Class 10: Audio Measurement

Theme: Audio

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

Mathias Rask (mathiasrask@ps.au.dk)

Aarhus University

Why Audio?

Why should we care about audio data?

Why Audio?

Why should we care about audio data?

- Independent effect: Nonverbal speech contains information in and off itself (e.g. applause and jeering)

Why Audio?

Why should we care about audio data?

- Independent effect: Nonverbal speech contains information in and off itself (e.g. applause and jeering)
- Interaction effect: The meaning of words is fundamentally changed by how we deliver them (e.g. intonation)

Why Audio?

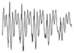
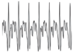
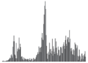
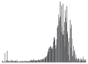
Why should we care about audio data?

- Independent effect: Nonverbal speech contains information in and off itself (e.g. applause and jeering)
- Interaction effect: The meaning of words is fundamentally changed by how we deliver them (e.g. intonation)

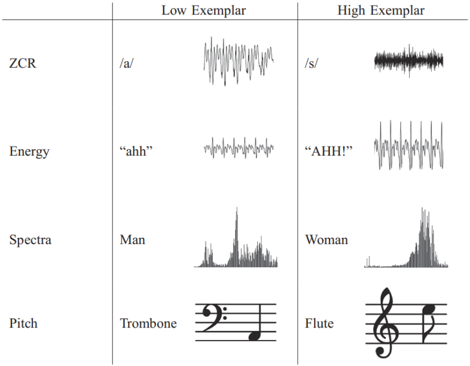
$\rightsquigarrow$ It's not only what you say, but it's also how you say it

Exploiting Variation in Audio Signals

Exploiting Variation in Audio Signals

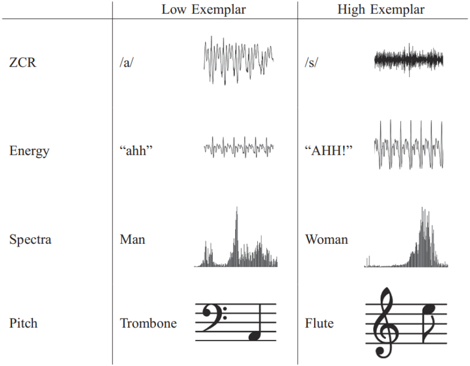
	Low Exemplar		High Exemplar	
ZCR	/a/		/s/	
Energy	“ahh”		“AHH!”	
Spectra	Man		Woman	
Pitch	Trombone		Flute	

Exploiting Variation in Audio Signals



Potential measures: Deception, sarcasm, skepticism, accent, attitude intensity, ...

Exploiting Variation in Audio Signals



Potential measures: Deception, sarcasm, skepticism, accent, attitude intensity, ...

How can you apply audio data outside this class?

Today's Menu

Measurement Approaches

- Theorizing
- Learning

Today's Menu

Measurement Approaches

- Theorizing
- Learning

Bias and Measurement Error

Today's Menu

- Measurement Approaches
  - Theorizing
  - Learning
- Bias and Measurement Error
- Lab

Table of Contents

Theorizing vs. Learning

Theorizing vs. Learning

Theorizing

Theorizing vs. Learning

Theorizing

- Linguistic, psychological, phonetic theory

Theorizing vs. Learning

Theorizing

- Linguistic, psychological, phonetic theory

Learning

Theorizing vs. Learning

Theorizing

- Linguistic, psychological, phonetic theory

Learning

- Hidden Markov Models and Neural Nets

Theorizing vs. Learning

Theorizing

- Linguistic, psychological, phonetic theory

Learning

- Hidden Markov Models and Neural Nets

Levels:

Theorizing vs. Learning

Theorizing

- Linguistic, psychological, phonetic theory

Learning

- Hidden Markov Models and Neural Nets

Levels:

- Semantically meaningful units (e.g. speeches, sentences, and words)

Theorizing vs. Learning

Theorizing

- Linguistic, psychological, phonetic theory

Learning

- Hidden Markov Models and Neural Nets

Levels:

- Semantically meaningful units (e.g. speeches, sentences, and words)
- Temporally fixed units

Theorizing vs. Learning

Theorizing

- Linguistic, psychological, phonetic theory

Learning

- Hidden Markov Models and Neural Nets

Levels:

- Semantically meaningful units (e.g. speeches, sentences, and words)
- Temporally fixed units

$\rightsquigarrow$ How do you decide upon the ‘right’ level of analysis?

