CS 224S / LINGUIST 285
Spoken Language Processing

Andrew Maas
Stanford University
Spring 2022

Lecture 18: Flirtation, Intoxication

Original slides by Dan Jurafsky

Poster session next week!

- **Tuesday May 31 5:30pm – 7:30pm**
  Hewlett Lawn
- Print your poster (9 printed slides works!)
  - We provide poster boards + easels
  - Set up your poster by 5:30
- Present your poster 5:30-7:30
  - 2 mins walk through for audience
  - At least one person attend poster during session

Outline for today

- Extracting social meaning with supervised ML
  - Emotion recognition as a motivating example
- Case study: Flirtation
- Case study: Alcohol intoxication

Supervised ML as an analysis tool

1. Define your task / hypothesis (e.g. detecting alcohol intoxication in speech)
2. Collect/access data and annotations (y) for your task. Ensure data is representative for your problem
3. Define and test your modeling approach and inputs (x)
    1. Optimize to find x,f() to improve f(x)=y for your data
4. Analyze results, feature importance, model weights etc.

Supervised ML as an analysis tool

- Examples we will explore today::
  - Flirtation / interpersonal stance
  - Intoxication
- Framework is highly flexible. Requires careful decisions about datasets and what conclusions to draw
- Use cases:
  - Accurately recognize **y** in a live/ongoing system
    - Validate **x** is reasonable/ethical. Build best possible **f(x)**. Favor accuracy
  - Infer findings/relationships predictive of **y**
    - Careful analysis of **x, f(x)**. Favor interpretability

Supervised ML analysis choices

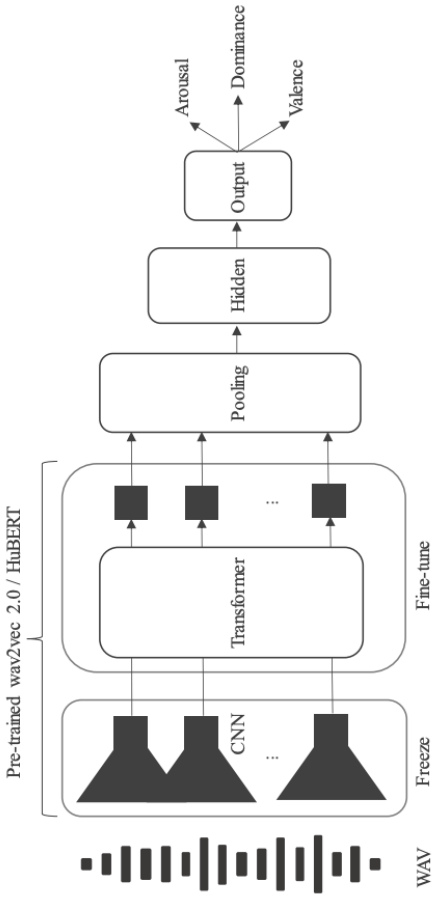
- Focus on prediction accuracy vs interpretable findings?
- What feature set? Use text + audio inputs?
  - Using foundation model features (wav2vec, Hubert, Bert) can improve predictive accuracy, makes interpretation hard
  - Often text features easiest to interpret (might require ASR)
- Data labeling + data collection define what the system will extract
  - Do label categories represent the right concept?
  - Is data labeled consistently and accurately?
  - Does training set represent true data distribution to study? Does it cover input variations

Emotion recognition with foundation model features

Table 1: *State-of-the-art 4-class emotion recognition performance on IEMOCAP using transformer-based architectures ranked by unweighted average recall (UAR) / weighted average recall (WAR). The table encodes whether the base or large (L) architecture was used as well as whether the pre-trained model was fine-tuned for speech recognition (FT-SR). The column FT-D marks if the transformer layers were further fine-tuned during the down-stream classification task.*

	Work	Model	L	FT-SR	FT-D	UAR	WAR
1	Krishna [27]	w2v2-L	✓			60.0	
2	Yuan *et al.* [28]*	w2v2-L	✓			62.5	62.6
3	Wang *et al.* [23]	w2v2-b					63.4
4	Yang *et al.* [29]	w2v2-b				63.4	
5	Pepino *et al.* [30]	w2v2-b		✓		63.8	
6	Wang *et al.* [23]	hubert-b					64.9
7	Yang *et al.* [29]	hubert-b				64.9	
8	Wang *et al.* [23]	w2v2-L	✓				65.6
9	Yang *et al.* [29]	w2v2-L	✓			65.6	
10	Pepino *et al.* [30]	w2v2-b				67.2	
11	Wang *et al.* [23]	hubert-L	✓				67.6
12	Yang *et al.* [29]	hubert-L	✓			67.6	
13	Chen and Rudnicky [31]	w2v2-b			✓	69.9	
14	Makiuchi *et al.* [32]	w2v2-L	✓			70.7	
15	Wang *et al.* [23]	w2v2-b		✓	✓		73.8
16	Chen and Rudnicky [31]	w2v2-b			✓	74.3	
17	Wang *et al.* [23]	hubert-b			✓		76.6
18	Wang *et al.* [23]	w2v2-L	✓	✓	✓		76.8
19	Wang *et al.* [23]	w2v2-b			✓		77.0
20	Wang *et al.* [23]	w2v2-L	✓		✓		77.5
21	Wang *et al.* [23]	hubert-L	✓	✓	✓		79.0
22	Wang *et al.* [23]	hubert-L	✓		✓		79.6

* For a fair comparison we report the result on utterance-level. Authors report better performance on phonetic level, though.



(Wagner *et al*, 2022)

Emotion recognition with foundation model features

1. We see a roughly 10% better performance with models where the weights of the pre-trained model were not frozen during the down-stream task.
2. Using a pre-trained model fine-tuned for speech recognition does not help with the down-stream task (e. g. row 15 vs row 19 −3.2%).
3. When the base and the large architecture of the same model type are tested within the same study, the large one yields better results (e. g. row 17 vs row 22 +3.0%), though the difference can be quite small (e. g. row 19 vs row 20 +.5%).
4. Likewise, in that case HuBERT outperforms wav2vec 2.0 (e. g. row 22 vs row 20: +2.1%).
5. When performing a fine-tuning of the transformer layers, a simple average pooling in combination with a linear classifier built over wav2vec 2.0 or HuBERT as proposed by Wang *et al.* [23] seems sufficient and shows best performance in the ranking. However, some of the more complex models like the cross-representation encoder-decoder model proposed by Makiuchi *et al.* [32] only report results without fine-tuning the pre-trained model during the down-stream task.

