointment schedulers or call centers
  - Ang et al 2002
- Detecting "interest", "certainty", "confusion" in on-line tutors
  - Kate Forbes and Diane Litman. In press. [Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogues with both Human and Computer Tutors](#) (with Diane Litman), Speech Communication.
  - Pon-Barry
- Synthesizing emotion for text-to-speech
  - Eide et al

## Data and tasks for Emotion Detection

- Scripted speech
  - Acted emotions, often using 6 emotions
  - Controls for words, focus on acoustic/prosodic differences
  - Features:
    - F0/pitch
    - Energy
    - speaking rate
- Spontaneous speech
  - More natural, harder to control
  - Dialogue
  - Kinds of emotion focused on:
    - frustration,
    - annoyance,
    - certainty/uncertainty
    - "activation/hot spots"

## Example 1: Emotional Prosody Speech and Transcripts Corpus (EPSaT)

- Collected by Julia Hirschberg, Jennifer Venditti at Columbia University
- 8 actors read short dates and numbers in 15 emotional styles

Slide from Jackson-Liacombe

## EPSaT Examples


---

- happy 🗣️
- sad 🗣️
- angry 🗣️
- confident 🗣️
- frustrated
- friendly
- interested

anxious

bored

encouraging



Slide from Jackson Liscombe

## Detecting EPSaT Emotions

---

- Liscombe et al 2003

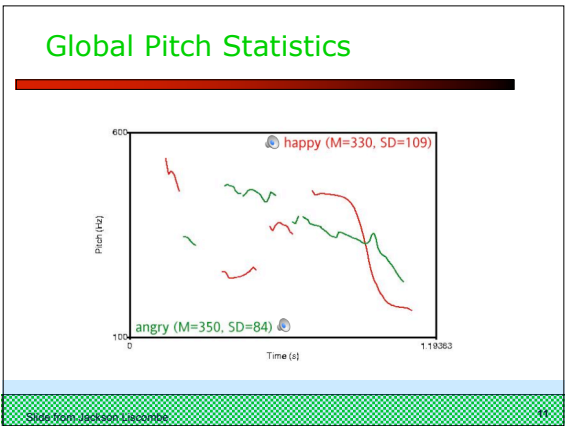
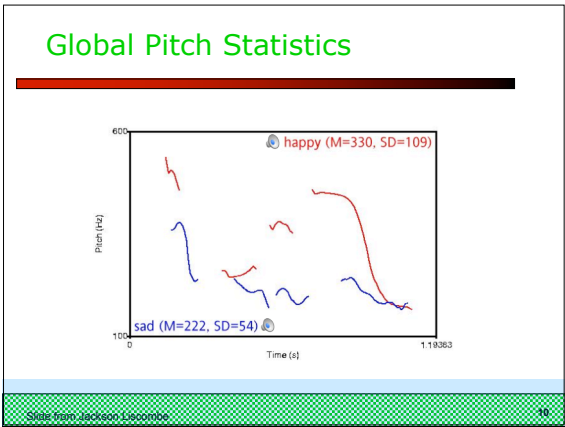
Slide from Jackson Liscombe

## Liscombe et al. Features

---

- Automatic Acoustic-prosodic
  - [Davitz, 1964] [Huttar, 1968]
  - Global characterization
    - pitch
    - loudness
    - speaking rate

Slide from Jackson Liscombe



## Liscombe et al. Features

---

- Automatic Acoustic-prosodic
  - [Davitz, 1964] [Huttar, 1968]
- ToBI Contours
  - [Mozziconacci & Hermes, 1999]
- Spectral Tilt
  - [Banse & Scherer, 1996] [Ang et al., 2002]

Slide from Jackson Liscombe

## Liscombe et al. Experiment

- RIPPER 90/10 split
- Binary Classification for Each Emotion
- Results
  - 62% average baseline
  - 75% average accuracy
  - Acoustic-prosodic features for activation
  - /H-L%/ for negative; /L-L%/ for positive
  - Spectral tilt for valence?

