

CS 224S/LING 281

Speech Recognition,

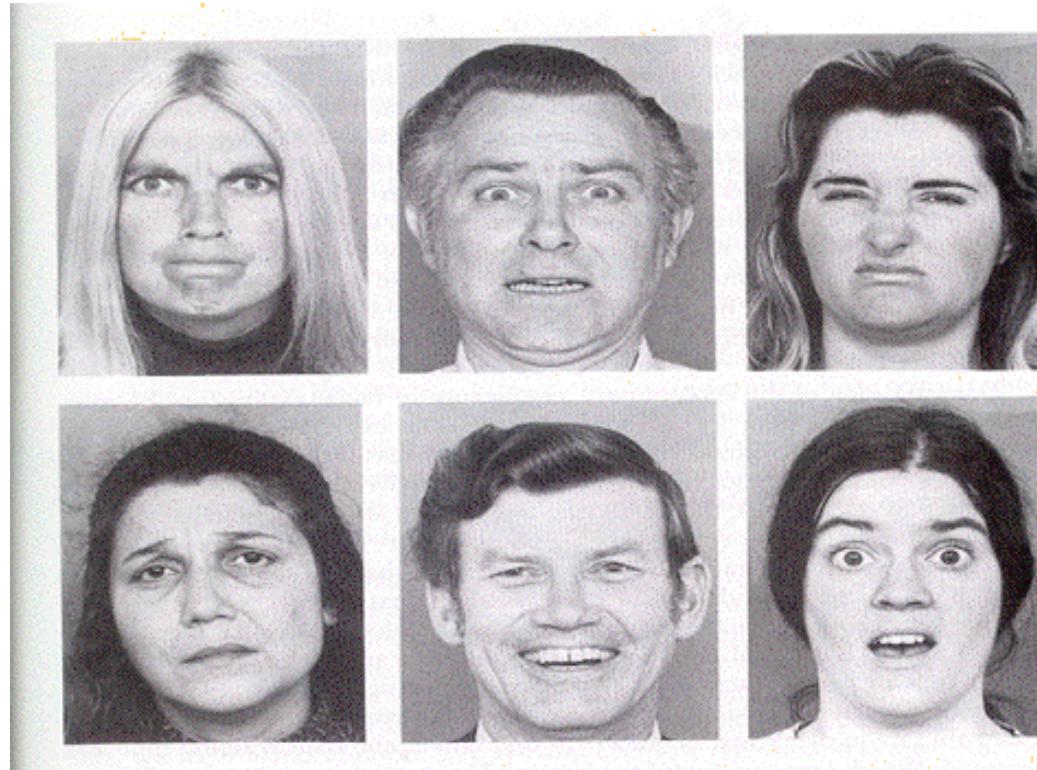
Synthesis, and Dialogue

Dan Jurafsky

Lecture 18: Emotion

In the last 20 years

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





Disgust

Anger

Sadness

Happiness



Fear

Surprise

Ekman's 6 basic emotions

Surprise, happiness, anger, fear, disgust, sadness



Why Emotion Detection from Speech?

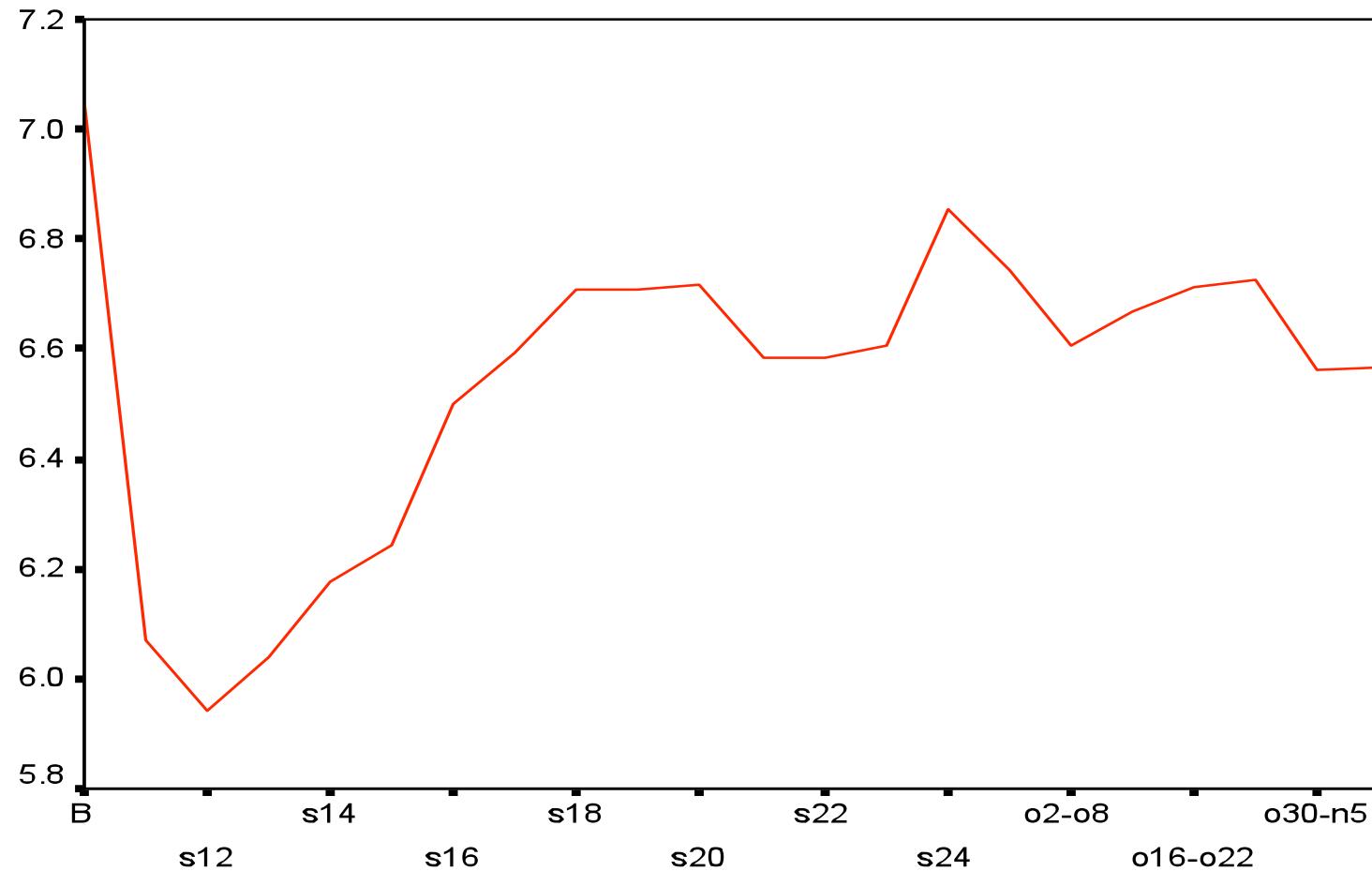
- Detecting frustration of callers to a help line
- Detecting stress in drivers or pilots
- Detecting “interest”, “certainty”, “confusion” in on-line tutors
 - ◆ Pacing/Positive feedback
- Synthesizing emotion for text-to-speech
 - ◆ On-line literacy tutors in the children’s storybook domain
- Lie detection