(Wagner *et al*, 2022)

Impact of fine tuning and text features

CCC = concordance correlation coefficient. Ranges [-1,1] with 1 being perfect prediction/true score correlation

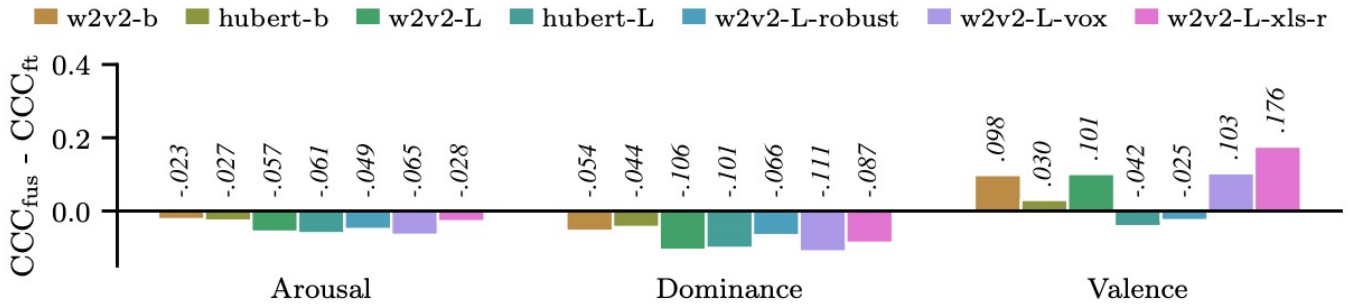


Figure 6: *Text & audio fusion results for arousal, dominance, and valence prediction on MSP-Podcast. Embeddings from the already fine-tuned models are concatenated with BERT embeddings extracted from automatic transcriptions, whereupon a two-layer feed-forward neural network is trained. We show the difference to results with the fine-tuned (ft) models from Figure 2.*

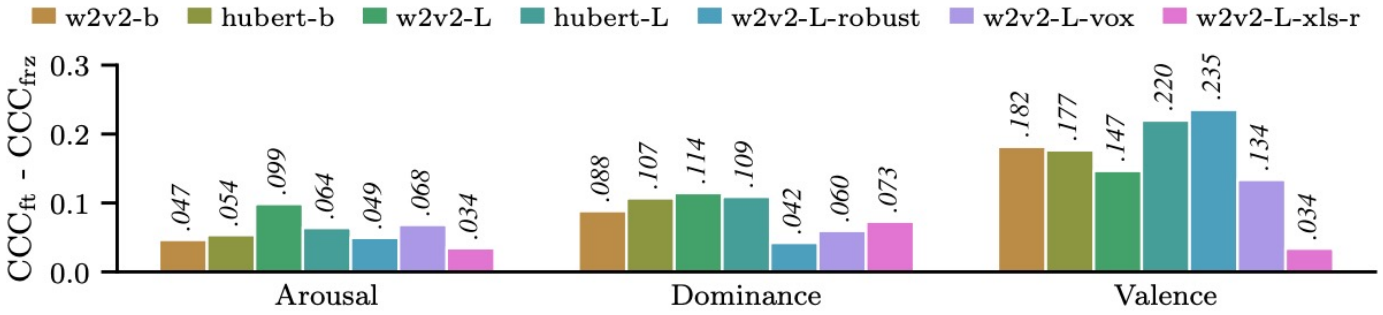


Figure 9: *Difference of fine-tuned (ft) to frozen (frz) CCC performance for arousal, dominance, and valence prediction on MSP-Podcast. The fine-tuned results are from Figure 2, where transformer and output layers are jointly trained. For the frozen results, we keep all transformer layers frozen and simply train the output head. Results show that fine-tuning the transformer layer is worth the computational cost it incurs.*

(Wagner *et al*, 2022)

Questions in Emotion Recognition

- How do we know what emotional speech is?
  - Acted speech vs. natural (hand labeled) corpora
- What can we classify?
  - Distinguish among multiple 'classic' emotions
  - Distinguish
    - Valence: is it positive or negative?
    - Activation: how strongly is it felt? (sad/despair)
- What features best predict emotions?
- What techniques best to use in classification?

Slide from Julia Hirschberg

Social Signal Processing = Affect/Emotion Detection

- Detecting frustration of callers to a help line
- Detecting stress in drivers or pilots
- Detecting depression, intoxication
- Detecting interest, certainty, confusion in on-line tutors
  - Pacing/Positive feedback
- Hot spots in meeting summarizers/browsers

Outline for today

- Extracting social meaning with supervised ML
  - Emotion recognition as a motivating example
- Case study: Flirtation
- Case study: Alcohol intoxication

Interpersonal Stance: Our Goals

- Friendliness
- Assertiveness
- Flirtation
- Awkwardness

Methodology

- Given speech and text from a conversation
- Can we tell if a speaker is
  - Awkward?
  - Flirtatious?
  - Friendly?
- Dataset:
  - 1000 4-minute “speed-dates”
  - Each subject rated their partner for these styles
  - The following segment has been lightly signal-processed:

(Jurafsky, Ranganath & McFarland. 2009)

(Ranganath, Jurafsky & McFarland. 2009)

speed dating *noun*

Menu

speed dating [uncountable]

an event at which you meet and talk to a lot of different people for only a few minutes at at a time. People do this in order to try to meet someone and have a romantic relationship.

E.J. Finkel, P.W. Eastwick. 2008. Speed-dating. *Current Directions in Psychological Science*, 17 (3) (2008), p. 193

Place, S. S., Todd, P. M., Penke, L., & Asendorpf, J. B. (2009). The ability to judge the romantic interest of others. *Psychological Science*, 20(1), 22-26.