Slide from Jackson Liscombe

13

## EPSaT Discussion

1. How is emotion communicated through speech?
  - Confirmed usefulness of acoustic-prosodic features
  - Novel findings: pitch contour and spectral tilt

Slide from Jackson Liscombe

14

## Example 2 - Ang 2002

- Ang Shriberg Stolcke 2002 "Prosody-based automatic detection of annoyance and frustration in human-computer dialog"
- Prosody-Based detection of annoyance/ frustration in human computer dialog
- DARPA Communicator Project Travel Planning Data
  - NIST June 2000 collection: 392 dialogs, 7515 utts
  - CMU 1/2001-8/2001 data: 205 dialogs, 5619 utts
  - CU 11/1999-6/2001 data: 240 dialogs, 8765 utts
- Considers contributions of prosody, language model, and speaking style
- Questions
  - How frequent is annoyance and frustration in Communicator dialogs?
  - How reliably can humans label it?
  - How well can machines detect it?
  - What prosodic or other features are useful?

Slide from Shriberg, Ang, Stolcke

15

## Data Annotation

- 5 undergrads with different backgrounds (emotion should be judged by 'average Joe').
- Labeling jointly funded by SRI and ICSI.
- Each dialog labeled by 2+ people independently in 1st pass (July-Sept 2001), after calibration.
- 2nd "Consensus" pass for all disagreements, by two of the same labelers (Oct-Nov 2001).
- Used customized Rochester Dialog Annotation Tool (DAT), produces SGML output.

Slide from Shriberg, Ang, Stolcke

16

## Data Labeling

- **Emotion:** neutral, annoyed, frustrated, tired/disappointed, amused/surprised, no-speech/NA
- **Speaking style:** hyperarticulation, perceived pausing between words or syllables, raised voice
- **Repeats and corrections:** repeat/rephrase, repeat/rephrase with correction, correction only
- **Miscellaneous useful events:** self-talk, noise, non-native speaker, speaker switches, etc.

Slide from Shriberg, Ang, Stolcke

17

## Emotion Samples

- **Neutral**
  - July 30  1
  - Yes  2
- **Disappointed/tired**
  - No  6
- **Amused/surprised**
  - No  7
- **Annoyed**
  - Yes  3
  - Late morning (HYP)  8
- **Frustrated**
  - Yes  4
  - No  5
  - No, I am ... (HYP)  5
  - There is no Manila...  9
  -  10

Slide from Shriberg, Ang, Stolcke

18

## Emotion Class Distribution

	Count	%
Neutral	17994	.831
Annoyed	1794	.083
No-speech	1437	.066
Frustrated	176	.008
Amused	127	.006
Tired	125	.006
TOTAL	21653	

To get enough data, we grouped annoyed and frustrated, versus else (with speech)

Slide from Shriberg, Ang, Stolcke

19

## Prosodic Model

- Used CART-style decision trees as classifiers
- Downsampled to equal class priors (due to low rate of frustration, and to normalize across sites)
- Automatically extracted prosodic features based on recognizer word alignments
- Used automatic feature-subset selection to avoid problem of greedy tree algorithm
- Used 3/4 for train, 1/4th for test, no call overlap

Slide from Shriberg, Ang, Stolcke

20

## Prosodic Features

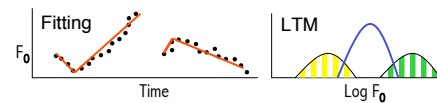
- Duration and speaking rate features**
  - duration of phones, vowels, syllables
  - normalized by phone/vowel means in training data
  - normalized by speaker (all utterances, first 5 only)
  - speaking rate (vowels/time)
- Pause features**
  - duration and count of utterance-internal pauses at various threshold durations
  - ratio of speech frames to total utt-internal frames

Slide from Shriberg, Ang, Stolcke

21

## Prosodic Features (cont.)

- Pitch features**
  - F0-fitting approach developed at SRI (Sónmez)
  - LTM model of F0 estimates speaker's F0 range