Some systems

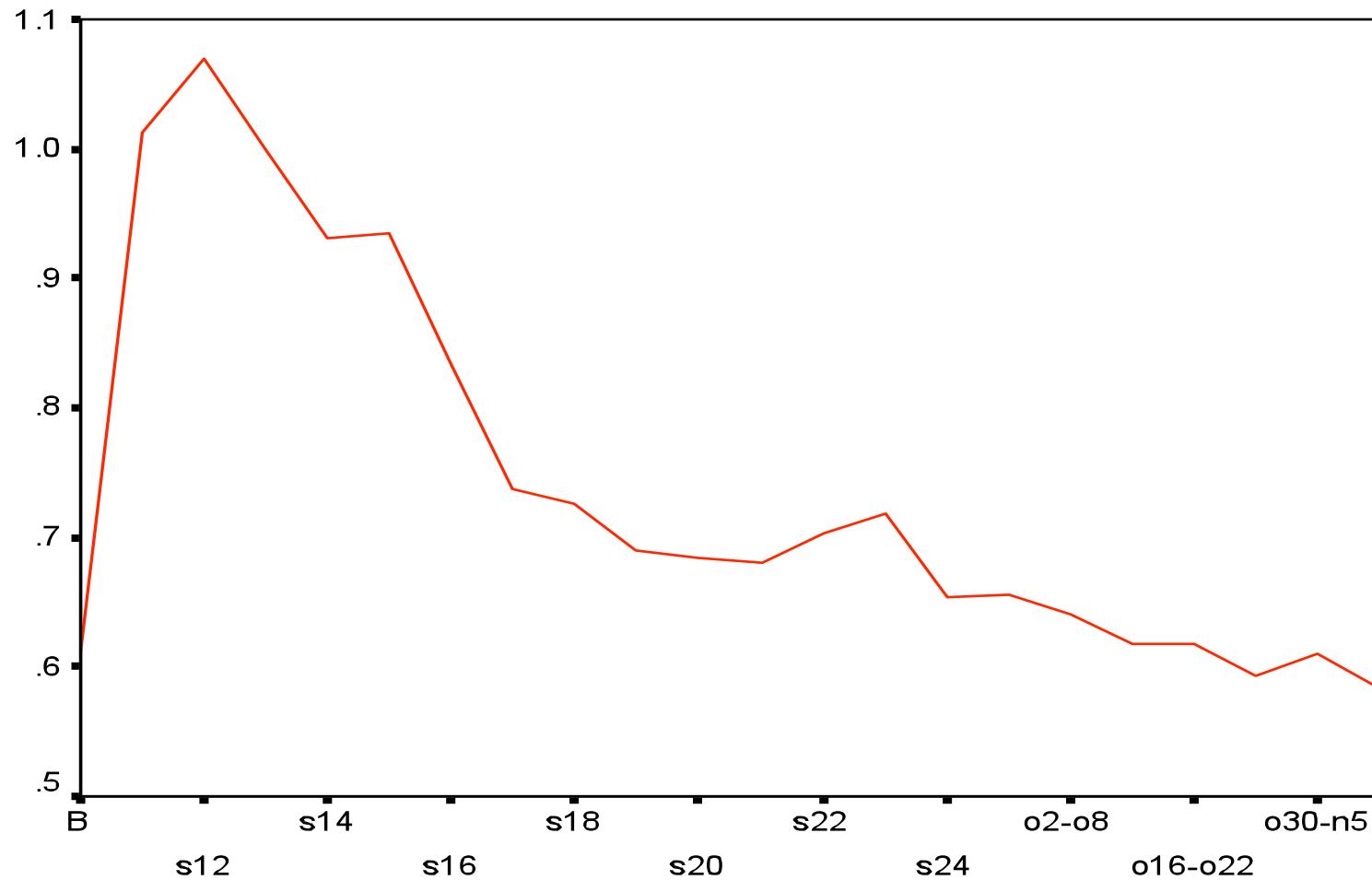
- Detecting acted emotions, holding words constant
 - ◆ Hirschberg EPoS corpus
- Detecting frustration of callers to appointment schedulers or call centers
 - ◆ Ang et al 2002
- Detecting “interest”, “certainty”, “confusion” in on-line tutors
 - ◆ Kate Forbes and Diane Litman. In press.
Recognizing Student 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

Text

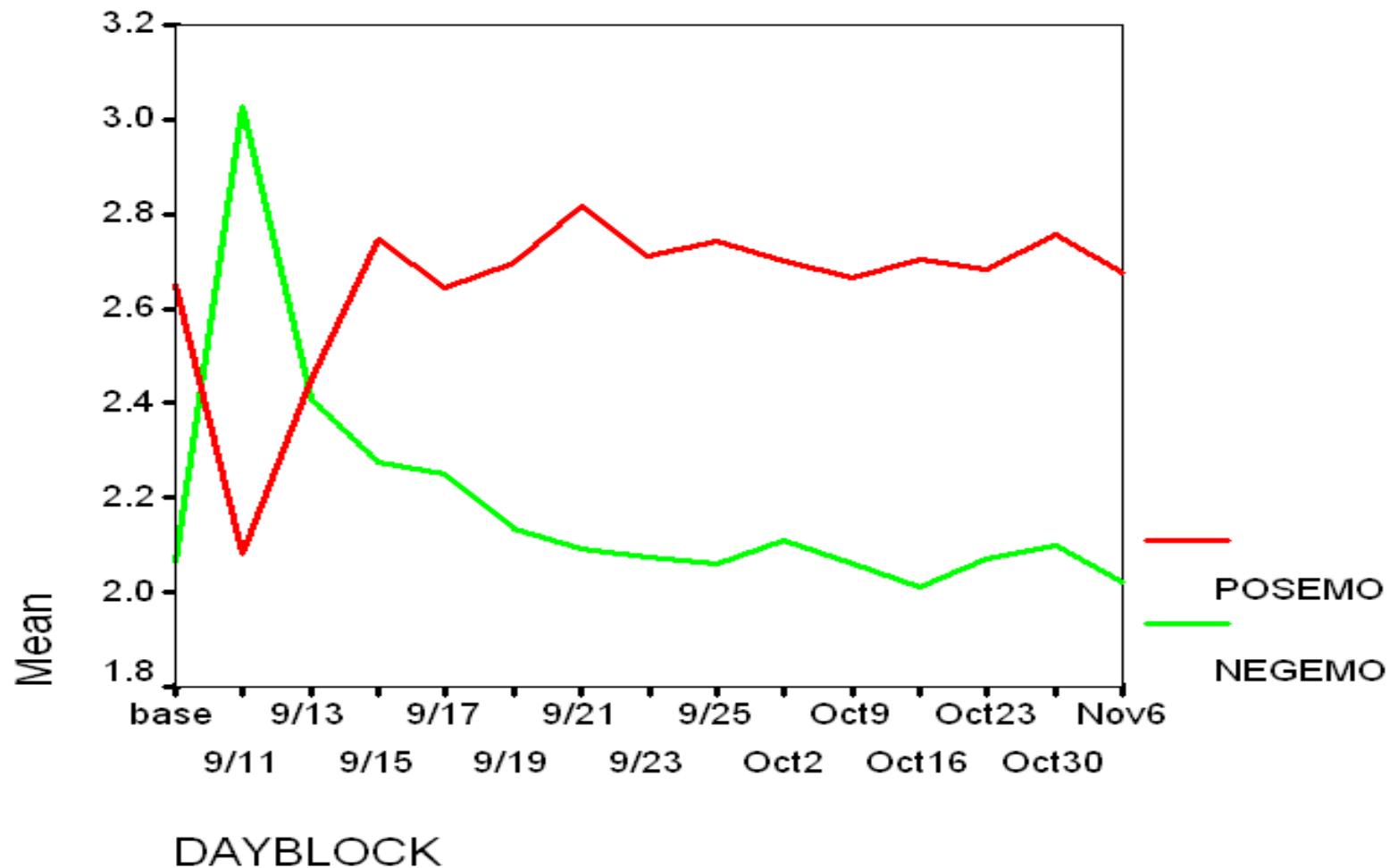
Pennebaker Livejournal.com Study: I, me, my



Pennebaker September 11 LiveJournal.com study: We, us, our



Pennebaker LiveJournal.com study: Positive and negative emotion words



Personality

- Mairesse & Walker '07 PERSONAGE (PERSONAlity GEnerator)
 - ◆ 'Big 5' personality trait model: extroversion, neuroticism, agreeableness, conscientiousness, openness to experience
 - ◆ Attempts to generate “extroverted” language based on traits associated with extroversion in psychology literature
 - ◆ Demo:
 - <http://mi.eng.cam.ac.uk/~farm2/personality/demo.html>: find your personality type

