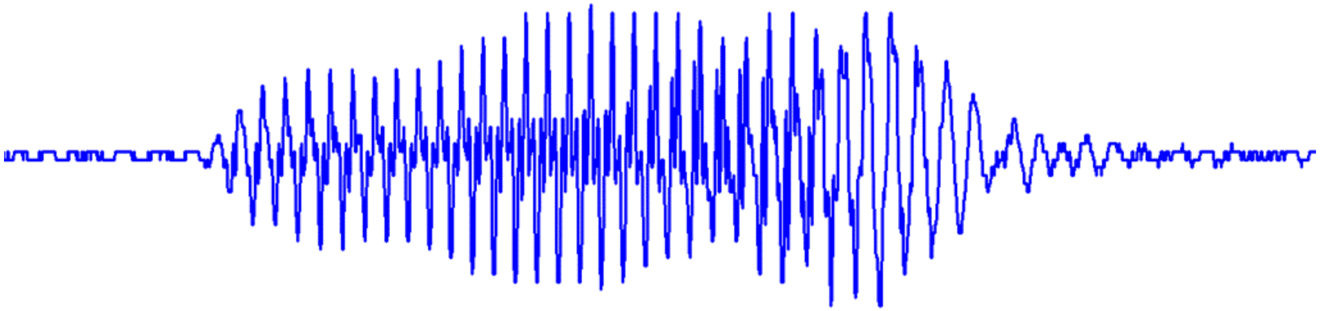
MONASH University

Engineering

Paralinguistic Speech Analysis

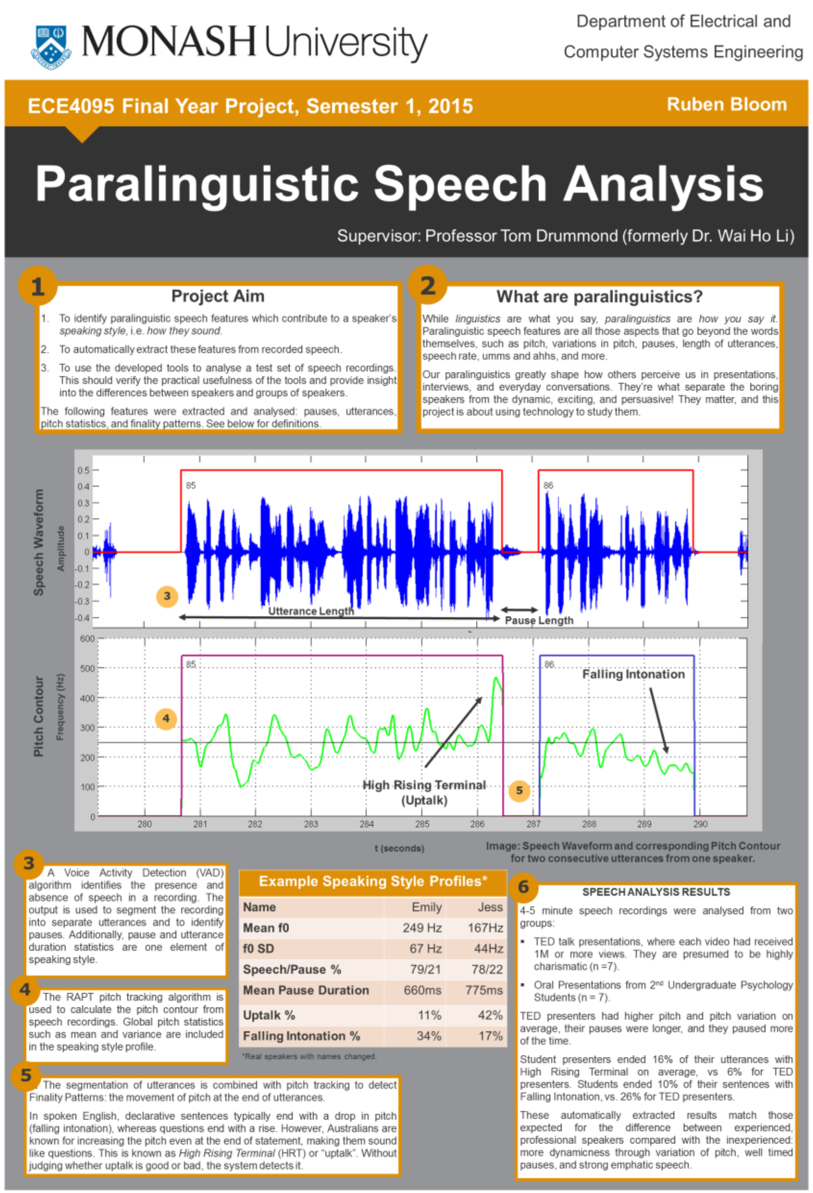
Ruben Bloom
ECE4095
30th June 2015



Significant Contributions

- My supervisors, Professor Tom Drummond and Dr. Wai Ho Li provided invaluable guidance and planning for this project.

Poster



MONASH University

Department of Electrical and
Computer Systems Engineering

ECE4095 Final Year Project, Semester 1, 2015

Ruben Bloom

Paralinguistic Speech Analysis

Supervisor: Professor Tom Drummond (formerly Dr. Wai Ho Li)

1 Project Aim

1. To identify paralinguistic speech features which contribute to a speaker's *speaking style*, i.e. *how they sound.*
2. To automatically extract these features from recorded speech.
3. To use the developed tools to analyse a test set of speech recordings. This should verify the practical usefulness of the tools and provide insight into the differences between speakers and groups of speakers.