M.E. Ireland, R.B. Slatcher, P.W. Eastwick, L.E. Scissors, E.J. Finkel, J.W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22 (1) (2011), p. 39

Extracting social meaning

- Stance
  - Friendly, flirt, awkward, assertive
- Social Bond
  - Clicking or Connection
  - Romantic Interest
- **946 4-minute dates**
  - ~800K words, hand-transcribed
  - ~60 hours, from shoulder sash recorders
  - 3 events, 20x20=400 dates x 3
  - Date perceptions, demographics, preferences

Data annotation

- Each speaker wore a microphone
- So each date had two recordings
- The wavefile from each speaker was manually segmented into 4-minute dates
- Professional transcription service produced:
  - words, laughter, disfluencies
  - timestamps for turn beginning and end (1 second)
    - for 10% of the dates, timestamp at 0.1 second granularity
  - using both recordings

Study 1:
What we attempted to predict

- Conversational style:
  - How often did they behave in the following ways on this date?
    - On a scale of 1-10 (1=never, 10=constantly)

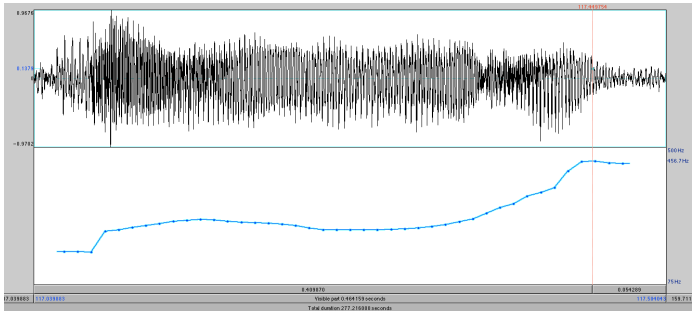
awkward

friendly

flirtatious

assertive

Features



- Prosodic
  - pitch (min, mean, max, std)
  - intensity (min, max, mean, std)
  - duration of turn
  - rate of speech (words per second)
- Lexical
  - negation words (don't, didn't, won't, can't, not, never)
  - hedges (kind of, sort of, probably, I don't know)
  - personal pronouns (I, you, we, us)
- Dialog
  - questions
  - backchannels ("uh-huh", "yeah")
  - appreciations ("Wow!", "That's great!")
  - sympathy ("That's awful!" "Oh, that sucks!")

Engineering studies
16 binary classifiers

- Female ±Awkward, Male ±Awkward,
- Female ±Friendly, Male ±Friendly,
- Female ±Flirtatious, Male ±Flirtatious,
- Female ±Assertive, Male ±Assertive

- Each study run twice, on:
  - self-assessed
  - alter-assessed
- Multiple classifier experiments
  - L1-regularized logistic regression
  - SVM w/RBF kernel

Test set

- For each of the 16 experiments
  - Sort all 946 dates
  - Choose top 10% as positive class
  - Choose bottom 10% as negative class
  - ignore 80% of dates in the middle!
- 5-fold cross-validation within this small training and test set

- Goal: distinguishing social interactants who are reported to exhibit (or not exhibit) clear social intentions or styles

Results using SVM Classifier

Using my speech to predict what my date says about me

	Male speaker	Female speaker
Flirting	65%	78%
Friendly	71	64
Awkward	67	67
Assertive	65	69

Results using SVM Classifier

- Using my speech to predict what I say about myself

	Male speaker	Female speaker
Flirting	66%	74%
Friendly	76	71
Awkward	63	67
Assertive	73	64

What do flirters do?

- Women when flirting:
  - raise pitch ceiling
  - talk faster
  - say “I” and “like”, use more hedges
  - laugh at themselves
- Men when flirting:
  - raise their pitch floor
  - laugh at their date (teasing?)
  - say “you”
  - don’t use words related to academics
  - say “um”, “I mean”, “you know”

Unlikely words for male flirting

academia

interview

teacher

phd

advisor

lab

research

management

finish

Assertive

- Assertive men
  - talk more
  - use more negative emotion
  - lower their pitch floor
  - use more agreements and appreciations
  - use more “um”, “you”
  - use less negation
- Assertive women:
  - use more negation (“no”, “didn’t”, “don’t”)
  - talk about academics
  - are less sympathetic
  - accommodate more (content words)
  - use more “I” and “I mean”
  - use less negative emotion

What makes an awkward conversationalist?

- Awkward people:
  - use more hedges
  - ask more questions
- Awkward men
  - don’t talk about academics
  - do swear or use negative emotion
- Awkward women:
  - do talk about academics
  - talk more, and talk faster
  - don’t laugh at their date
  - don’t use “I”

(Actionable?) conclusions

- How to date::
- Don’t talk about your advisor
- Focus on the empowered party
- Flirting women raise pitch ceiling – flirting men raise pitch floor

What makes someone seem friendly? "Collaborative conversational style"

Related to the "collaborative floor" of Edelsky (1981), Coates (1996)

- **Friendly people:**
  - laugh at themselves
  - don't use negative emotions
- **Friendly men**
  - are sympathetic and agree more often
  - don't interrupt
  - don't use hedges
- **Friendly women:**
  - higher max pitch
  - laugh at their date

Intoxication

Hollien et al 2001

- Methods:
  - 35 young adults, 19 males, 16 females
  - given series of doses of alcohol
  - speech collected at 4 BAC stages 
    - Rainbow passage
    - difficult words (buttercup, shapupie)
    - extemp speech (“Tell us about your favorite TV program)
    - head-mounted mics
  - Investigated:
    - F0 mean and variance
    - duration/rate of speech
    - intensity
    - disfluencies