- Many features to capture pitch range, contour shape & size, slopes, locations of interest
- Normalized using LTM parameters by speaker, using all utts in a call, or only first 5 utts

Slide from Shriberg, Ang, Stolcke

22

## Features (cont.)

- Spectral tilt features**
  - average of 1st cepstral coefficient
  - average slope of linear fit to magnitude spectrum
  - difference in log energies btw high and low bands
  - extracted from longest normalized vowel region
- Other (nonprosodic) features**
  - position of utterance in dialog
  - whether utterance is a repeat or correction
  - to check correlations: hand-coded style features including hyperarticulation

Slide from Shriberg, Ang, Stolcke

23

## Language Model Features

- Train 3-gram LM on data from each class
- LM used word classes (AIRLINE, CITY, etc.) from SRI Communicator recognizer
- Given a test utterance, chose class that has highest LM likelihood (assumes equal priors)
- In prosodic decision tree, use sign of the likelihood difference as input feature
- Finer-grained LM scores cause overtraining

Slide from Shriberg, Ang, Stolcke

24

## Results: Human and Machine

	Accuracy (%) (chance = 50%)	Kappa (Acc-C)/(1-C)
Each Human with Other Human, overall	71.7	.38
Human with Human "Consensus" (biased)	84.2	.68
<b>Baseline</b> ▶ Prosodic Decision Tree with Consensus	75.6	.51
Tree with Consensus, no repeat/correction	72.9	.46
Tree with Consensus, repeat/correction only	68.7	.37
Language Model features only	63.8	.28

Slide from Shriberg, Ang, Stolcke

25

## Results (cont.)

- H-H labels agree 72%, **complex decision task**
  - inherent continuum
  - speaker differences
  - relative vs. absolute judgements?
- H labels agree 84% with "consensus" (biased)
- Tree model agrees 76% with consensus-- *better than original labelers with each other*
- Prosodic model makes use of a dialog state feature, but without it it's still better than H-H
- Language model features alone are not good predictors (dialog feature alone is better)

Slide from Shriberg, Ang, Stolcke

26

## Predictors of Annoyed/Frustrated

- **Prosodic: Pitch features:**
  - high maximum fitted F0 in longest normalized vowel
  - high speaker-norm. (1st 5 utts) ratio of F0 rises/falls
  - maximum F0 close to speaker's estimated F0 "topline"
  - minimum fitted F0 late in utterance (no "?" intonation)
- **Prosodic: Duration and speaking rate features**
  - long maximum phone-normalized phone duration
  - long max phone- & speaker- norm.(1st 5 utts) vowel
  - low syllable-rate (slower speech)
- **Other:**
  - utterance is repeat, rephrase, explicit correction
  - utterance is after 5-7th in dialog

Slide from Shriberg, Ang, Stolcke

27

## Effect of Class Definition

	Accuracy (%) (chance = 50%)	Entropy Reduction
Baseline prosody model Consensus labels A,F vs. N,else	75.6	21.6
Tokens on which labelers originally agreed A,F vs. N,else	78.3	26.4
All tokens Consensus labels F vs. A,N,else	82.7	37.0

For **less ambiguous** tokens, or **more extreme** tokens performance is significantly better than baseline

Slide from Shriberg, Ang, Stolcke

28

## Ang et al '02 Conclusions

- Emotion labeling is a complex decision task
- Cases that labelers independently agree on are classified with high accuracy
  - Extreme emotion (e.g. 'frustration') is classified even more accurately
- Classifiers rely heavily on prosodic features, particularly duration and stylized pitch
  - Speaker normalizations help
- Two nonprosodic features are important: utterance position and repeat/correction
  - Language model is an imperfect surrogate feature for the underlying important feature repeat/correction