The following features were extracted and analysed: pauses, utterances, pitch statistics, and finality patterns. See below for definitions.

2 What are paralinguistics?

While *linguistics* are what you say, *paralinguistics* are *how you say it*. Paralinguistic speech features are all those aspects that go beyond the words themselves, such as pitch, variations in pitch, pauses, length of utterances, speech rate, umms and ahhs, and more.

Our paralinguistics greatly shape how others perceive us in presentations, interviews, and everyday conversations. They're what separate the boring speakers from the dynamic, exciting, and persuasive! They matter, and this project is about using technology to study them.

Image: Speech Waveform and corresponding Pitch Contour for two consecutive utterances from one speaker.

3 A Voice Activity Detection (VAD) algorithm identifies the presence and absence of speech in a recording. The output is used to segment the recording into separate utterances and to identify pauses. Additionally, pause and utterance duration statistics are one element of speaking style.

4 The RAPT pitch tracking algorithm is used to calculate the pitch contour from speech recordings. Global pitch statistics such as mean and variance are included in the speaking style profile.

Example Speaking Style Profiles*

Name	Emily	Jess
Mean f0	249 Hz	167Hz
f0 SD	67 Hz	44Hz
Speech/Pause %	79/21	78/22
Mean Pause Duration	660ms	775ms
Uptalk %	11%	42%
Falling Intonation %	34%	17%

*Real speakers with names changed.

5 The segmentation of utterances is combined with pitch tracking to detect Finality Patterns: the movement of pitch at the end of utterances.

In spoken English, declarative sentences typically end with a drop in pitch (falling intonation), whereas questions end with a rise. However, Australians are known for increasing the pitch even at the end of statement, making them sound like questions. This is known as *High Rising Terminal* (HRT) or "uptalk". Without judging whether uptalk is good or bad, the system detects it.

6 SPEECH ANALYSIS RESULTS

4-5 minute speech recordings were analysed from two groups:

- TED talk presentations, where each video had received 1M or more views. They are presumed to be highly charismatic (n =7).
- Oral Presentations from 2nd Undergraduate Psychology Students (n = 7).

TED presenters had higher pitch and pitch variation on average, their pauses were longer, and they paused more of the time.

Student presenters ended 16% of their utterances with High Rising Terminal on average, vs 6% for TED presenters. Students ended 10% of their sentences with Falling Intonation, vs. 26% for TED presenters.

These automatically extracted results match those expected for the difference between experienced, professional speakers compared with the inexperienced: more dynamicness through variation of pitch, well timed pauses, and strong emphatic speech.

It's not just what you say, but how you say it.

Intro to Computational Paralinguistics

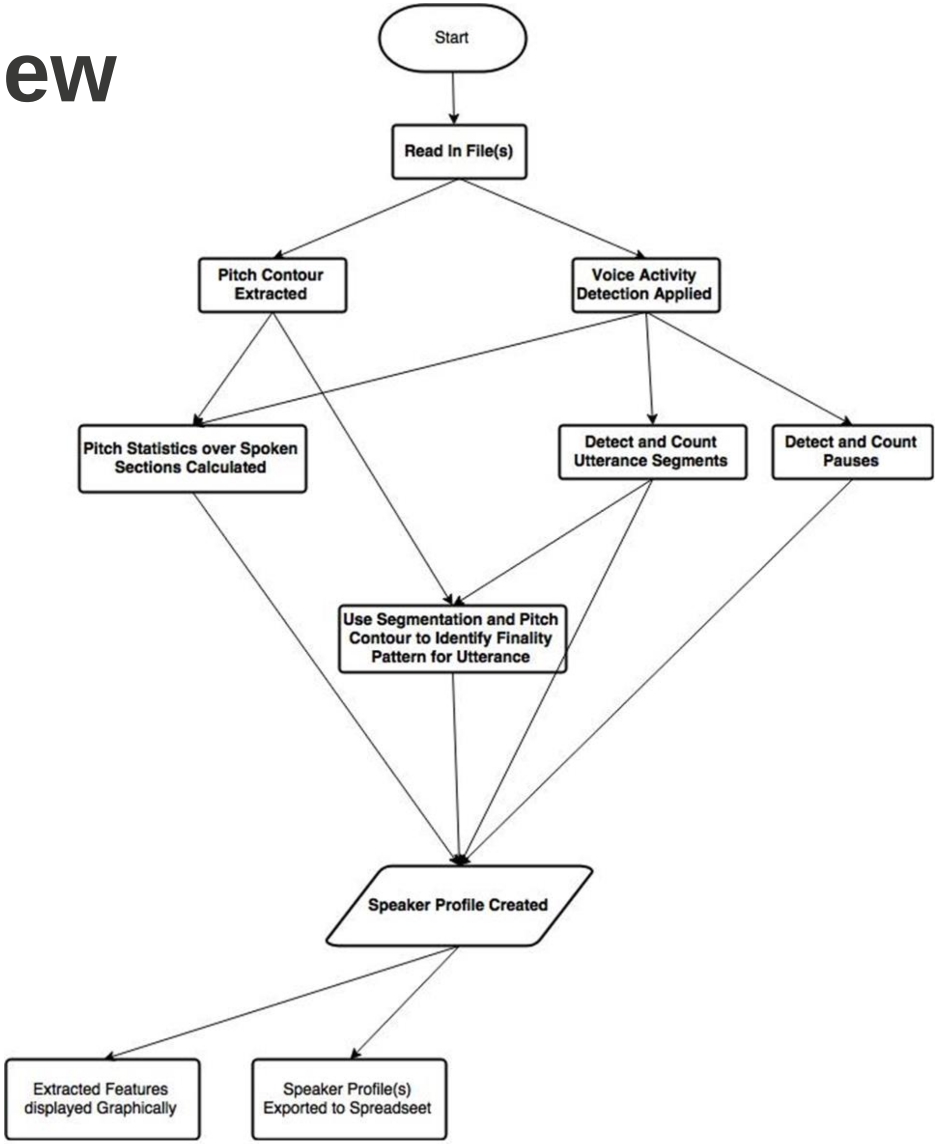
- WHAT *are* paralinguistics?
- All aspects of speech beyond the words!
  - Pitch
  - Pauses
  - Speech Rate
  - Pitch Patterns
  - Umms and Ahhs
- Can be extracted from speech waveform automatically!
- Automatic Speech Recognition for non-verbals!