NLG modules	Introvert findings	Extravert findings	Parameter	Intro	Extra
Content selection and structure	Single topic Strict selection Problem talk, dissatisfaction	Many topics Think out loud*	VERBOSITY RESTATEMENTS REPETITIONS CONTENT POLARITY REPETITIONS POLARITY CLAIM POLARITY CONCESSIONS CONCESSIONS POLARITY POLARISATION POSITIVE CONTENT FIRST	low low low low low low avg low low low	high high low high high high avg high high high
Syntactic templates selection	Few self-references Elaborated constructions Many articles	Many self-references Simple constructions* Few articles	SELF-REFERENCES CLAIM COMPLEXITY	low high	high low
Aggregation Operations	Many words per sentence/clause Many unfilled pauses	Few words per sentence/clause Few unfilled pauses	RELATIVE CLAUSES WITH CUE WORD CONJUNCTION PERIOD ...	high high low high	low low high low
Pragmatic transformations	Many nouns, adjectives, prepositions (explicit) Many negations Many tentative words Formal Realism No politeness form Lower word count	Many verbs, adverbs, pronouns (implicit) Few negations Few tentative words Informal Exaggeration* Positive face redressment* Higher word count	SUBJECT IMPLICITTNESS NEGATION INSERTION DOWNTONER HEDGES: ·SORT OF, SOMEWHAT, QUITE, RATHER, ERR, I THINK THAT, IT SEEMS THAT, IT SEEMS TO ME THAT, I MEAN ·AROUND ·KIND OF, LIKE ACKNOWLEDGMENTS: ·YEAH ·RIGHT, OK, I SEE, WELL EMPHASIZER HEDGES: ·REALLY, BASICALLY, ACTUALLY, JUST HAVE, JUST IS, EXCLAMATION ·YOU KNOW TAG QUESTION INSERTION HEDGE VARIATION HEDGE REPETITION	low high high low avg low high low high low high low high low low low low low	high low high low avg high low high high high low
Lexical choice	Rich Few positive emotion words Many negative emotion words	Poor Many positive emotion words Few negative emotion words	LEXICON FREQUENCY <i>see polarity parameters</i> <i>see polarity parameters</i>	low	high

Table 2: Summary of language cues for extraversion, based on Dewaele and Furnham (1999); Furnham (1990); Mehl et al. (2006); Oberlander and Gill (2006); Pennebaker and King (1999), as well as PERSON-AGE's corresponding generation parameters. Asterisks indicate hypotheses, rather than results. For details on aggregation parameters, see Section 4.2.