Table of Contents

Theory-Driven Measures

Theory-Driven Measures

Theory-driven audio research is largely built upon the firmly established link between:

$$\text{emotional arousal} \rightsquigarrow \text{pitch}$$

Theory-Driven Measures

Theory-driven audio research is largely built upon the firmly established link between:

$$\text{emotional arousal} \rightsquigarrow \text{pitch}$$

- *Political science*: Emotional arousal is a distinct dimension of affect/emotions that carries information about political behavior

Theory-Driven Measures

Theory-driven audio research is largely built upon the firmly established link between:

$$\text{emotional arousal} \rightsquigarrow \text{pitch}$$

- *Political science*: Emotional arousal is a distinct dimension of affect/emotions that carries information about political behavior
- *Psychology*: Variation in pitch is consistently linked to a speaker's level of emotional activation

Theory-Driven Measures

Theory-driven audio research is largely built upon the firmly established link between:

$$\text{emotional arousal} \rightsquigarrow \text{pitch}$$

- *Political science*: Emotional arousal is a distinct dimension of affect/emotions that carries information about political behavior
- *Psychology*: Variation in pitch is consistently linked to a speaker's level of emotional activation

$\rightsquigarrow$ Pitch carries politically relevant information

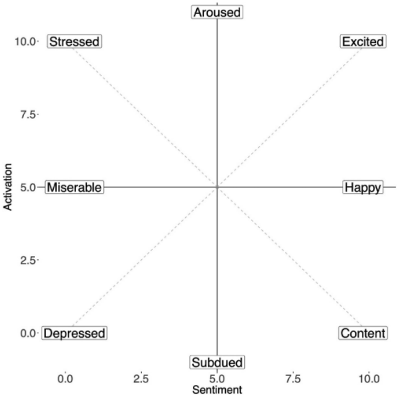
Pitch and Emotional Arousal

Pitch and Emotional Arousal

The link between pitch and emotional arousal is based on a continuous model of emotions:

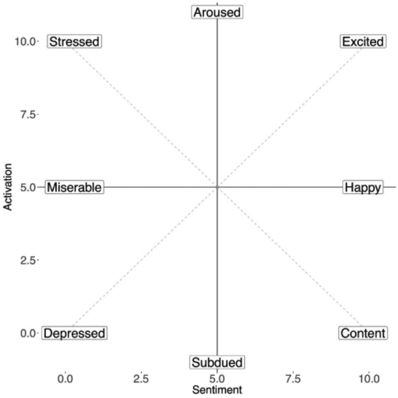
Pitch and Emotional Arousal

The link between pitch and emotional arousal is based on a continuous model of emotions:



Pitch and Emotional Arousal

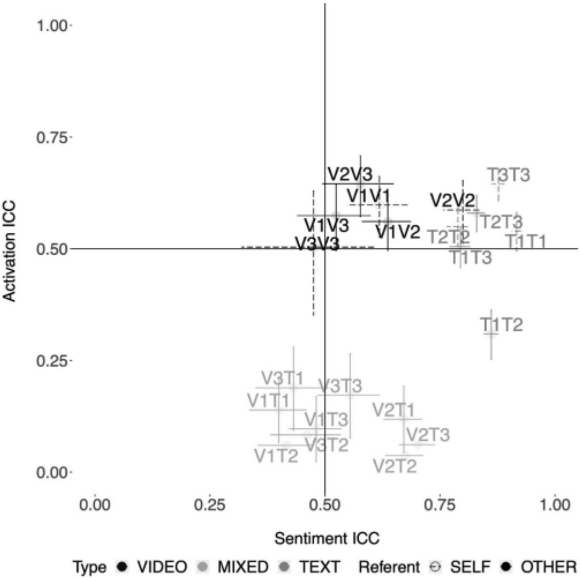
The link between pitch and emotional arousal is based on a continuous model of emotions:



⇝ pitch is indicative of different behaviors in different contexts

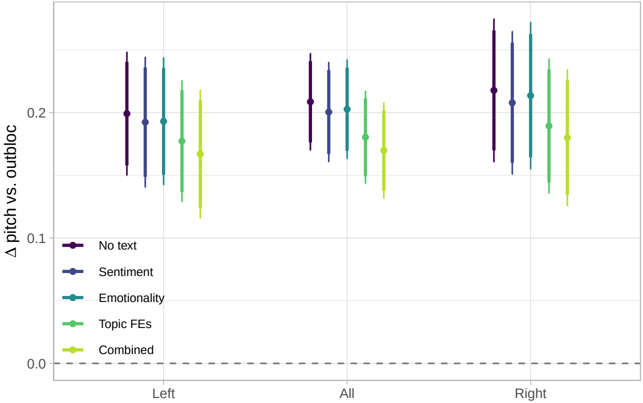
Conveying Emotions (Cochrane *et al.*, 2022)

Conveying Emotions (Cochrane *et al.*, 2022)

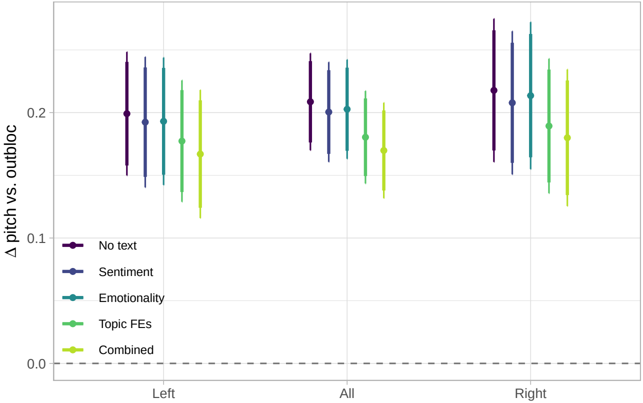


Nonverbal Signals of Partisan Conflict and Polarization (Rask and Hjorth, 2023)