The Goal: Speaking Style

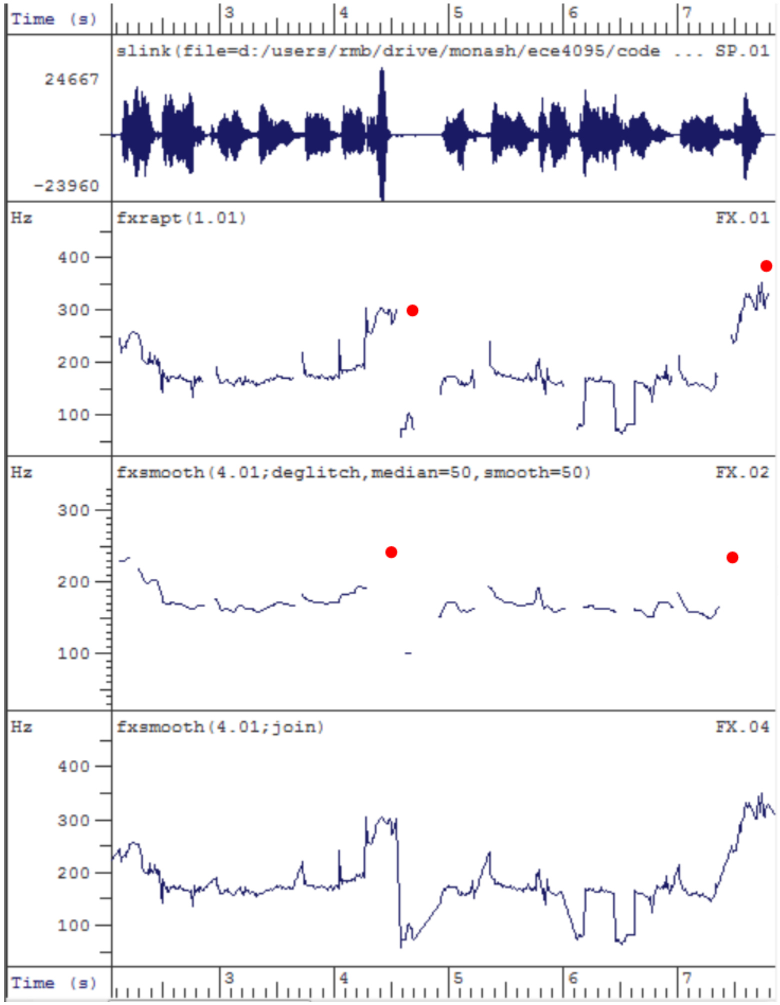
- IDENTIFY
- EXTRACT
- ANALYSE

for SCIENCE!!
for ENGINEERING!!!

System Overview


Start
Read In File(s)
Pitch Contour Extracted
Voice Activity Detection Applied
Pitch Statistics over Spoken Sections Calculated
Detect and Count Utterance Segments
Detect and Count Pauses
Use Segmentation and Pitch Contour to Identify Finality Pattern for Utterance
Speaker Profile Created
Extracted Features displayed Graphically
Speaker Profile(s) Exported to Spreadseet