Speech

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

EPSaT Examples

happy 

sad 

angry 

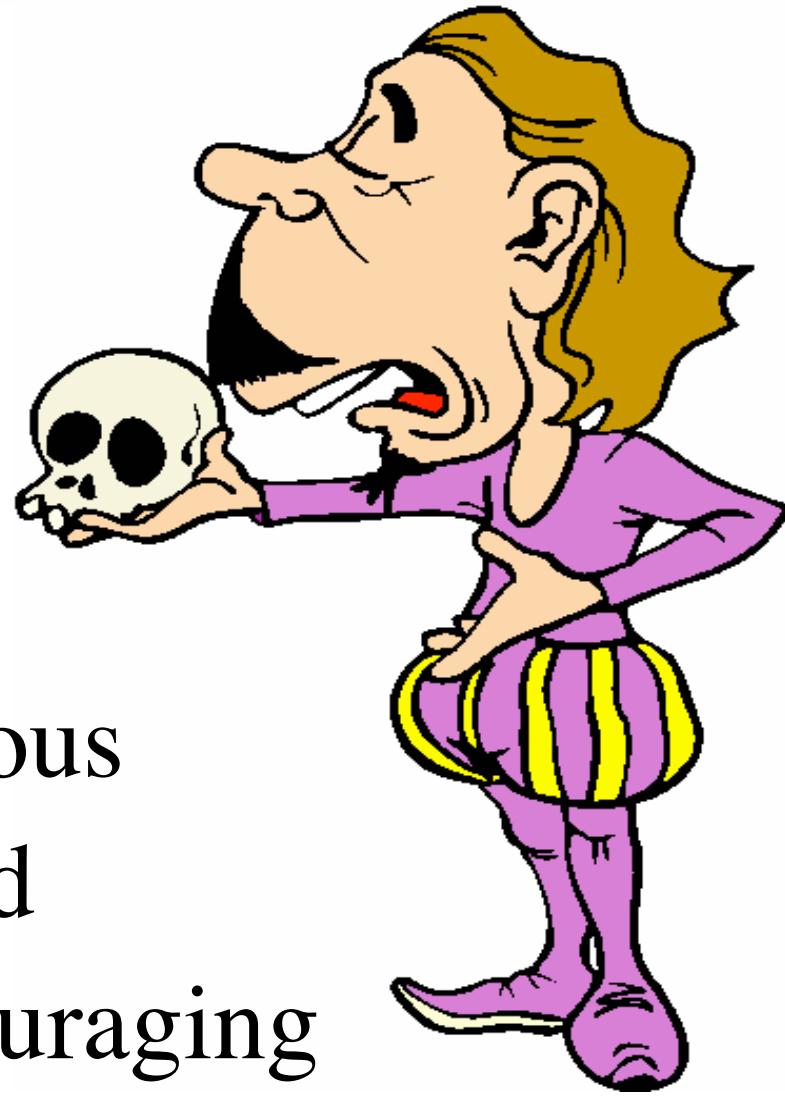
confident

frustrated

friendly

interested

anxious
bored
encouraging



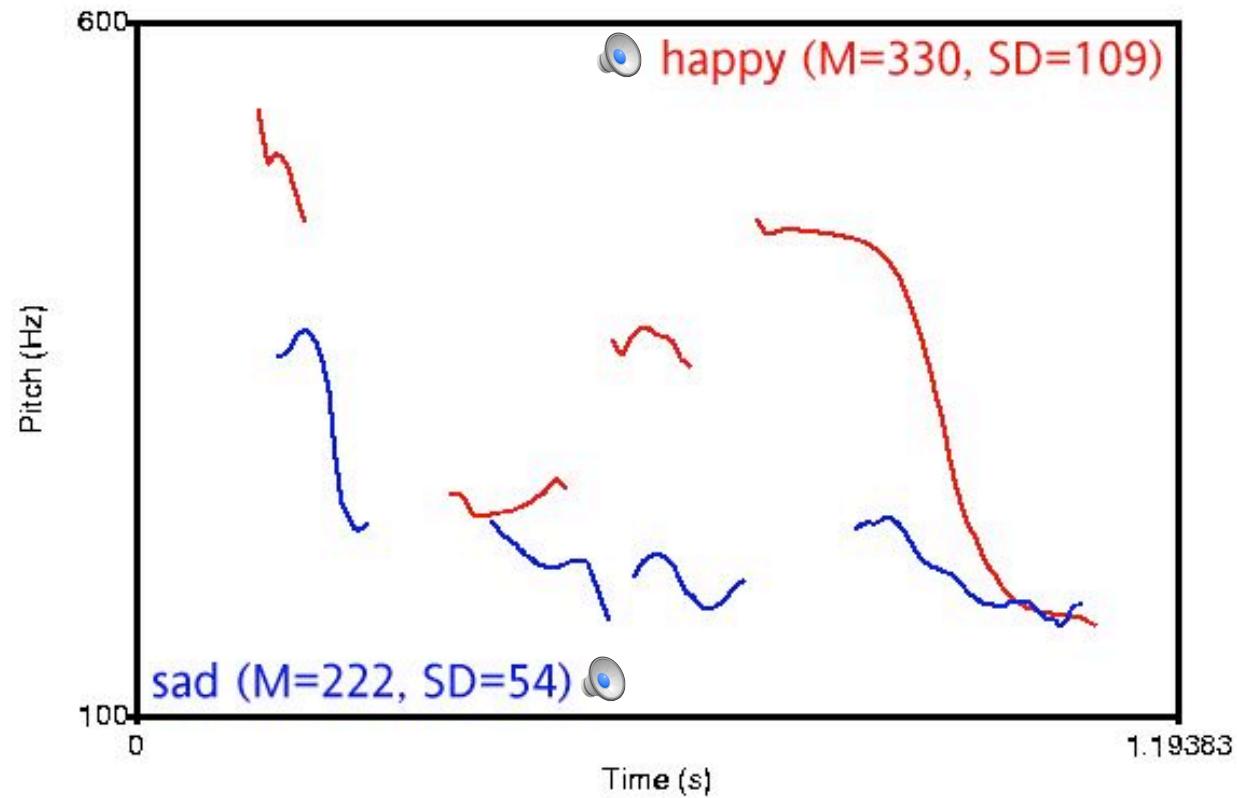
Detecting EPSaT Emotions

- Liscombe et al 2003

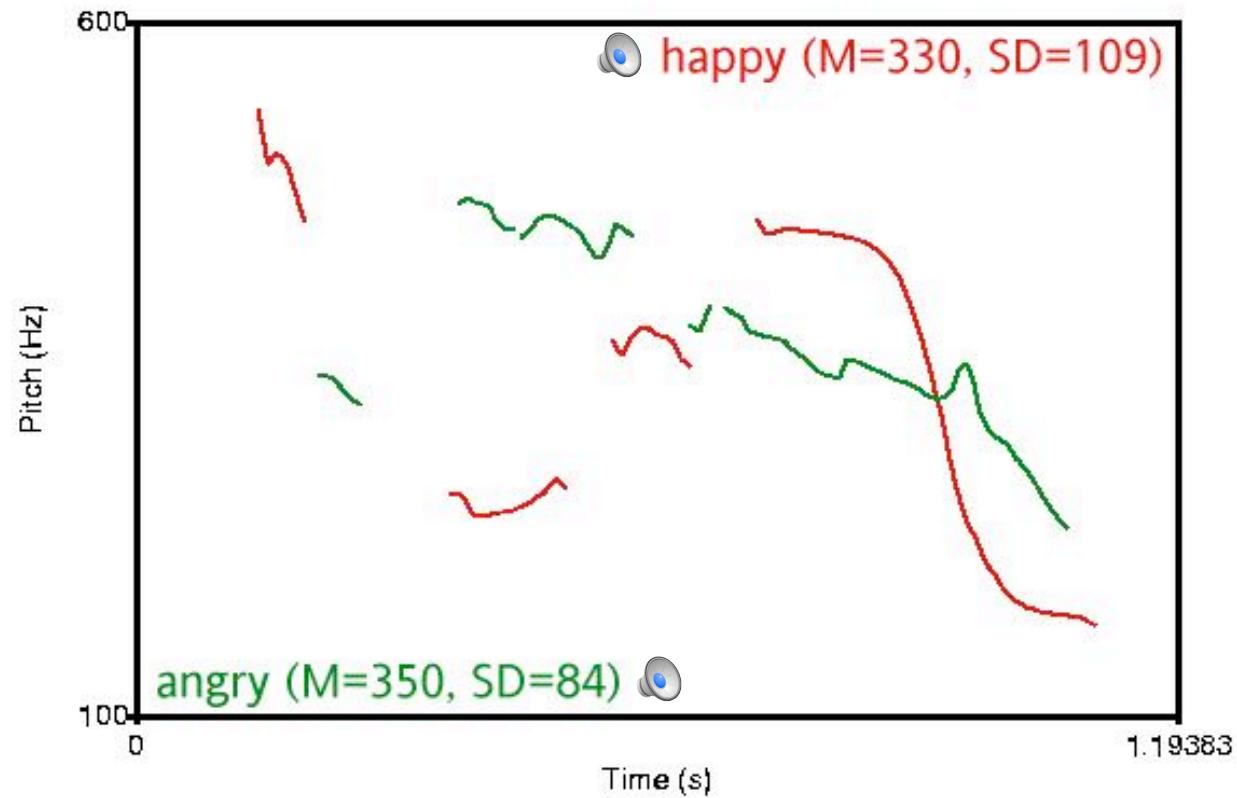
Liscombe et al. Features

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

Global Pitch Statistics



Global Pitch Statistics



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]

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?

EPSaT Discussion

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

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?

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.

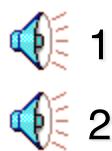
Data Labeling

- **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.

Emotion Samples

- **Neutral**

- ♦ *July 30*
- ♦ *Yes*



- **Disappointed/tired**

- ♦ *No*



- **Amused/surprised**

- ♦ *No*



- **Annoyed**

- ♦ *Yes*
- ♦ *Late morning (HYP)*



- **Frustrated**

- ♦ *Yes*
- ♦ *No*
- ♦ *No, I am ... (HYP)*
- ♦ *There is no Manila...*



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)

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

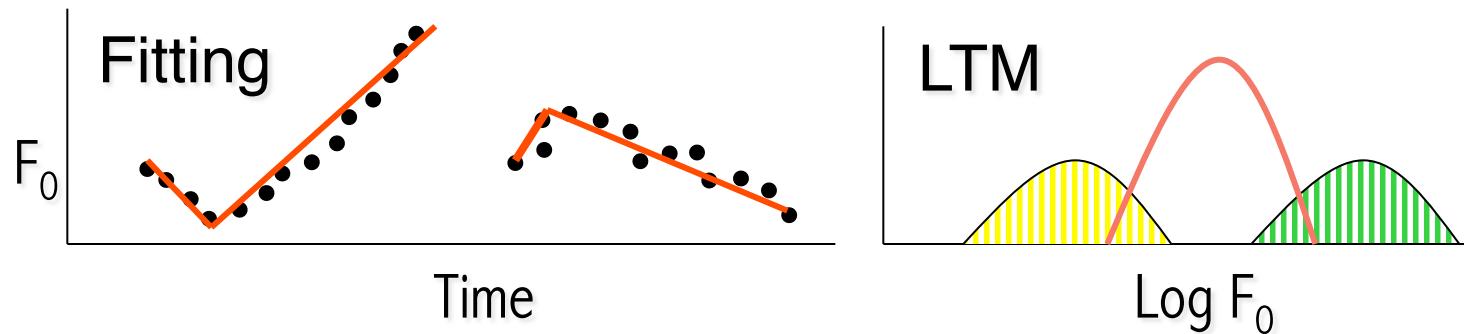
Prosodic Features

- **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

Prosodic Features (cont.)

- **Pitch features**

- F0-fitting approach developed at SRI (Sönmez)
- 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

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

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

Results: Human and Machine

Baseline ►

	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
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

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)

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

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

Ang et al '02 Conclusions

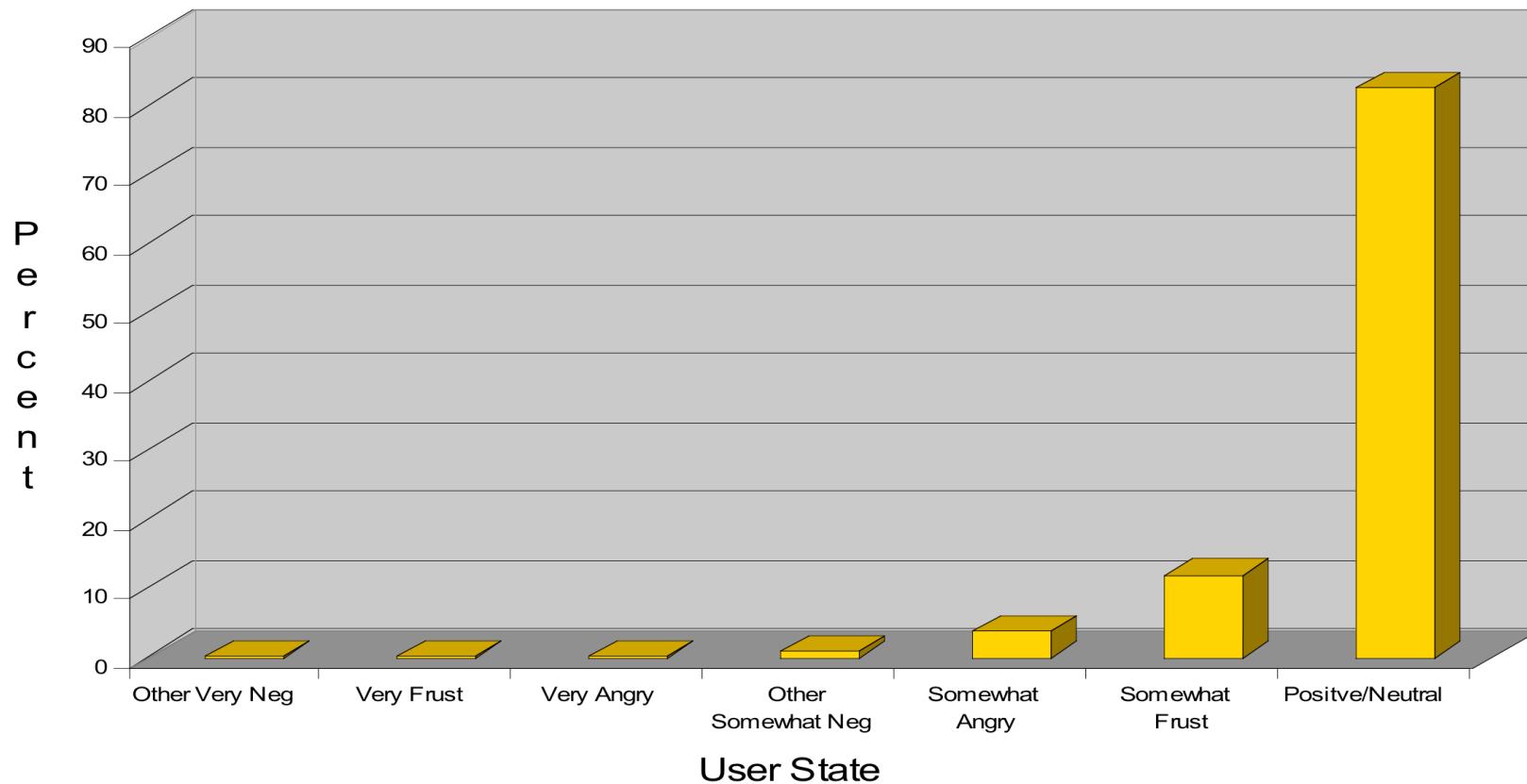
- Emotion labeling is a complex decision task
- Cases that labelers independently agree on are classified with high accuracy
 - Extreme emotion (e.g. 'frustration') is classified even more accurately
- Classifiers rely heavily on prosodic features, particularly duration and stylized pitch
 - Speaker 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

Example 3: “How May I Help YouSM” (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)!