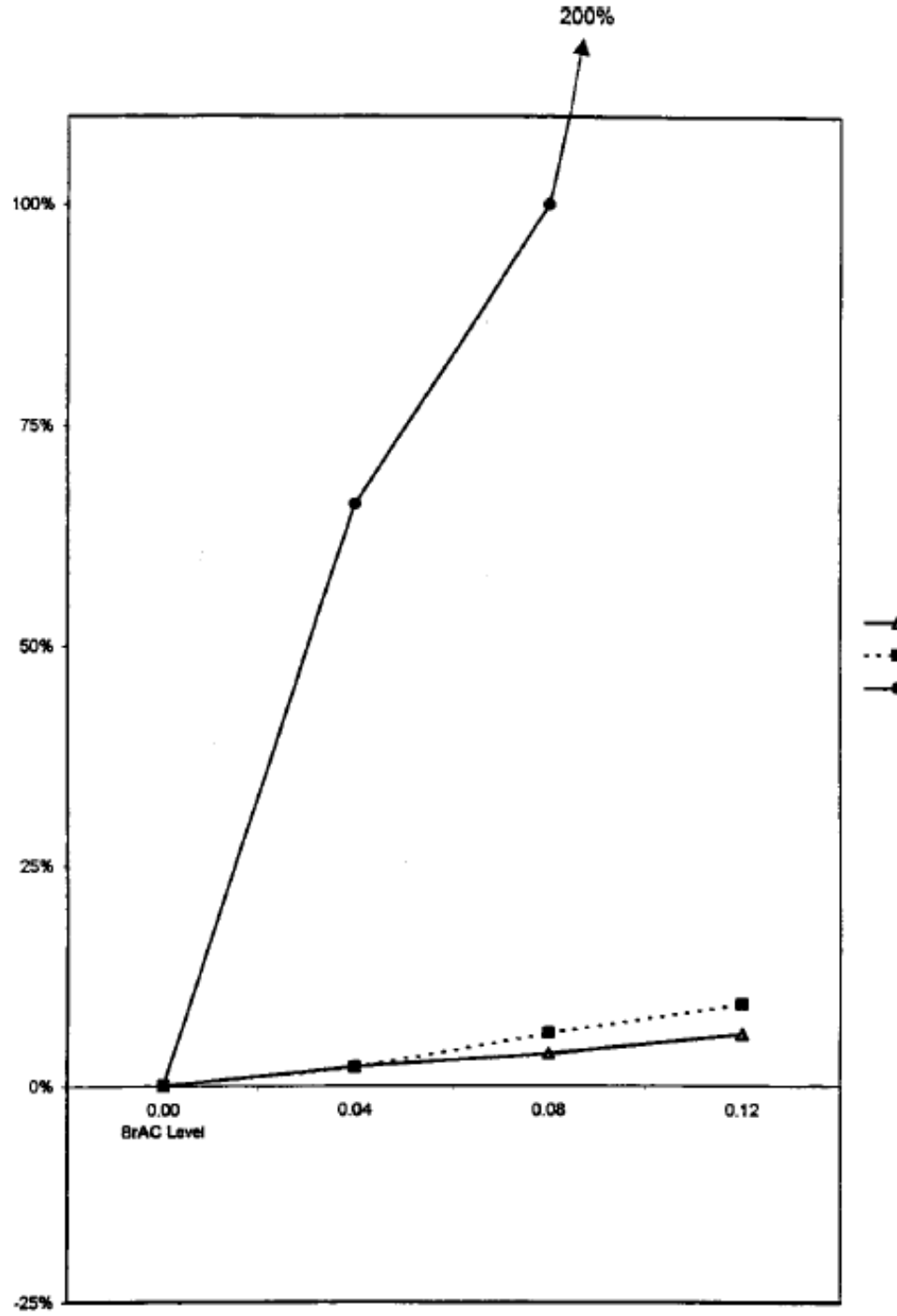
Hollien et al 2001 Results: Duration

Group	Level of intoxication (BrAC) 0.00	0.04	0.08	0.12	Shift (0.00–0.12)
Men					
Mean (s)	25.3	25.8	26.8	27.6	+2.3
S.D. (s)	2.9	2.5	2.1	2.5	
Women					
Mean (s)	25.1	25.5	25.7	27.5	+2.4
S.D. (s)	2.2	2.2	2.4	2.7	

Hollien et al 2001 Results: Disfluencies

Subjects	*N*	Experimental condition (BrAC) 0.00	0.04	0.08	0.12
Males	19				
Mean		3.2	4.7	6.5	8.6
SD		2.0	2.6	3.1	3.4
Females	16				
Mean		2.2	3.5	4.7	6.1
SD		1.7	2.2	2.7	3.0
Mean	35	2.7	4.1	5.6	7.4

Hollien et al 2001 Results: Magnitudes



A famous case study

- Johnson, K., Pisoni, D. & Bernacki, R. (1990) Do voice recordings reveal whether a person is intoxicated?: A case study. Phonetica. 47: 215-237.

Exxon Valdez

Exxon Valdez oil spill

From Wikipedia, the free encyclopedia Coordinates: 60.83333°N 146.86667°W

The ***Exxon Valdez* oil spill** occurred in Prince William Sound, Alaska, on March 24, 1989, when the *Exxon Valdez*, an oil tanker bound for Long Beach, California, struck Prince William Sound's Bligh Reef and spilled 260,000 to 750,000 barrels (41,000 to 119,000 m^3) of crude oil.[1][2] It is considered to be one of the most devastating human-caused environmental disasters.[3] As

***Exxon Valdez* oil spill**

3 days after *Exxon Valdez* ran aground

Location	Prince William Sound, Alaska
Coordinates	60.83333°N 146.86667°W
Date	24 March 1989
Cause	

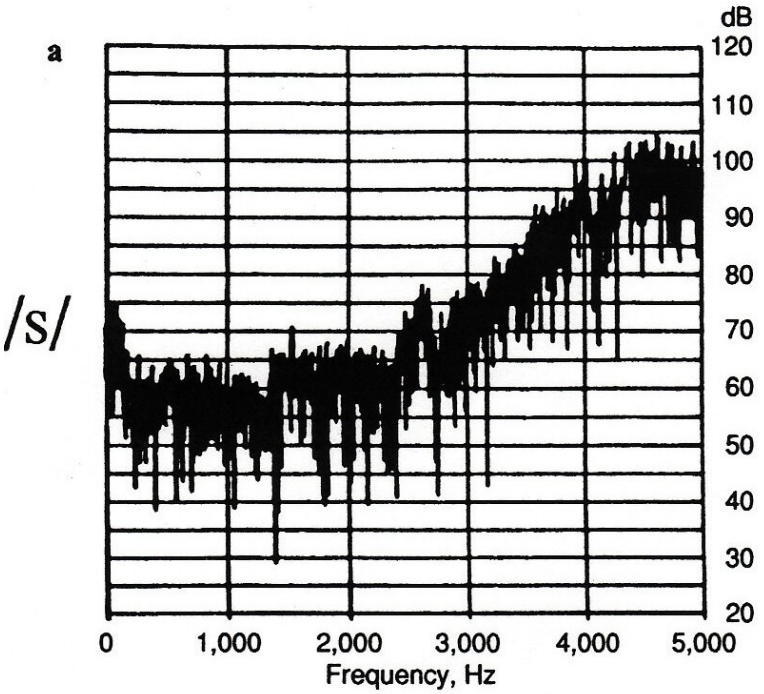
Was Captain Hazelwood drunk?