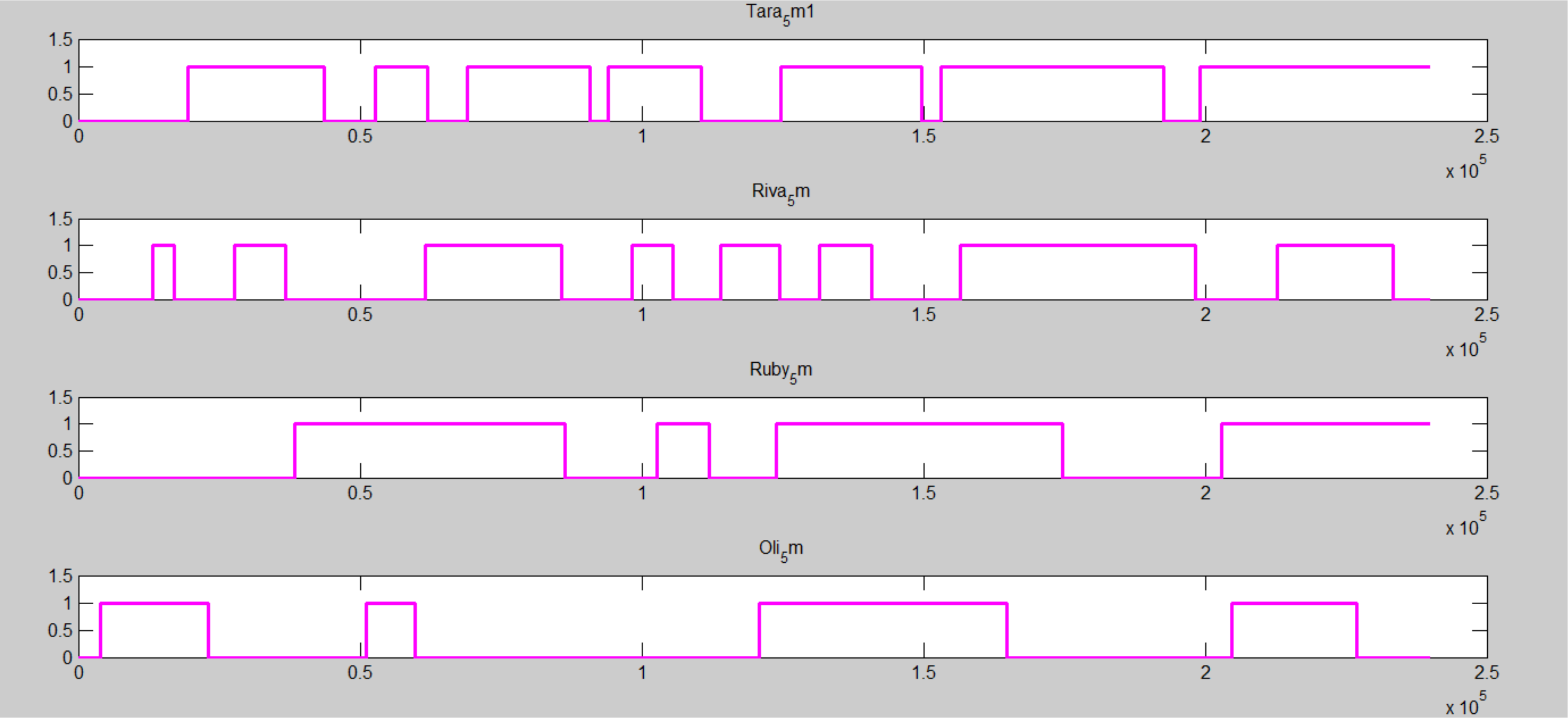
Pitch


Time (s)
3
4
5
6
7
slink(file=d:/users/rmb/drive/monash/ece4095/code ... SP.01
24667
-23960
Hz
fxrapt(1.01)
FX.01
400
300
200
100
Hz
fxsmooth(4.01;deglitch,median=50,smooth=50)
FX.02
300
200
100
Hz
fxsmooth(4.01;join)
FX.04
400
300
200
100
Time (s)
3
4
5
6
7

Pauses and Utterances

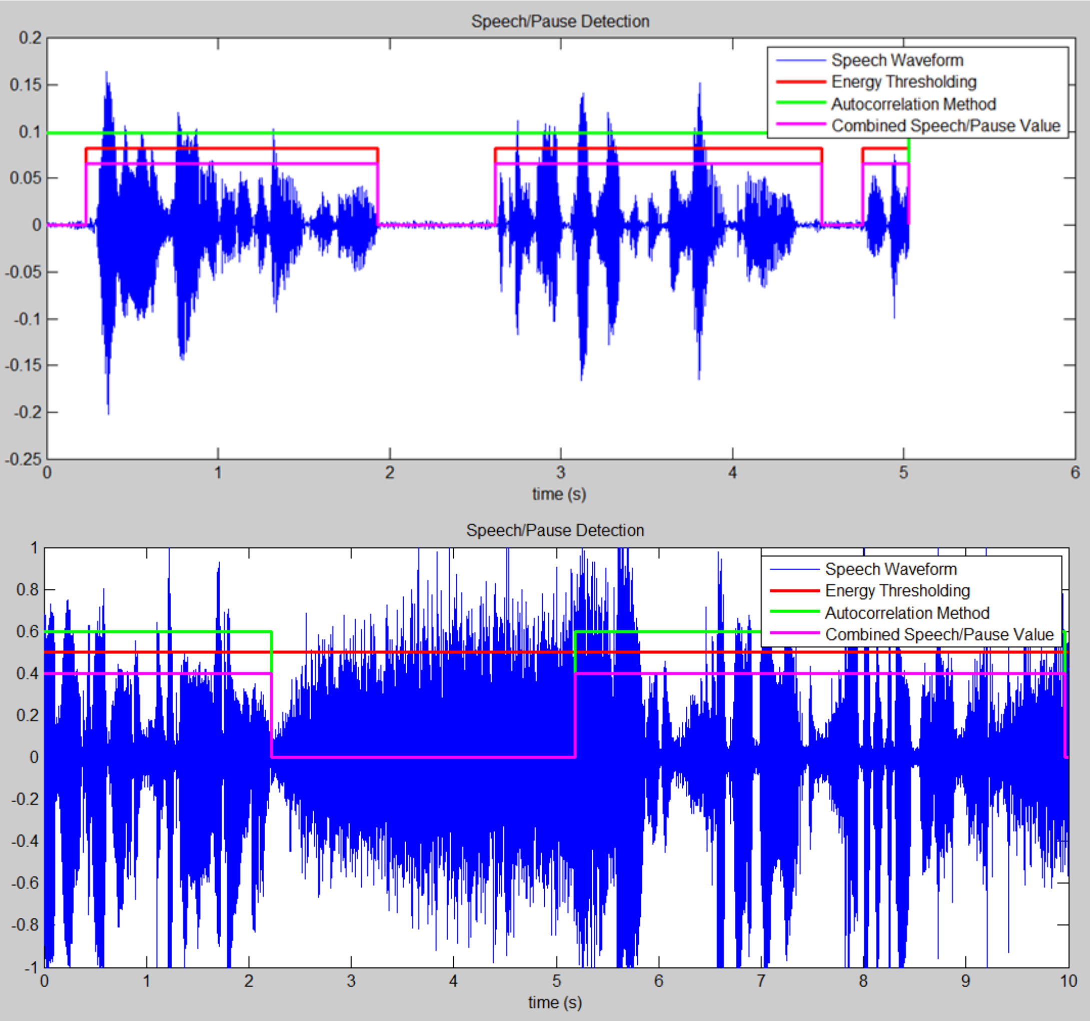
- Why. They. Matter.
- whytheymatter . . .