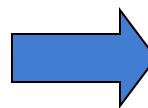
User Emotion Distribution

Emotion Label Distribution



Emotion Labels

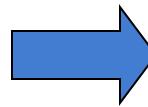
- Original Set:
 - ◆ Positive/Neutral



- ◆ Somewhat Frustrated
- ◆ Very Frustrated
- ◆ Somewhat Angry
- ◆ Very Angry
- ◆ Other Somewhat Negative
- ◆ Very Negative

- Reduced Set:
 - ◆ Positive

- ◆ Negative



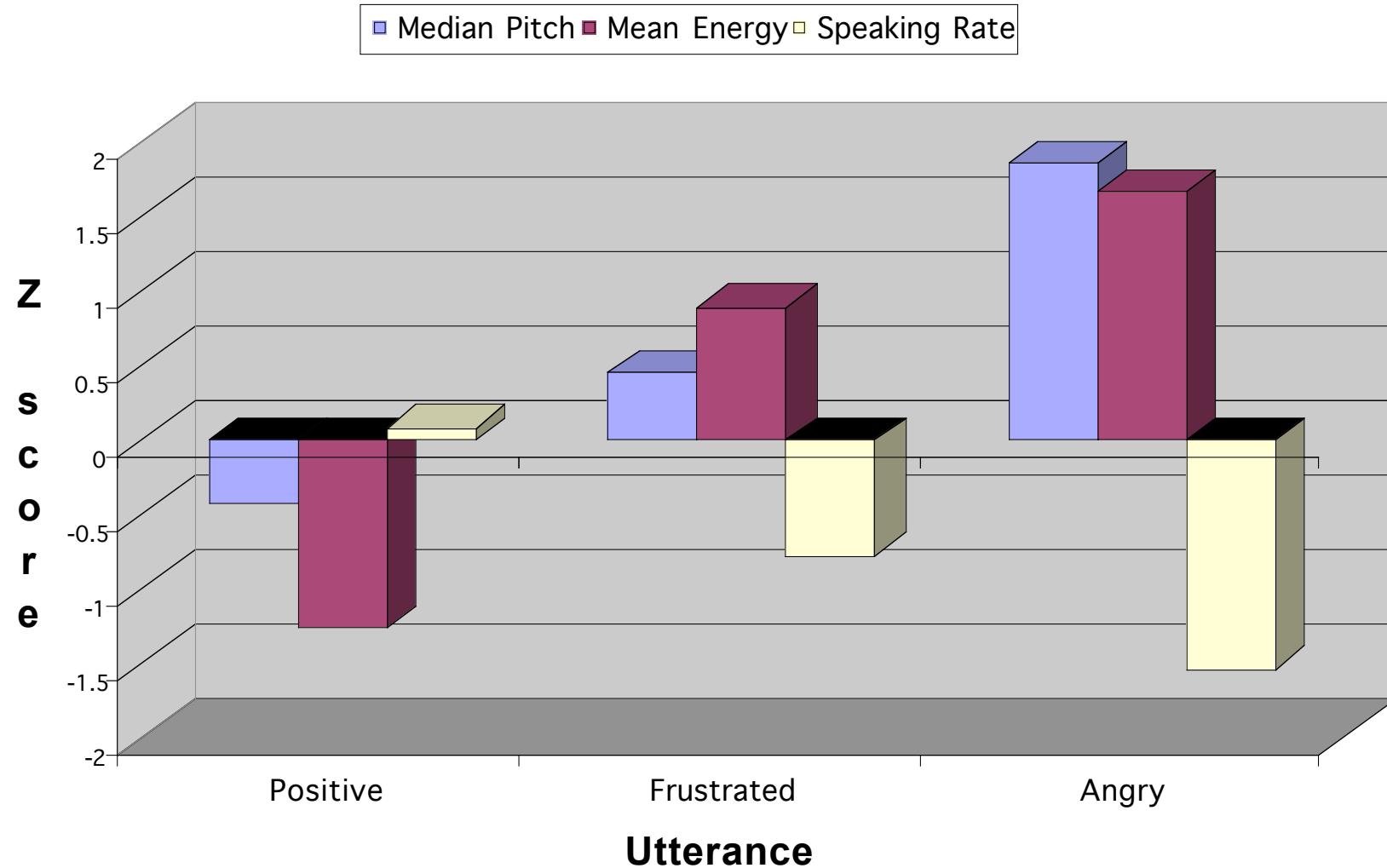
Pitch/Energy/Speaking Rate : Change

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	

20020221/0221080552atf1536

Slide from Jackson Liscombe

Pitch/Energy/Speaking Rate : Change



20020221/0221080552atf1536

Slide from Jackson Liscombe

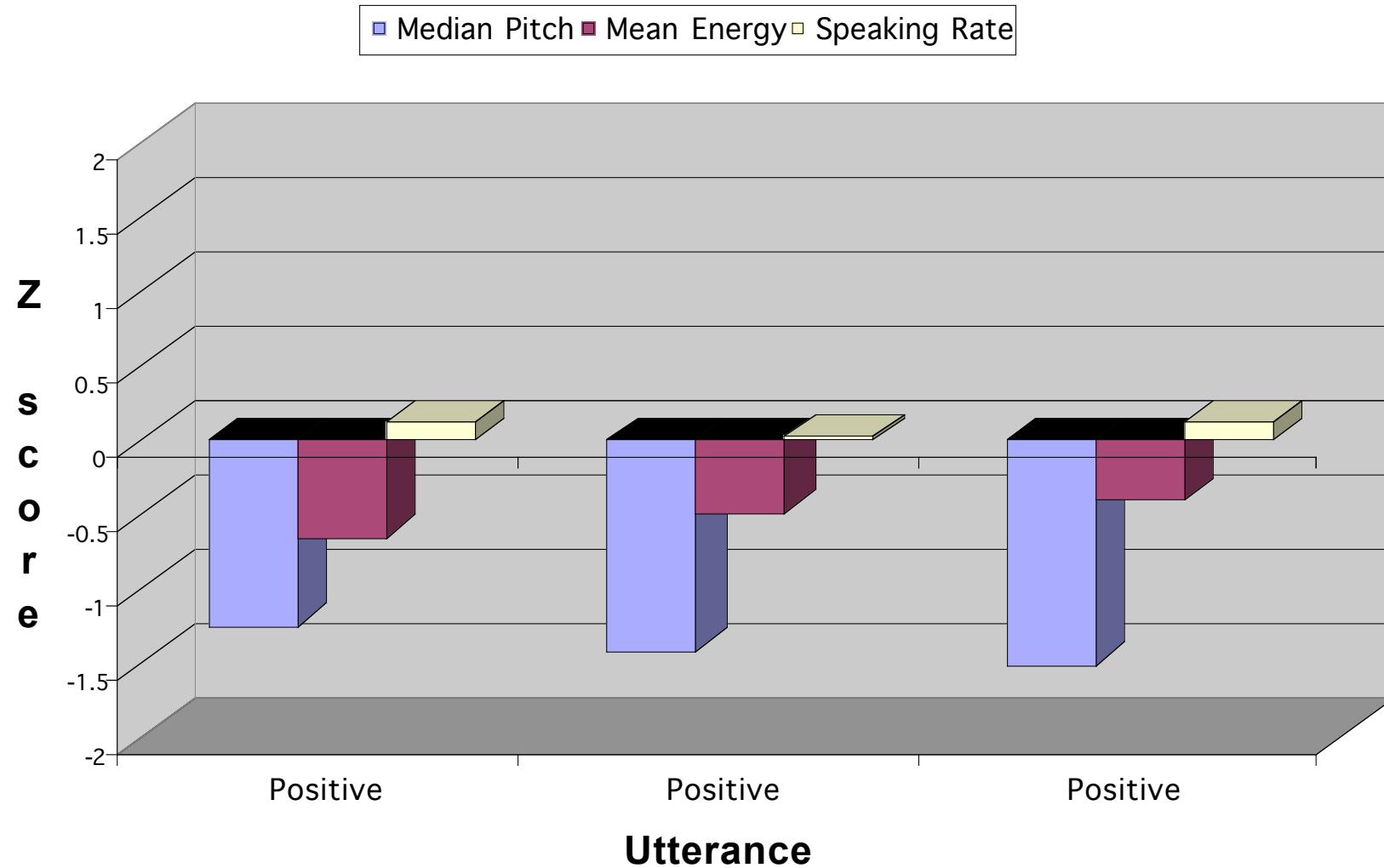
Pitch/Energy/Speaking Rate : No Change

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 your bill?		
User:	yes I would	Positive/Neutral	
Agent:	Are you calling from your home phone?		
User:	yes I am	Positive/Neutral	