- Not clear if this is relevant, since seems like other questionable corporate things were going on:
  - he was asleep below deck
  - the third mate was in charge of the wheelhouse
  - the ship's radar was broken
- But is a well-studied case

Johnson et al examined 3 kinds of cues

- Segmental Effects (phoneme, syllable, word level)
- Disfluencies
- Suprasegmental Effects (stress, intonation, etc.)

Keith Johnsons /s/ and/ʃ/



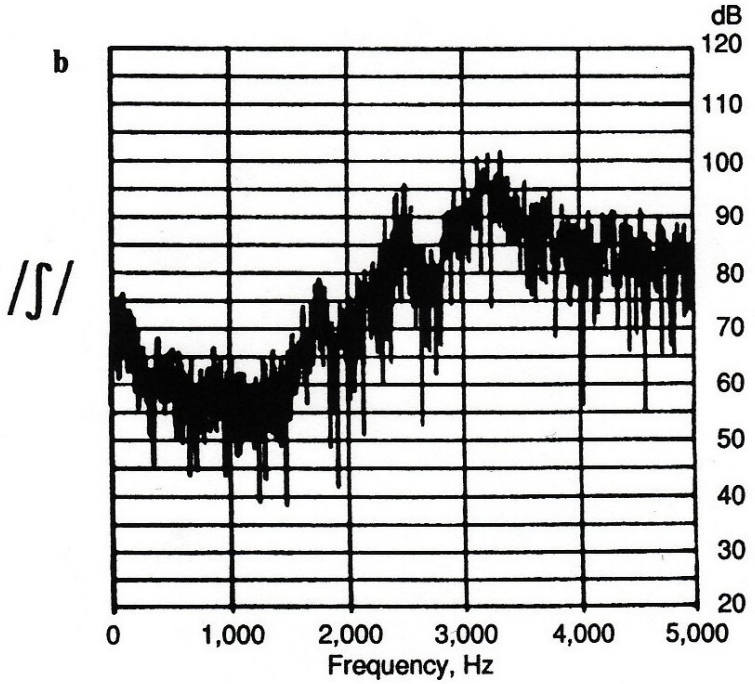


Fig. 1. Power spectra of /s/ (**a**) and /ʃ/ (**b**) produced by K. J. in a quiet recording booth with recording equipment responsive up to 5,000 Hz.

e.g. “sun” vs “shun”