Pauses and Utterances



VAD: Detecting Pauses and Utterances

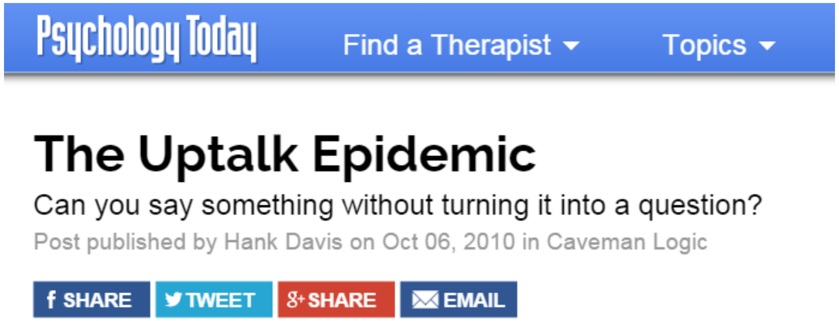
- VAD
- (Can do with ASR, but didn't)
- Three attempts



Finality Patterns

- Pitch movement at the end of utterance.
- English: Up, Down, Neutral.
- Up for statements: High Rising Terminal/Uptalk

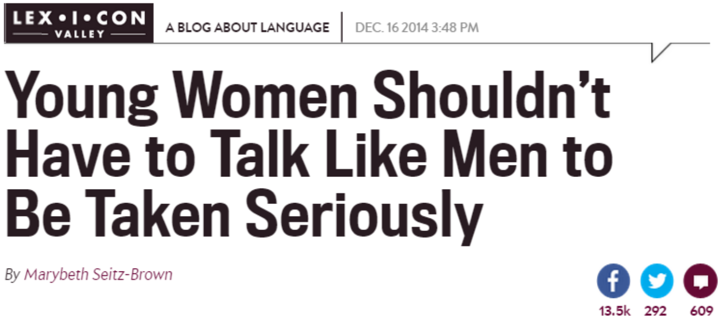
Uptalk: A Hot Topic



The Uptalk Epidemic

Can you say something without turning it into a question?

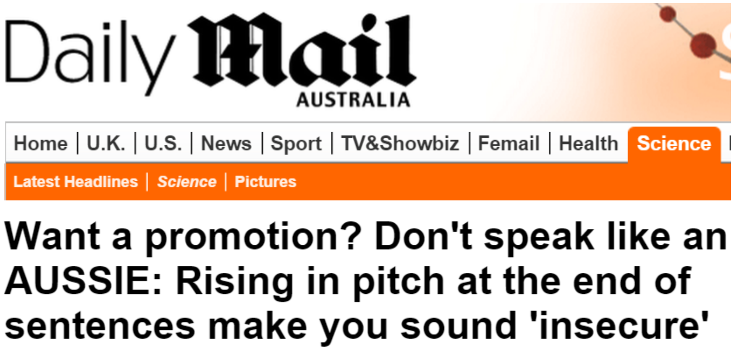
Post published by Hank Davis on Oct 06, 2010 in Caveman Logic