20020317/0317220739atf9051

Slide from Jackson Liscombe

Pitch/Energy/Speaking Rate : No Change



20020317/0317220739atf9051

Slide from Jackson Liscombe

HMIHY Features

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

Lexical Features

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

Prosodic Features

- Pitch (F0)
 1. Overall minimum
 2. overall maximum
 3. overall median
 4. overall standard deviation
 5. mean absolute slope
 6. slope of final vowel
 7. longest vowel mean
- Other
 8. local jitter over longest vowel
- Energy
 9. overall minimum
 10. overall maximum
 11. overall mean
 12. overall standard deviation
 13. longest vowel mean
- Speaking Rate
 14. vowels per second
 15. mean vowel length
 16. ratio voiced frames to total frames
 17. percent internal silence

Contextual Features

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

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%

HMIHY Discussion

1. How is emotion communicated through speech?

- ◆ Novel features improve performance
 - transcription
 - pragmatics
 - context

Intelligent Tutoring Spoken Dialogue System

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

Why2 - Atlas

http://www.cs.pitt.edu/itspoke

Google

ITSPKE



[pr01_sess00_prob58]

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?

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 key's so he will fall faster than the keys.

Submit

Task 1

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

PROBLEM (TYPED): If a car is able to accelerate at 2 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 m/s^2 .

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

TUTOR₁: Uh let us talk of one car first.

STUDENT₁: ok. (*EMOTION = NEUTRAL*)

TUTOR₂: If there is a car, what is it that exerts force on the car such that it accelerates forward?

STUDENT₂: The engine (*EMOTION = POSITIVE*)

TUTOR₃: Uh well engine is part of the car, so how can it exert force on itself?

STUDENT₃: um... (*EMOTION = NEGATIVE*)