/ʃ/: Captain Hazelwood

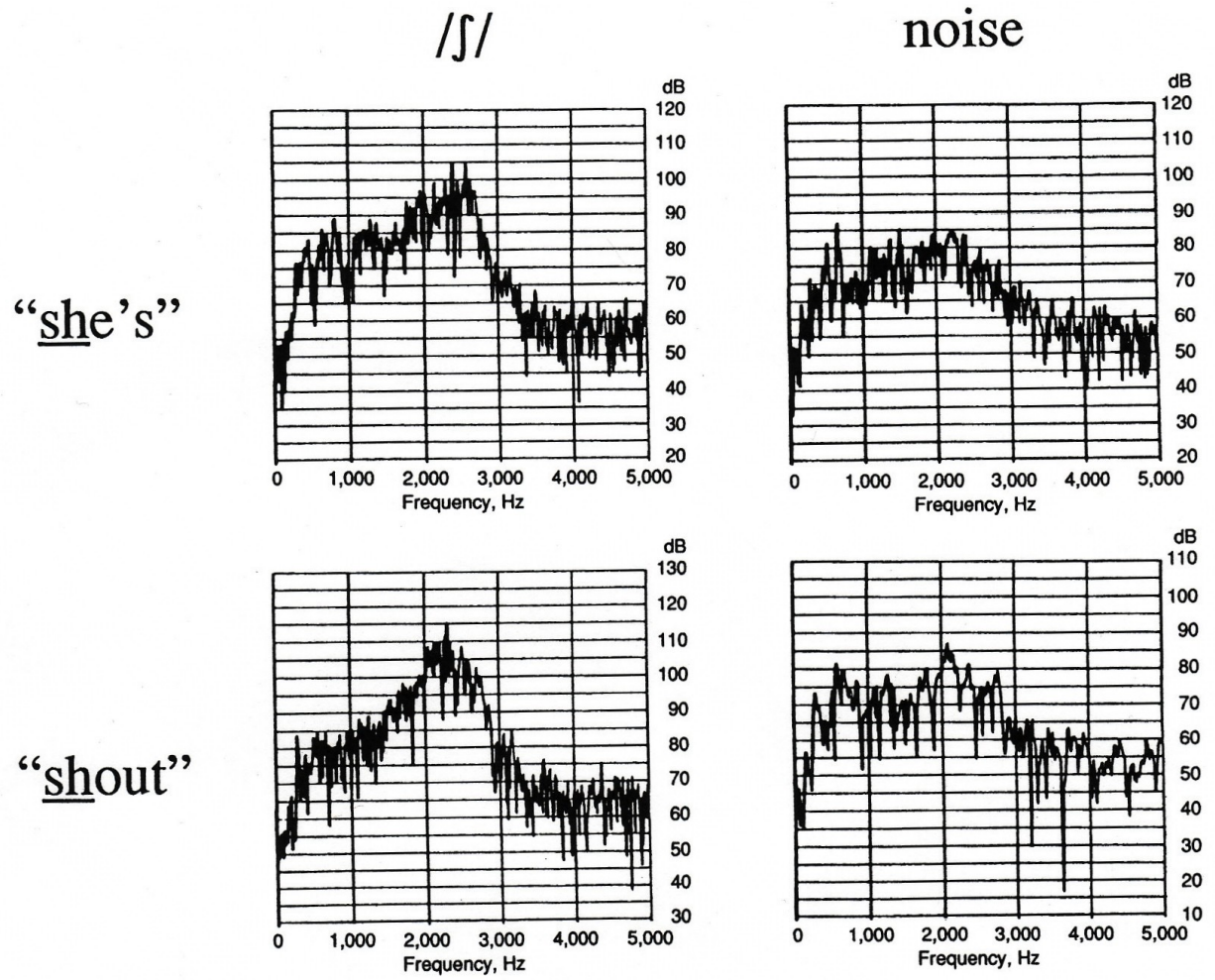
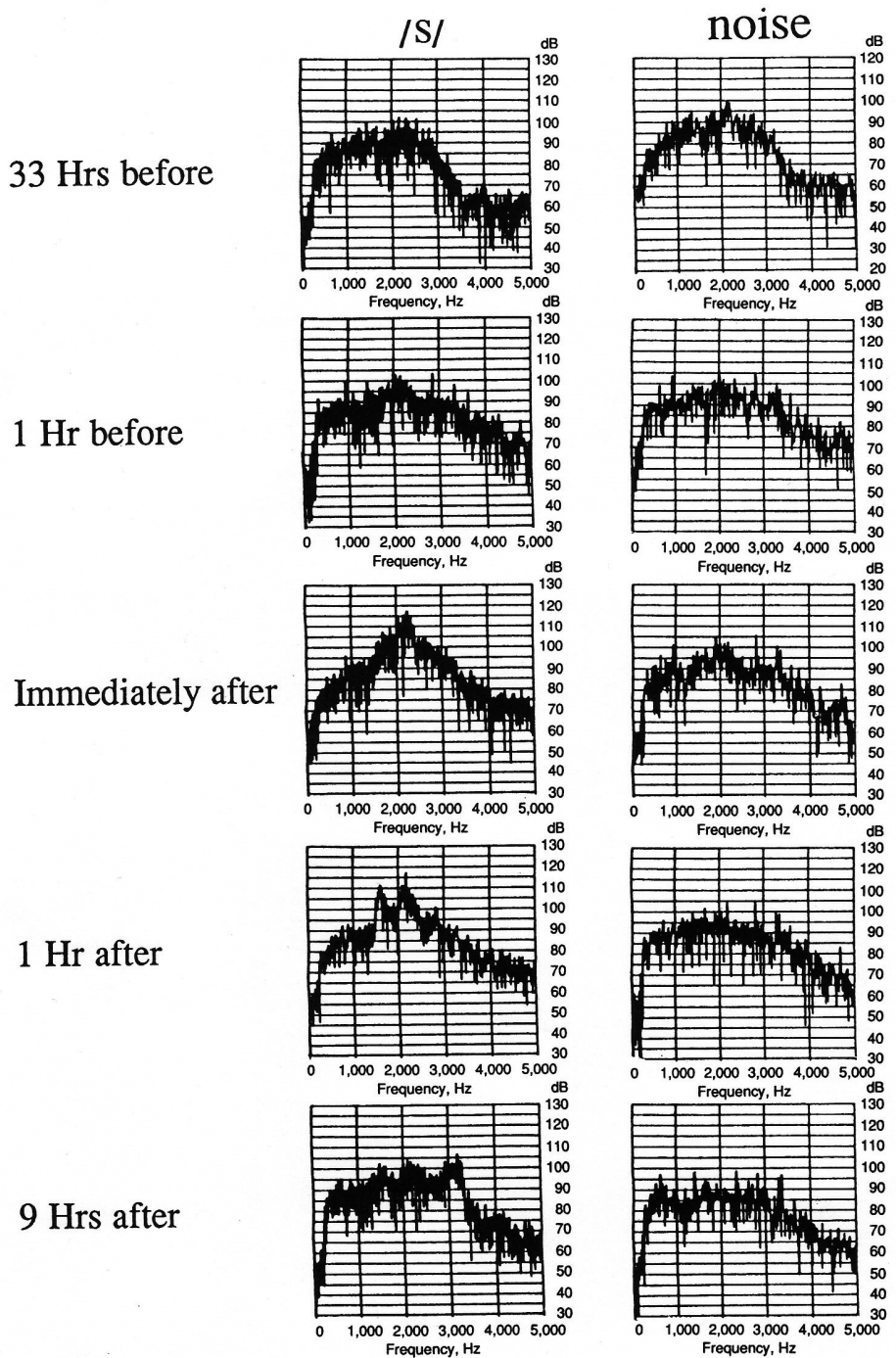


Fig. 2. Power spectra of /ʃ/ produced by Captain Hazelwood in the words *she's* and *shout* recorded 33 h before the accident. Each spectrum is paired with a spectrum of the background noise from a nearby open-mike pause.


/s/
noise
33 Hrs before
1 Hr before
Immediately after
1 Hr after
9 Hrs after
dB
130
120
110
100
90
80
70
60
50
40
30
20
0 1,000 2,000 3,000 4,000 5,000
Frequency, Hz