Slide from Shriberg, Ang, Stolcke

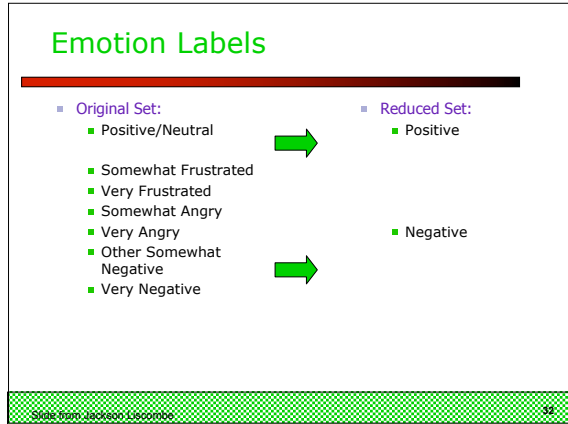
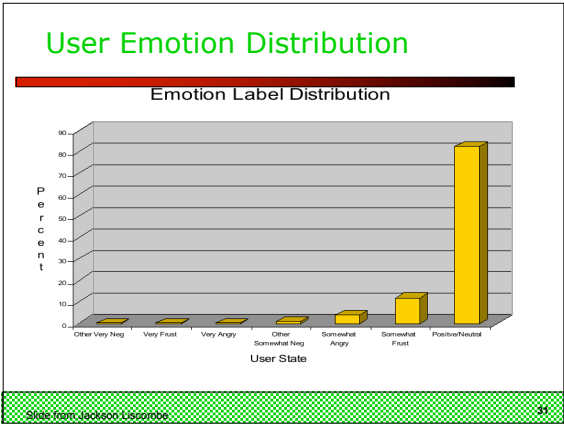
29

## Example 3: "How May I Help You<sup>SM</sup>" (HMIHY)

- Giuseppe Riccardi, Dilek Hakkani-Tür, AT&T Labs
- Liscombe, Riccardi, Hakkani-Tür (2004)
- Each turn in 20,000 turns (5690 dialogues) annotated for 7 emotions by one person
  - Positive/neutral, somewhat frustrated, very frustrated, somewhat angry, very angry, somewhat other negative, very other negative
  - Distribution was so skewed (73.1% labeled positive/neutral)
  - So classes were collapsed to **negative/nonnegative**
- Task is hard!
  - Subset of 627 turns labeled by 2 people: kappa .32 (full set) and .42 (reduced set)!

Slide from Jackson-Liscombe

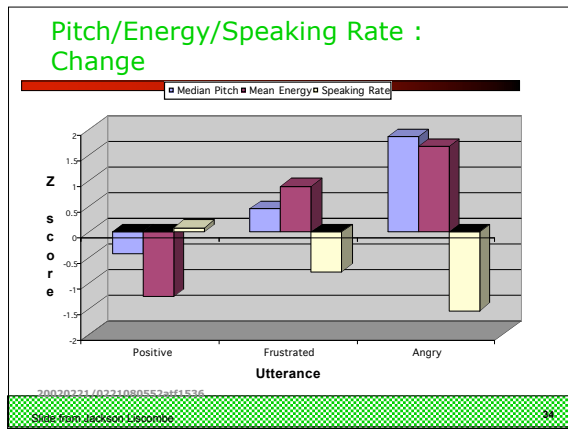
30



### Pitch/Energy/Speaking Rate : Change

Person	Transcription	Emotion	Play
Agent:	I'm sorry, I didn't understand that. How may I help you?		
User:	please an operator please [ .hst ] may I speak-	Positive/Neutral	
Agent:	I'm sorry we're having trouble. Can we try one more time? Just briefly tell me how I may help you.		
User:	may I speak to an assistant please	Somewhat Frustrated	
Agent:	Are you calling from your home phone?		
User:	yes	Somewhat Angry	