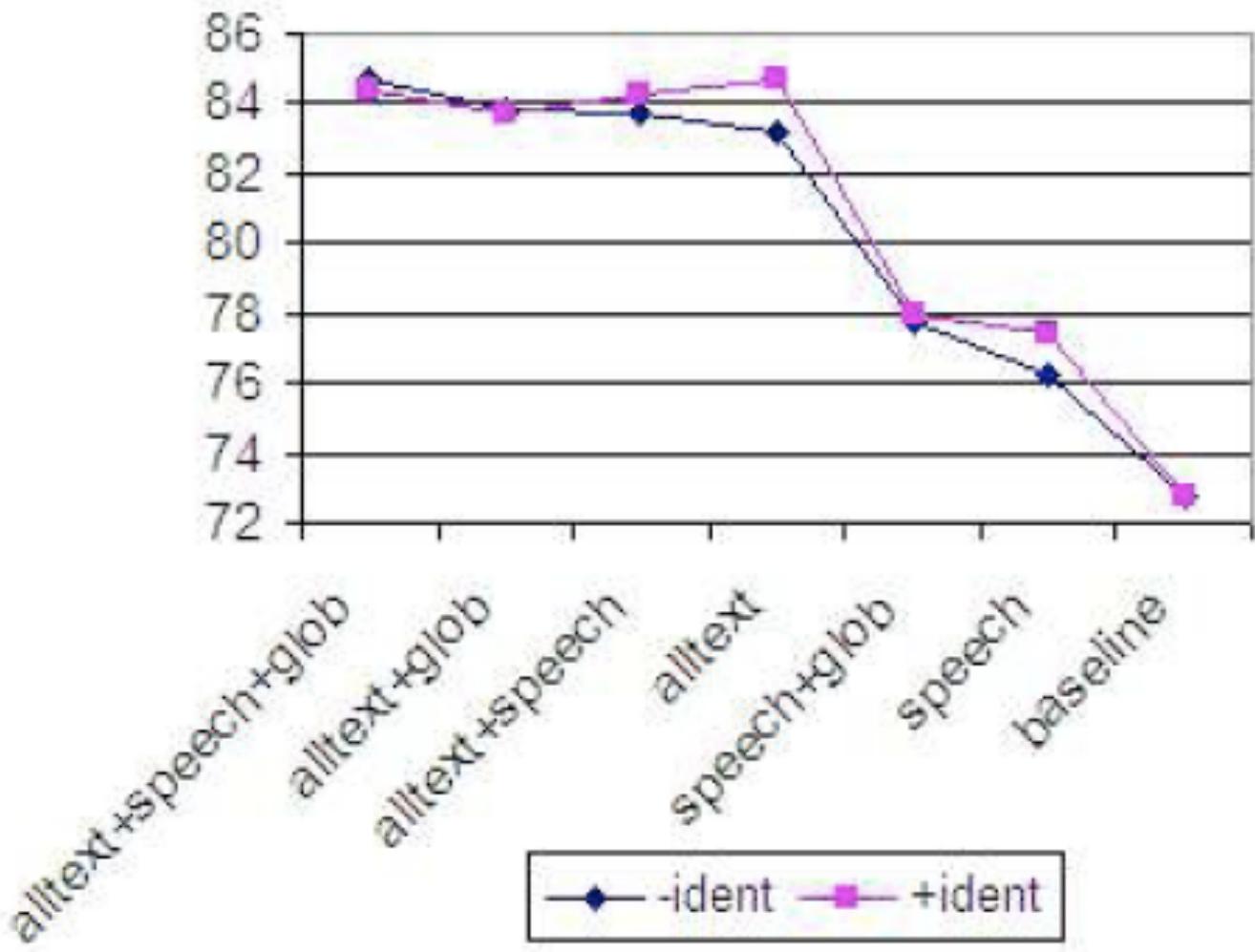
Acoustic-Prosodic Features

- 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

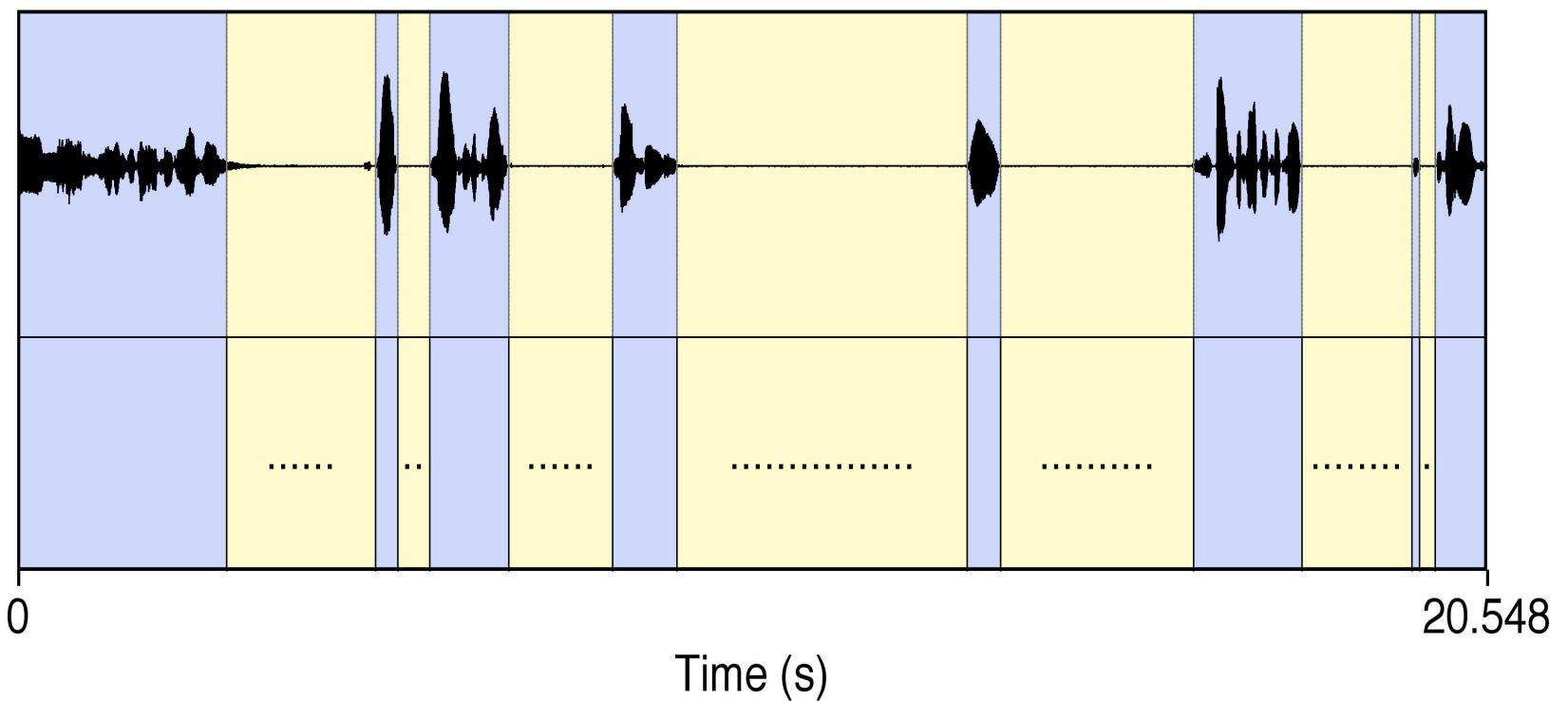


Liscombe et al: Uncertainty in ITSpoke

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] 🔊

Breath Group Segmentation



[71-67-1:92-113]

Liscombe et al: ITSpoke Features

- Automatic acoustic-prosodic
- Contextual
- Breath groups

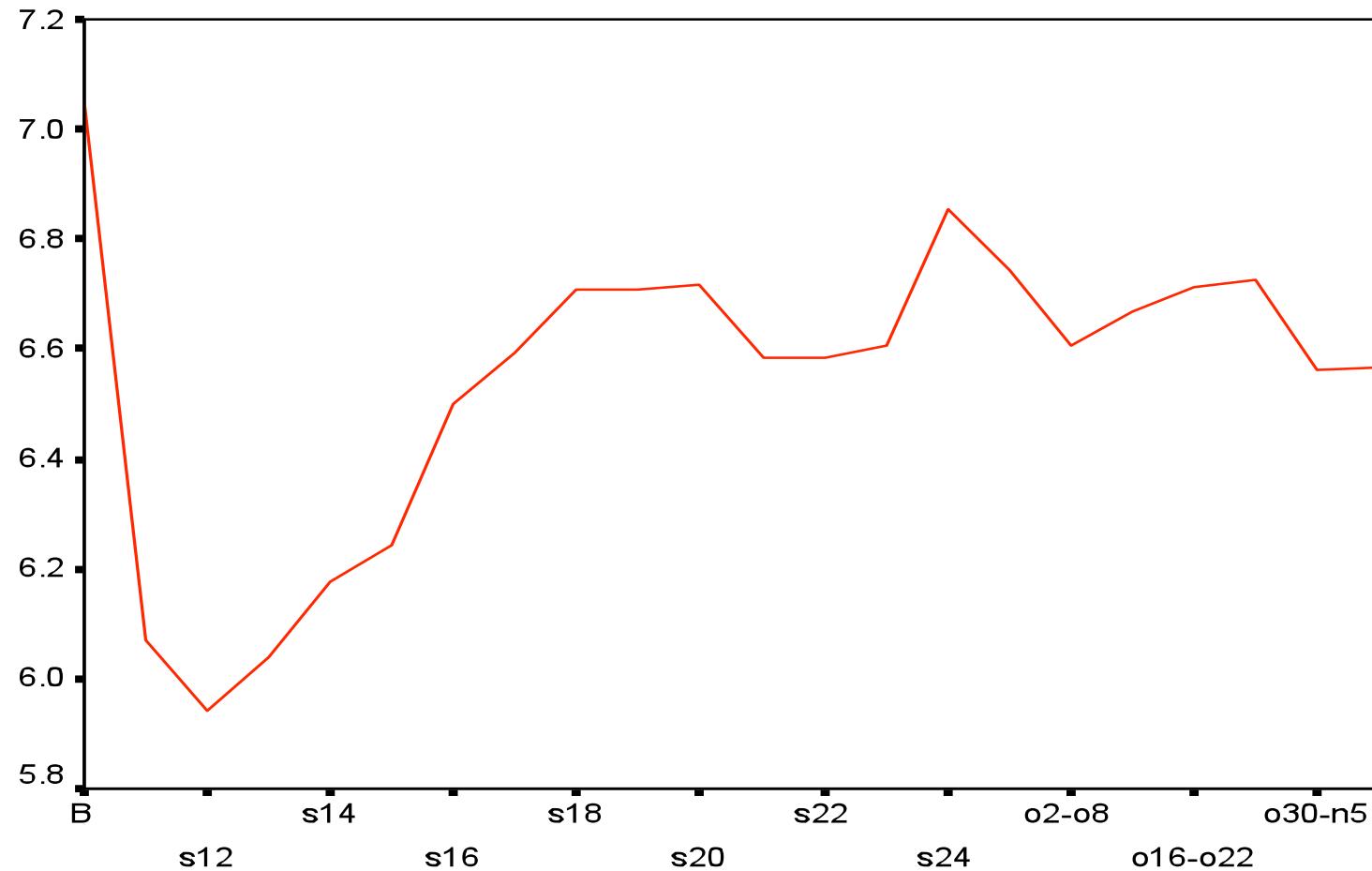
[Hirst & Christo, 1998]

Liscombe et al: ITSpoke Experiment

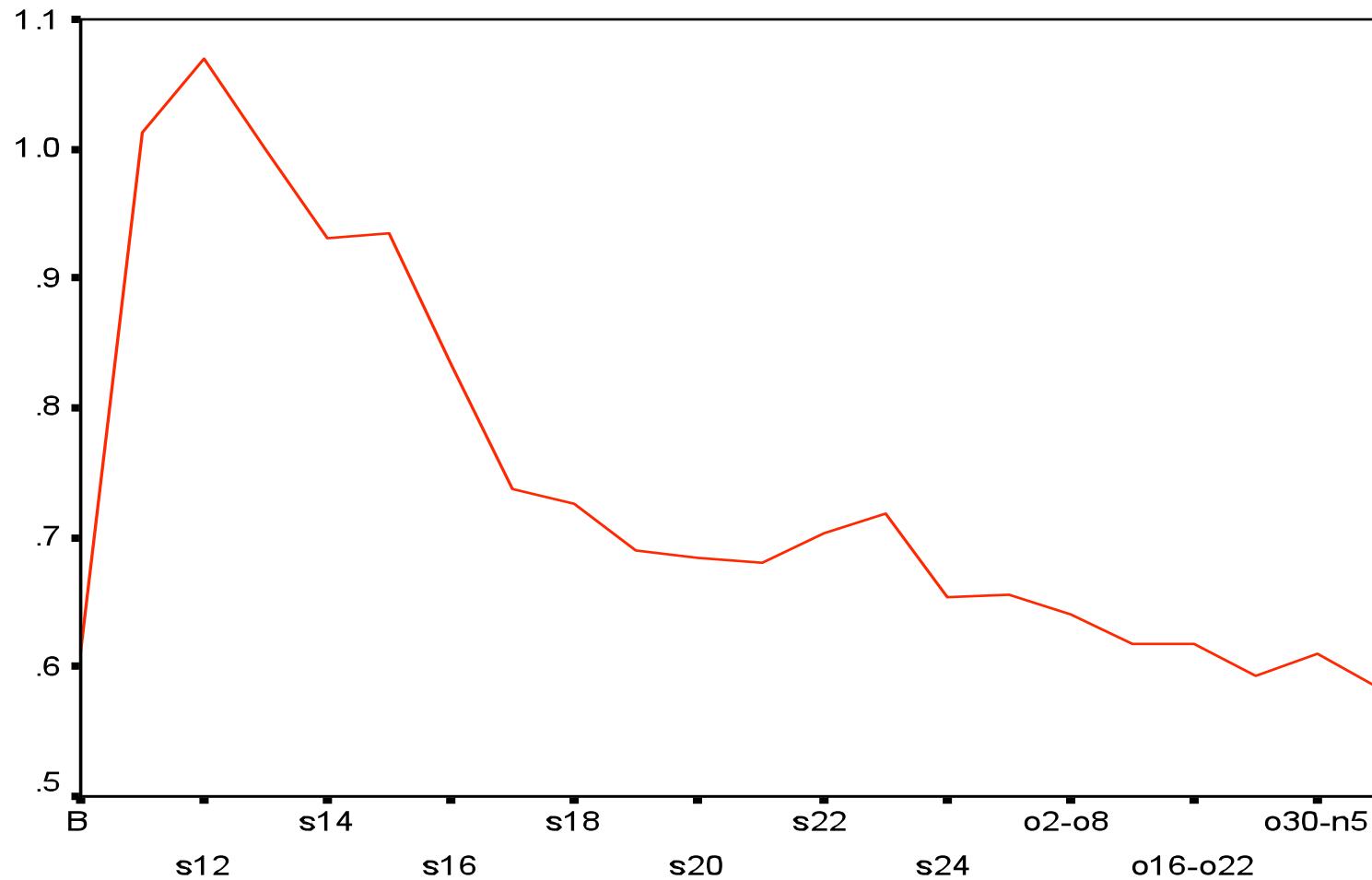
- Human-Human Corpus
- AdaBoost(C4.5) 90/10 split in WEKA
- Classes: *Uncertain* vs *Certain* vs *Neutral*
- Results:

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

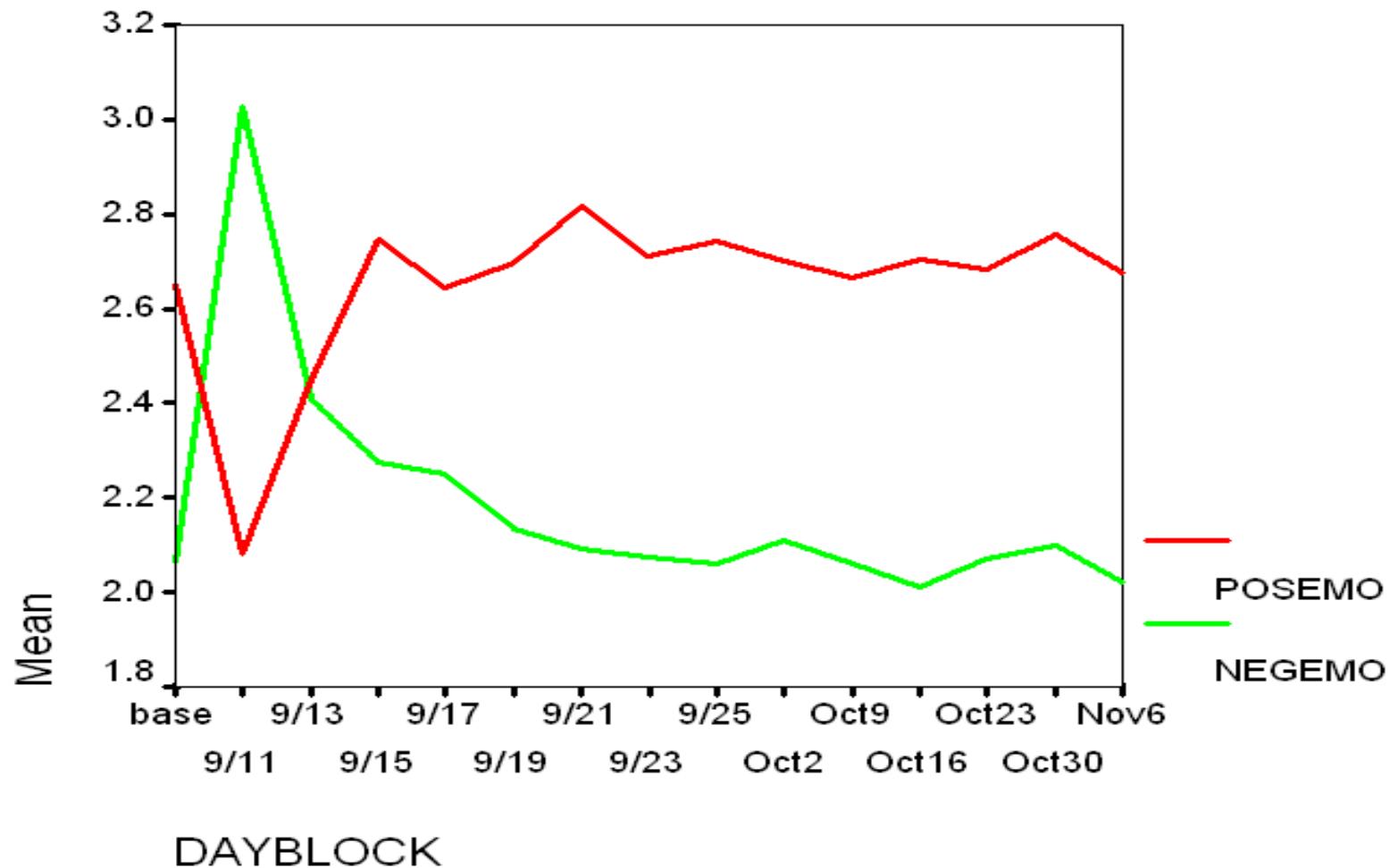
Pennebaker Livejournal.com Study: I, me, my



Pennebaker September 11 LiveJournal.com study: We, us, our



Pennebaker LiveJournal.com study: Positive and negative emotion words



Personality

- Mairesse & Walker '07 PERSONAGE (PERSONAlity GEnerator)
 - ◆ 'Big 5' personality trait model: extroversion, neuroticism, agreeableness, conscientiousness, openness to experience
 - ◆ Attempts to generate “extroverted” language based on traits associated with extroversion in psychology literature
 - ◆ Demo:
 - <http://mi.eng.cam.ac.uk/~farm2/personality/demo.html>: find your personality type

NLG modules	Introvert findings	Extravert findings	Parameter	Intro	Extra
Content selection and structure	Single topic Strict selection Problem talk, dissatisfaction	Many topics Think out loud*	VERBOSITY RESTATEMENTS REPETITIONS CONTENT POLARITY REPETITIONS POLARITY CLAIM POLARITY CONCESSIONS CONCESSIONS POLARITY POLARISATION POSITIVE CONTENT FIRST	low low low low low low avg low low low	high high low high high high avg high high high
Syntactic templates selection	Few self-references Elaborated constructions Many articles	Many self-references Simple constructions* Few articles	SELF-REFERENCES CLAIM COMPLEXITY	low high	high low
Aggregation Operations	Many words per sentence/clause Many unfilled pauses	Few words per sentence/clause Few unfilled pauses	RELATIVE CLAUSES WITH CUE WORD CONJUNCTION PERIOD ...	high high low high	low low high low
Pragmatic transformations	Many nouns, adjectives, prepositions (explicit) Many negations Many tentative words Formal Realism No politeness form Lower word count	Many verbs, adverbs, pronouns (implicit) Few negations Few tentative words Informal Exaggeration* Positive face redressment* Higher word count	SUBJECT IMPLICITTNESS NEGATION INSERTION DOWNTONER HEDGES: ·SORT OF, SOMEWHAT, QUITE, RATHER, ERR, I THINK THAT, IT SEEMS THAT, IT SEEMS TO ME THAT, I MEAN ·AROUND ·KIND OF, LIKE ACKNOWLEDGMENTS: ·YEAH ·RIGHT, OK, I SEE, WELL EMPHASIZER HEDGES: ·REALLY, BASICALLY, ACTUALLY, JUST HAVE, JUST IS, EXCLAMATION ·YOU KNOW TAG QUESTION INSERTION HEDGE VARIATION HEDGE REPETITION	low high high low avg low high low high low high low high low low low low low	high low low high high low high low high high high low
Lexical choice	Rich Few positive emotion words Many negative emotion words	Poor Many positive emotion words Few negative emotion words	LEXICON FREQUENCY <i>see polarity parameters</i> <i>see polarity parameters</i>	low	high

Table 2: Summary of language cues for extraversion, based on Dewaele and Furnham (1999); Furnham (1990); Mehl et al. (2006); Oberlander and Gill (2006); Pennebaker and King (1999), as well as PERSON-AGE's corresponding generation parameters. Asterisks indicate hypotheses, rather than results. For details on aggregation parameters, see Section 4.2.