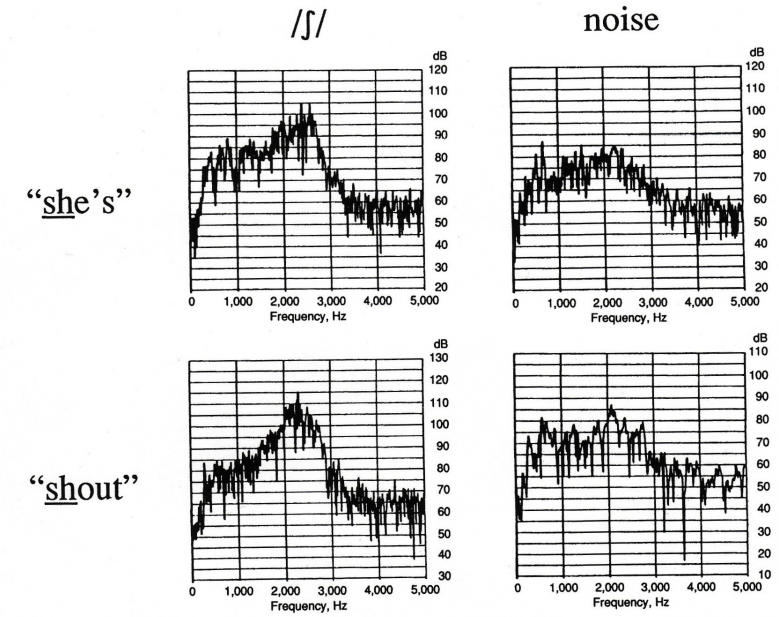
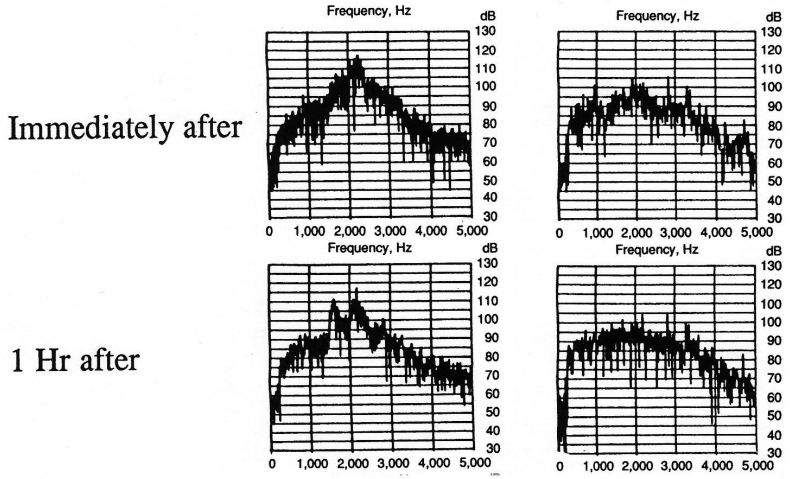
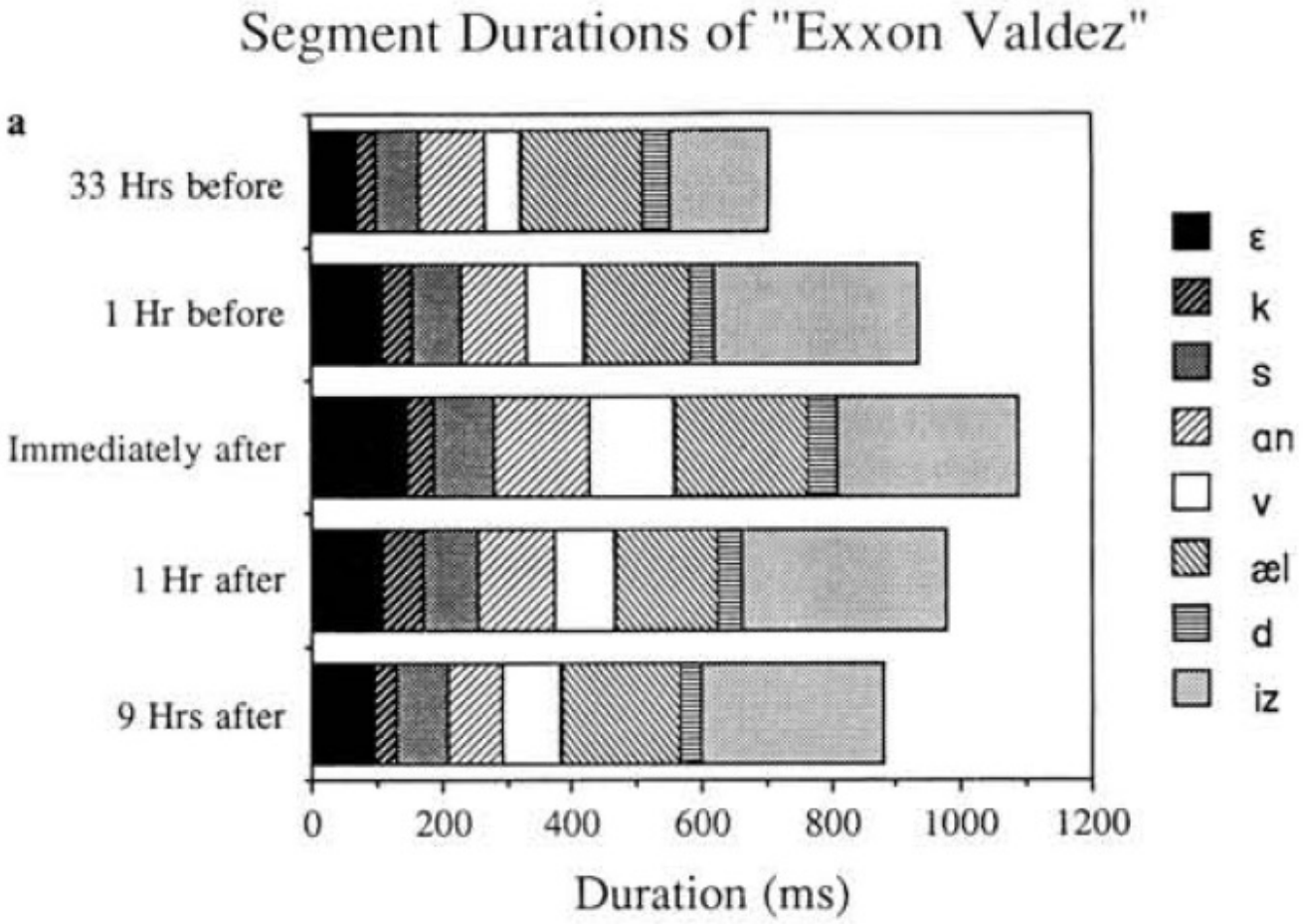


Fig. 2. Power spectra of /ʃ/ produced by Captain Hazelwood in the words *she's* and *shout* recorded 33 h before the accident. Each spectrum is paired with a spectrum of the background noise from a nearby open-mike pause.