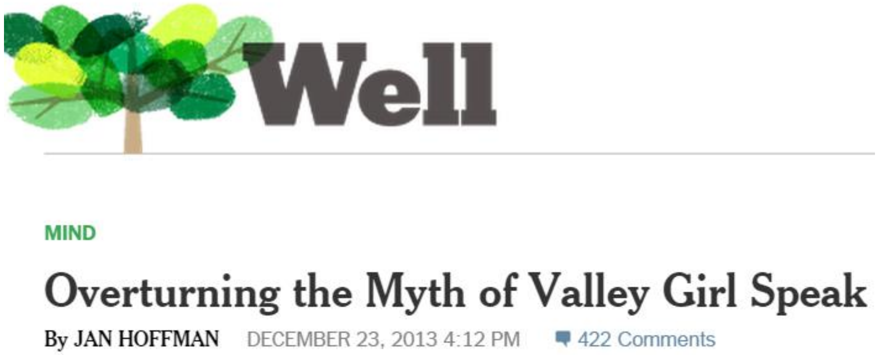
DEC. 16 2014 3:48 PM

Young Women Shouldn't Have to Talk Like Men to Be Taken Seriously

By Marybeth Seitz-Brown


Daily Mail
AUSTRALIA
Home | U.K. | U.S. | News | Sport | TV&Showbiz | Femail | Health | Science
Latest Headlines | Science | Pictures

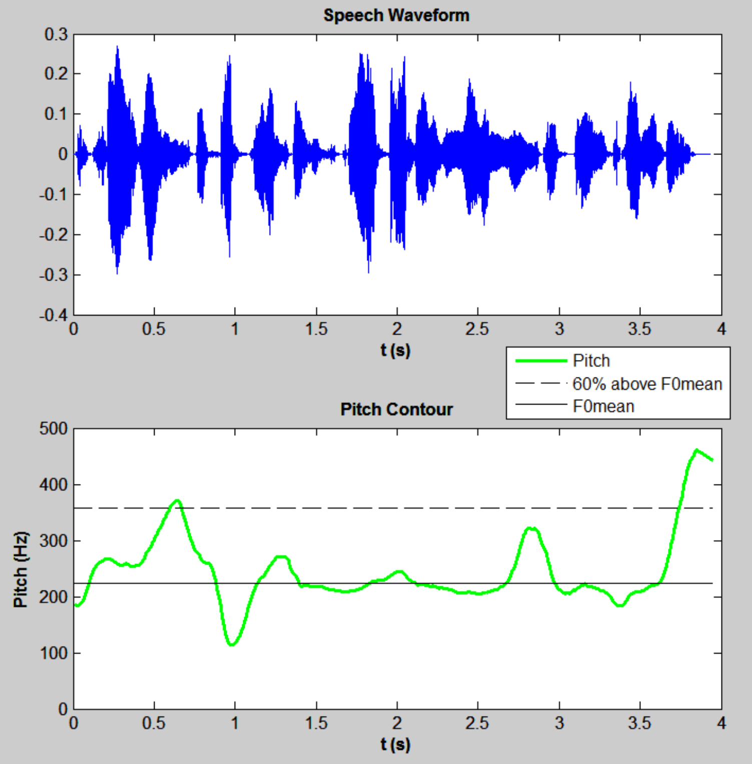
Want a promotion? Don't speak like an AUSSIE: Rising in pitch at the end of sentences make you sound 'insecure'