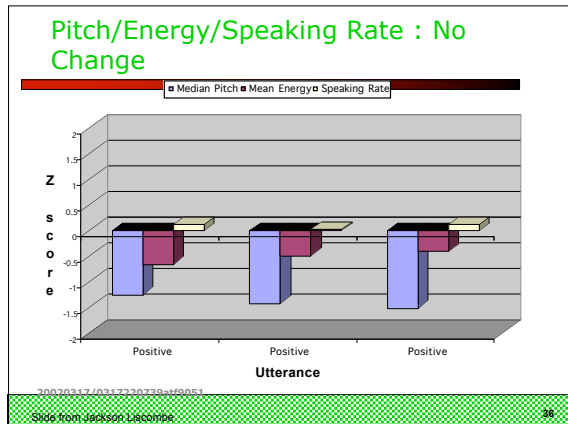
30030331 / 0311080552-m61536  
Slide from Jackson Licombe 33



### Pitch/Energy/Speaking Rate : No Change

Person	Transcription	Emotion	Play
Agent:	Hello, I am an AT&T automated assistant. You can speak naturally to me. How may I help you?		
User:	I need to find out about a number that I don't recognize	Positive/Neutral	
Agent:	Would you like to look up a number you don't recognize on you bill?		
User:	yes I would	Positive/Neutral	
Agent:	Are you calling from your home phone?		
User:	yes I am	Positive/Neutral	

30030317 / 031730730-m69051  
Slide from Jackson Licombe 35



## HMIHY Features

- Automatic Acoustic-prosodic
- Contextual [Cauldwell, 2000]
- Transcriptive [Schröder, 2003] [Brennan, 1995]
- Pragmatic [Ang et al., 2002] [Lee & Narayanan, 2005]

Slide from Jackson Liscombe

37

## Lexical Features

- Language Model (ngrams)
- Examples of words significantly correlated with negative user state ( $p < 0.001$ ):
  - 1<sup>st</sup> person pronouns: 'I', 'me'
  - requests for a human operator: 'person', 'talk', 'speak', 'human', 'machine'
  - billing-related words: 'dollars', 'cents'
  - curse words: ...

Slide from Jackson Liscombe

38

## Prosodic Features

- Pitch (F0)**
  - Overall minimum
  - overall maximum
  - overall median
  - overall standard deviation
  - mean absolute slope
  - slope of final vowel
  - longest vowel mean
- Energy**
  - overall minimum
  - overall maximum
  - overall mean
  - overall standard deviation
  - longest vowel mean
- Other**
  - local jitter over longest vowel
- Speaking Rate**
  - vowels per second
  - mean vowel length
  - ratio voiced frames to total frames
  - percent internal silence

Slide from Jackson Liscombe

39

## Contextual Features

- Lexical (2)**
  - edit distance with previous 2 turns
- Prosodic (34)**
  - 1<sup>st</sup> and 2<sup>nd</sup> order differentials for each feature
- Discourse (10)**
  - turn number
  - call type repetition with previous 2 turns
  - dialog act repetition with previous 2 turns
- Other (2)**
  - user state of previous 2 turns

Slide from Jackson Liscombe

40

## HMIHY Experiment

- Classes: *Negative vs. Non-negative*
  - Training size = 15,013 turns
  - Testing size = 5,000 turns
- Most frequent user state (*positive*) accounts for 73.1% of testing data
- Learning Algorithm Used:
  - BoosTexter
    - (boosting w/ weak learners)
  - continuous/discrete features
  - 2000 iterations
- Results:

Features	Accuracy
Baseline	73%
Acoustic-prosodic	75%
+ transcriptive	76%
+ pragmatic	77%
+ contextual	79%

Slide from Jackson Liscombe

41

## HMIHY Discussion

- How is emotion communicated through speech?
  - Novel features improve performance
    - transcription
    - pragmatics
    - context

Slide from Jackson Liscombe

42

## Intelligent Tutoring Spoken Dialogue System

- (ITSpoke)
- Diane Litman, Katherine Forbes-Riley, Scott Silliman, Mihai Rotaru, University of Pittsburgh, Julia Hirschberg, Jennifer Venditti, Columbia University

Slide from Jackson Lieberman

43

The screenshot shows a web browser window with the URL 'http://www.cs.pitt.edu/it spoke'. The page title is 'ITSpoke'. The main content area displays a physics problem: '58. Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys?'. Below the problem is a text box containing the solution: 'The keys will rise above the man's face because the same gravitational force is being applied to both, yet the man's mass is greater than the mass of the keys so he will fall faster than the keys.' There are three mouse cursor icons and the text '[pr01\_sess00\_prob58]' below the problem. A 'Submit' button is visible at the bottom right of the solution box.