Duration


Segment Durations of "Exxon Valdez"
a
33 Hrs before
1 Hr before
Immediately after
1 Hr after
9 Hrs after
0
200
400
600
800
1000
1200
Duration (ms)
ɛ
k
s
ɑn
v
æl
d
iz

Summary

Table 3. Summary of phenomena found in the analysis of the NTSB tape (numbers in parentheses indicate the time of recording)

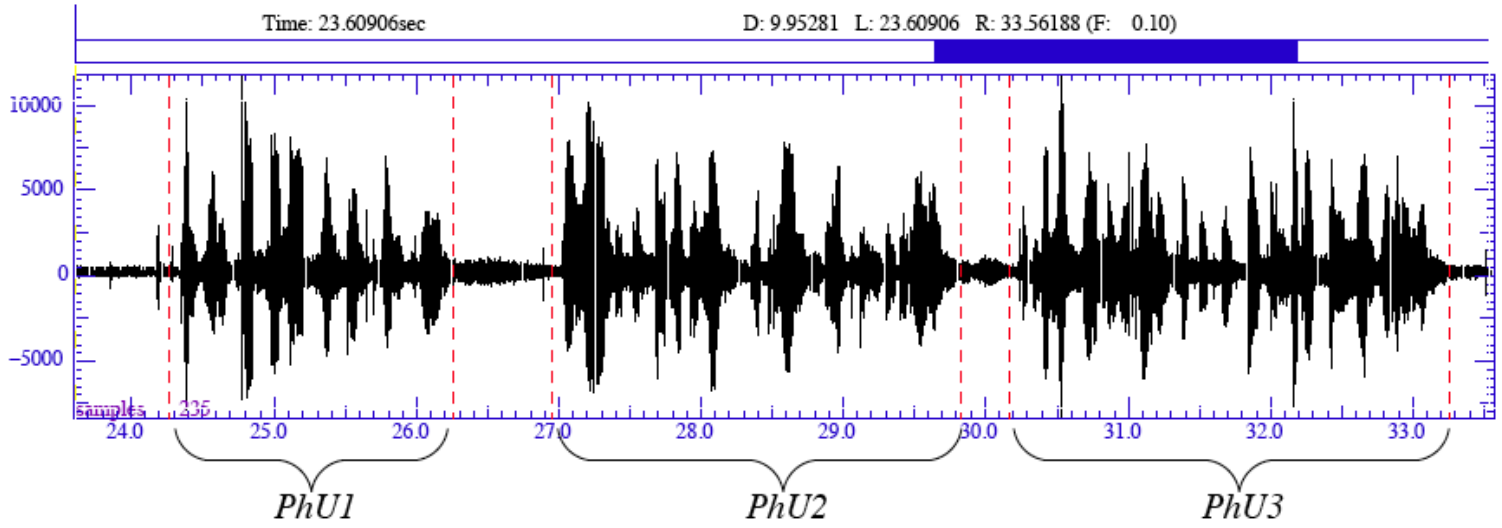
Gross effects	revisions (–1) Exxon Ba, uh Exxon Valdez (–1) departed disembarked (–1) I, we'll (–1) columbia gla, columbia bay
Segmental effects	misarticulation of /r/ and /l/ (0) northerly, little, drizzle, visibility (/s/ becomes /ʃ/ (fig. 3) final devoicing (e.g. /z/ → /s/) (–1,0,+1) Valdez → Valdes
Suprasegmental effects	reduced speaking rate (fig. 4, 5) mean change in pitch range (talker-dependent, fig. 6) increased F_0 jitter (fig. 6)

Problems

- If intoxicated speech, why wasn't s pronounced as sh 1 hour before?
- Other kinds of speaker state could cause drop in F0, slower speech, and disfluencies?
  - Stress, just having woken up, trauma….

Automatic Classification

- Use of prosodic speech characteristics
  for automated detection of alcohol intoxication
  Michael Levit, Richard Huber, Anton Batliner, Elmar Noeth
- Break utterance into phrases automatically, based on
  - fundamental frequency (where possible);
  - zero-crossing rate



Then use 4 classes of features

- Prosodic
  - F0 max, F0 min, energy max, energy min, pause length
- Duration of voiced regions, unvoiced regions, etc.
- Jitter and shimmer
  - jitter is variation in pitch
  - shimmer is variation in energy
- Average cepstrum and cepstral slope

Methods

- Alcoholized speech samples collected at the Police Academy of Hessen, Germany
  - 120 readings (87 minutes) of a fable
  - 33 male speakers
  - BAC between 0 and .24/mille

Alcohol Blood Level	0.0	< 0.4	< 0.8	< 1.2	< 1.6	< 2.0	< 2.4
Recordings	32	20	20	18	20	7	3

  - Binary task: above or below 0.8/mille
  - leave-one-out cross-validation
  - neural net classifier

Results of Levit et al.

- Used dev set to find best classifier
- This suggested two feature classes:
  - Prosodic features
  - Jitter/shimmer
- Results with this classifier
  - 62% phrase-accuracy
  - 69% for the whole speech sample
    - voting of the phrases

Alcohol Language Corpus

Florian Schiel et al 2009, 2010

- http://www.bas.uni-muenchen.de/forschung/Bas/BasALCeng.html
- 162 speakers (77 female, 85 male)
  - recorded in a car (sometimes with engine running)
  - command and control speech (“turn off the radio”)
  - spontaneous dialogue, monologue, question answering
  - read speech
  - counts of disfluencies, etc
- sample, drunk:
- sample, sober:

Automatic detection in ALC : Paralinguistic Challenge 2011

- Human: 66-72% (Schiel 2011, Ultes, Schmitt, Minker 2011)
- Machine: roughly 65%-70%
- Example features from winning system:

Bone, Daniel, Matthew Black, Ming Li, Angeliki Metallinou, Sungbok Lee, and Shrikanth S. Narayanan. 2011. Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors. In *INTERSPEECH*, pp. 3217-3220.

- Prosody (f0, duration, energy, jitter, shimmer)
- Spectral (MFCC, MFB log-energy, formants)
- Computed over whole utterance and small windows
- normalized phoneme duration
- iterative speaker normalizaton

Supervised ML as an analysis tool

1. Define your task / hypothesis (e.g. detecting alcohol intoxication in speech)
2. Collect/access data and annotations (y) for your task. Ensure data is representative for your problem
3. Define and test your modeling approach and inputs (x)
    1. Optimize to find x,f() to improve f(x)=y for your data
4. Analyze results, feature importance, model weights etc.

Poster session next week!

- **Tuesday May 31 5:30pm – 7:30pm**
  Hewlett Lawn
- Print your poster (9 printed slides works!)
  - We provide poster boards + easels
  - Set up your poster by 5:30
- Present your poster 5:30-7:30
  - 2 mins walk through for audience
  - At least one person attend poster during session