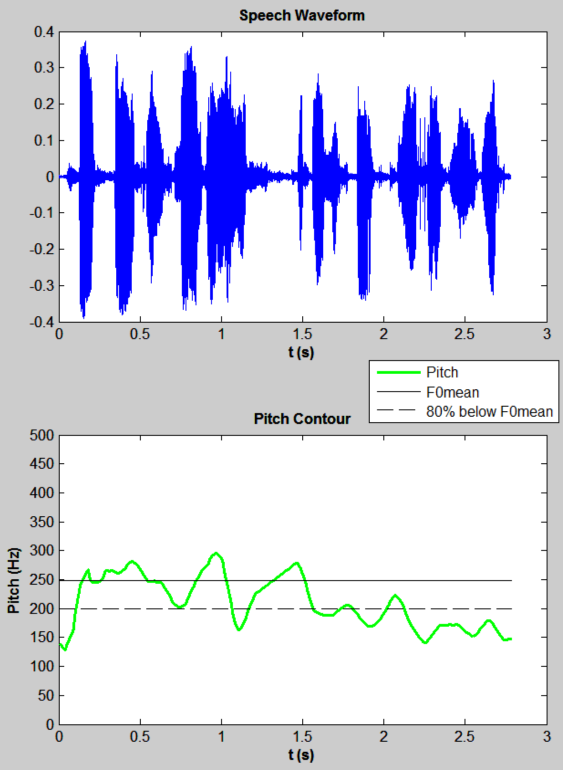


Overturning the Myth of Valley Girl Speak

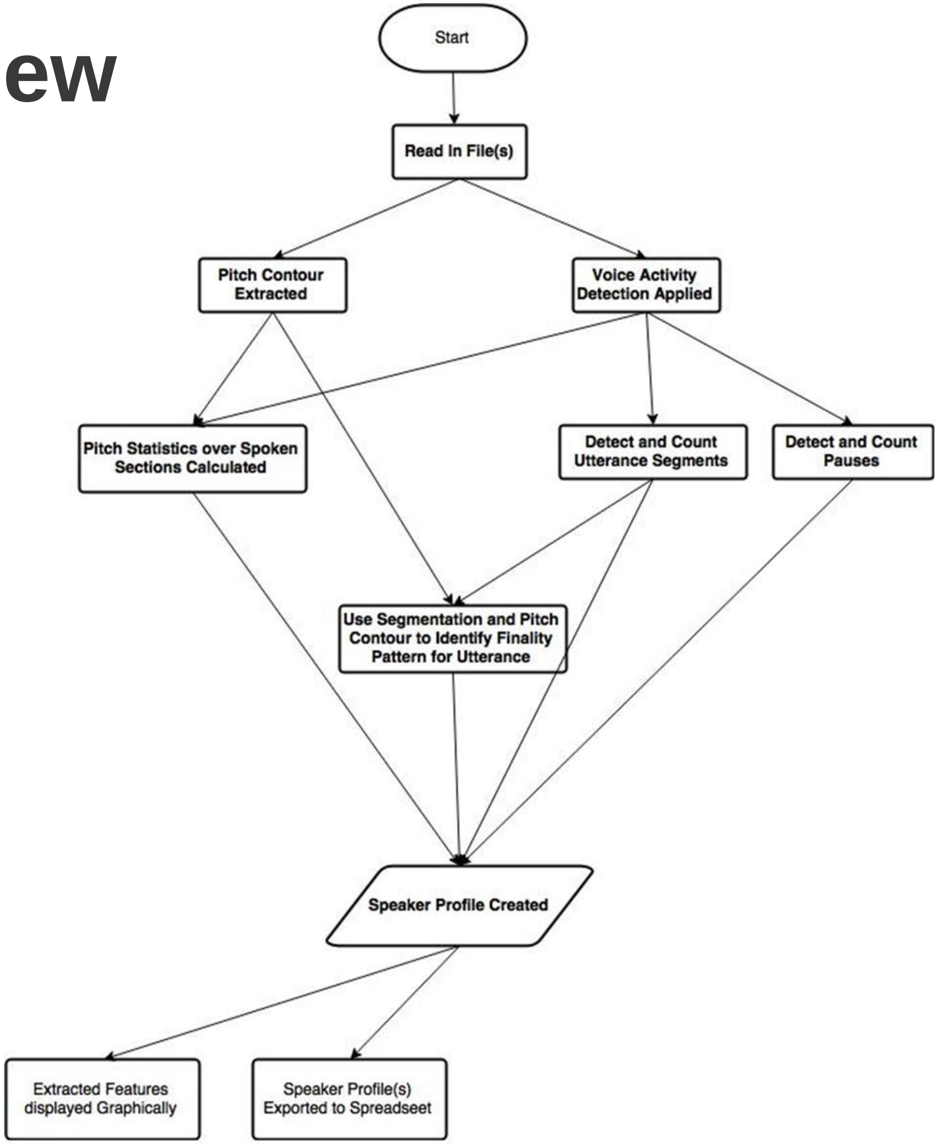
By JAN HOFFMAN DECEMBER 23, 2013 4:12 PM 422 Comments

Finality Patterns





System Overview


Start
Read In File(s)
Pitch Contour Extracted
Voice Activity Detection Applied
Pitch Statistics over Spoken Sections Calculated
Detect and Count Utterance Segments
Detect and Count Pauses
Use Segmentation and Pitch Contour to Identify Finality Pattern for Utterance
Speaker Profile Created
Extracted Features displayed Graphically
Speaker Profile(s) Exported to Spreadseet

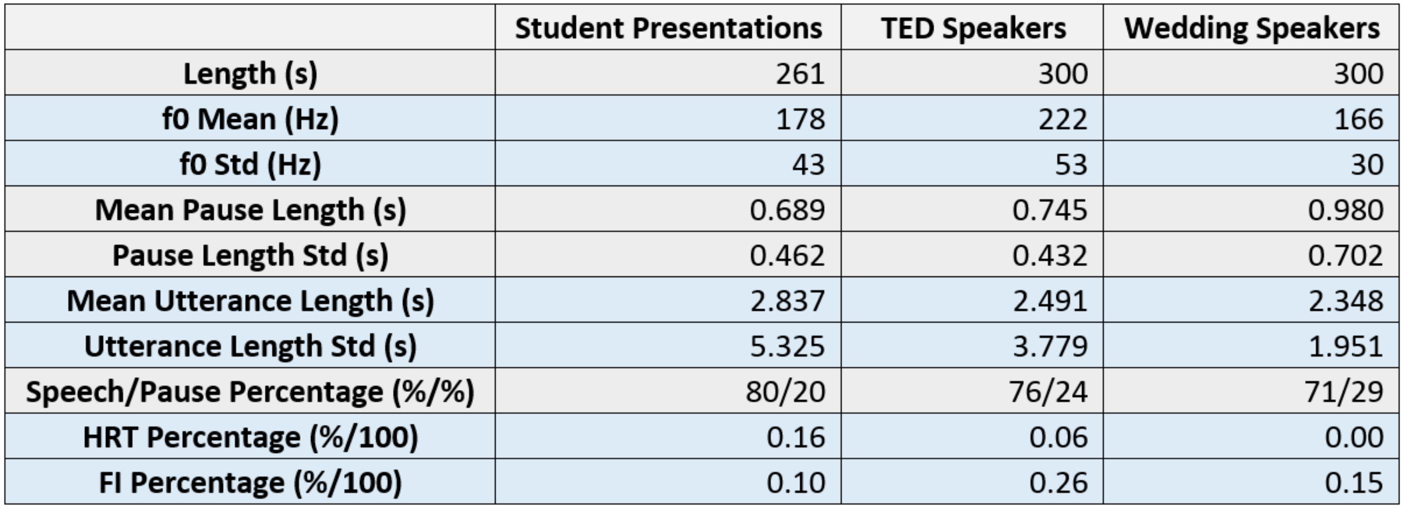
Speaking Style Comparisons

Three groups

1. Oral Presentations by 2nd Year Psychology Undergraduates (n = 7)
2. TED Talks with a minimum of 1M views each (n = 7)
3. Ceremonial wedding speeches (n = 6)