Slide from Jackson Lieberman

44

## Task 1

- Negative
  - Confused, bored, frustrated, uncertain
- Positive
  - Confident, interested, encouraged
- Neutral

45

**PROBLEM (TYPED):** If a car is able to accelerate at  $2 \text{ m/s}^2$ , what acceleration can it attain if it is towing another car of equal mass?

**ESSAY (TYPED):** The maximum acceleration a car can reach when towing a car behind it of equal mass will be halved. Therefore, the maximum acceleration will be  $1 \text{ m/s}^2$ .

**DIALOGUE (SPOKEN):** ... 9.1 min. into session ...

**TUTOR<sub>1</sub>:** Uh let us talk of one car first.

**STUDENT<sub>1</sub>:** ok. (*EMOTION = NEUTRAL*)

**TUTOR<sub>2</sub>:** If there is a car, what is it that exerts force on the car such that it accelerates forward?

**STUDENT<sub>2</sub>:** The engine (*EMOTION = POSITIVE*)

**TUTOR<sub>3</sub>:** Uh well engine is part of the car, so how can it exert force on itself?

**STUDENT<sub>3</sub>:** um... (*EMOTION = NEGATIVE*)

46

### Acoustic-Prosodic Features

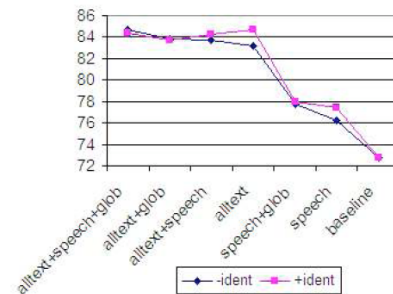
- 4 normalized fundamental frequency (f0) features: maximum, minimum, mean, standard deviation
- 4 normalized energy (RMS) features: maximum, minimum, mean, standard deviation
- 4 normalized temporal features: total turn duration, duration of pause prior to turn, speaking rate, amount of silence in turn

### Non-Acoustic-Prosodic Features

- lexical items in turn
- 6 automatic features: turn begin time, turn end time, isTemporalBarge-in, isTemporalOverlap, #words in turn, #syllables in turn
- 6 manual features: #false starts in turn, isPriorTutorQuestion, isQuestion, isSemanticBarge-in, #canonical expressions in turn, isGrounding

**Identifier Features:** subject, subject gender, problem

47



48

## Liscombe et al: Uncertainty in ITSpoke

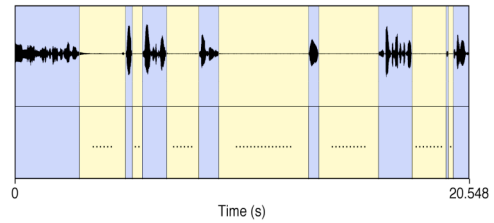
um <sigh> I don't even think I have an idea here ..... now .. mass isn't weight ..... mass is ..... the ..... space that an object takes up ..... is that mass?

[71-67-1:92-113]

Slide from Jackson Liscombe

49

## Breath Group Segmentation



[71-67-1:92-113]

Slide from Jackson Liscombe

50

## Liscombe et al: ITSpoke Features

- Automatic acoustic-prosodic
- Contextual
- Breath groups  
[Hirst & Christo, 1998]

Slide from Jackson Liscombe

51

## Liscombe et al: ITSpoke Experiment

- Human-Human Corpus
- AdaBoost(C4.5) 90/10 split in WEKA
- Classes: *Uncertain vs Certain vs Neutral*
- Results:

Features	Accuracy
Baseline	66%
Acoustic-prosodic	75%
+ contextual	76%
+ breath-groups	77%

Slide from Jackson Liscombe

52