Results: Group Differences

- Differences in expected direction from previous studies! (Rosenburg and Hirschburg, Stangert, Strangert and Gustafson)
- TED Speakers have higher mean pitch and pitch variation, longer and more pauses, and higher percentage of utterances with Falling Intonation.
- Australian Psychology Undergrads have high levels of High Rising Terminal.

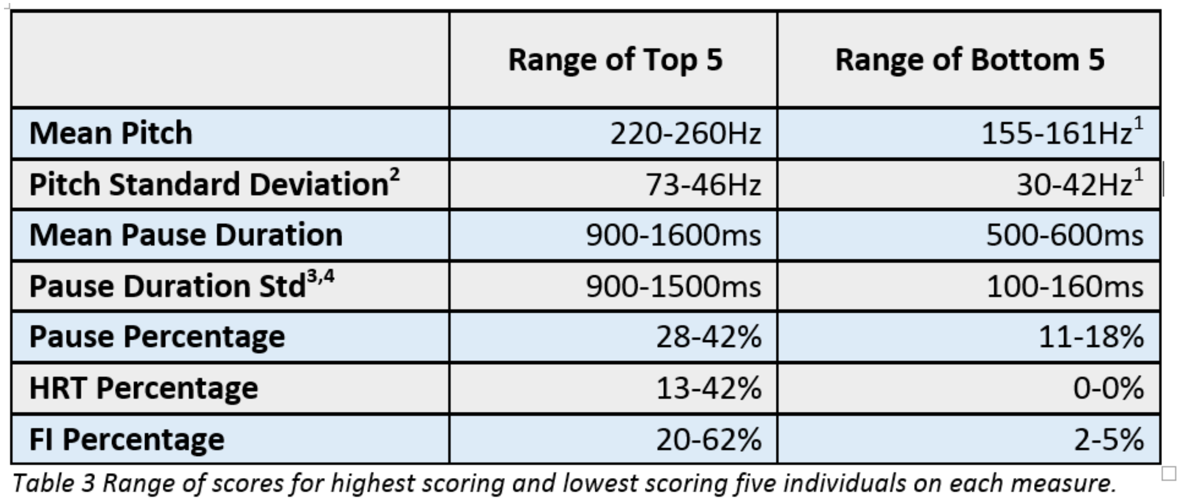


	Student Presentations	TED Speakers	Wedding Speakers
Length (s)	261	300	300
f0 Mean (Hz)	178	222	166
f0 Std (Hz)	43	53	30
Mean Pause Length (s)	0.689	0.745	0.980
Pause Length Std (s)	0.462	0.432	0.702
Mean Utterance Length (s)	2.837	2.491	2.348
Utterance Length Std (s)	5.325	3.779	1.951
Speech/Pause Percentage (%/%)	80/20	76/24	71/29
HRT Percentage (%/100)	0.16	0.06	0.00
FI Percentage (%/100)	0.10	0.26	0.15

Table 2 Mean of each speaker group for each variable listed.

Results: Individual Differences

- Dramatic differences!
- Occur across group boundaries: it is possible to be charismatic in many ways!
- Need more sophisticated measures to differentiate style for charisma.



	Range of Top 5	Range of Bottom 5
Mean Pitch	220-260Hz	155-161Hz[1]
Pitch Standard Deviation[2]	73-46Hz	30-42Hz[1]
Mean Pause Duration	900-1600ms	500-600ms
Pause Duration Std[3,4]	900-1500ms	100-160ms
Pause Percentage	28-42%	11-18%
HRT Percentage	13-42%	0-0%
FI Percentage	20-62%	2-5%

Table 3 Range of scores for highest scoring and lowest scoring five individuals on each measure.

Limitations

- Inaccuracies in segmentation have downstream effects.
- Small samples.
- Limited testing of feature extraction.
- No control of audio recording environment and resultant quality.

Outcomes

Overall Goal: Use computational paralinguistics to develop tools useful for the scientific and engineering analysis of speaking style.

- Speech features identified!
- Automatic extraction achieved!
- Tools for quantitatively identifying paralinguistic differences between speakers achieved!
- Sample analysis achieved!

Project changed from original specifications, but majority of requirements still met!

Future Directions

- Rigorous testing of extracted features.
- More speech features:
  - Speech Rate
  - Filled Pauses
  - Energy Variation (emphasis)
  - Whole pitch contours
- Improved Segmentation
- Better analysis of speech/pause rhythm, e.g. “frequency analysis”
- More speech types: interviews, political speeches, conversations
- Speaker Diarisation
- PCA on analysed data
- Combination with linguistic features from ASR for comprehensive speaker profile.

Graphical Output Demo & Spreadsheet