Nonverbal Signals of Partisan Conflict and Polarization (Rask and Hjorth, 2023)



Nonverbal Signals of Partisan Conflict and Polarization (Rask and Hjorth, 2023)



⇝ Pitch is consistently higher when talking to out-partisans

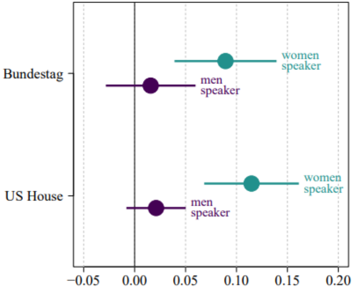
⇝ Indicates a nonverbal dimension of polarization

Nonverbal Signals of Issue Commitment (Rittmann, 2023)

Nonverbal Signals of Issue Commitment (Rittmann, 2023)

	US House	Bundestag
"Women" mentioned	0.021	0.016
	(0.015)	(0.022)
"Women" mentioned × Women Speaker	0.094*	0.073*
	(0.028)	(0.034)
R^2	0.000	0.000
Adj. R^2	−0.008	−0.035
Num. obs.	71198	33489

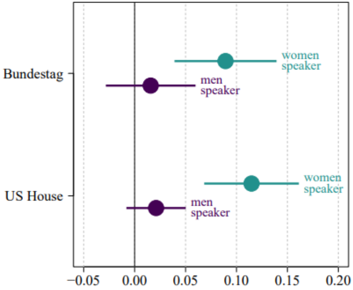
Legislator Fixed-Effects Models. $^*p < 0.05$.



Nonverbal Signals of Issue Commitment (Rittmann, 2023)

	US House	Bundestag
"Women" mentioned	0.021	0.016
	(0.015)	(0.022)
"Women" mentioned	0.094*	0.073*
× Women Speaker	(0.028)	(0.034)
R^2	0.000	0.000
Adj. R^2	−0.008	−0.035
Num. obs.	71198	33489

Legislator Fixed-Effects Models. $^*p < 0.05$.



- ⇝ Female politicians speak with a higher pitch when talking about women
- ⇝ Indicates a nonverbal dimension of political representation

Exercise

1. What are the implications of standardization?
2. The papers compute the pitch at the speech level. Is that meaningful? Why, why not? Contrast it with what we do when using text data.

Table of Contents

Learning From Audio

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

- Input X

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

- Input X
- Labels y

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

- Input X
- Labels y
- Train model h that learns the function $f : X \to y$

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

- Input X
- Labels y
- Train model h that learns the function $f : X \to y$
- Apply h out-of-sample to label new samples

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

- Input X
- Labels y
- Train model h that learns the function $f : X \to y$
- Apply h out-of-sample to label new samples

Two approaches:

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

- Input X
- Labels y
- Train model h that learns the function $f : X \rightarrow y$
- Apply h out-of-sample to label new samples

Two approaches:

- Machine learning: Hard-coding features

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

- Input X
- Labels y
- Train model h that learns the function $f : X \to y$
- Apply h out-of-sample to label new samples

Two approaches:

- Machine learning: Hard-coding features
- Deep learning: Learning features

Learning From Audio

The alternative to the theory-driven (rule-based?) approach is *learning from data*

- Input X
- Labels y
- Train model h that learns the function $f : X \to y$
- Apply h out-of-sample to label new samples

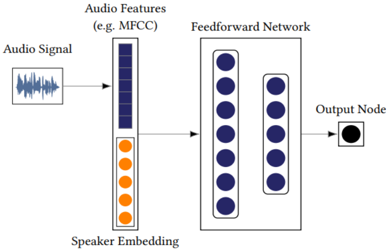
Two approaches:

- Machine learning: Hard-coding features
- Deep learning: Learning features

$\rightsquigarrow$ Works like a classical ML/DL pipeline, but the input differs

Exercise

Theorizing



Discuss the differences/similarities between the approaches used by Rheault and Borwein (2019) and Knox and Lucas (2021). Which approach(es) are used?

Learning

Data: Audio features ($\boldsymbol{X}^{\mathcal{C}}, \boldsymbol{X}^{\mathcal{T}}$), static metadata for primary corpus ($\boldsymbol{W}^{\text{stat},\mathcal{C}}$)
Result: Auditory parameters Θ, conversational flow parameters ζ
Procedure:
1. *Define problem.*
Analyst determines tones of interest and rubric for human coding. Human-coded tone labels are obtained for training set ($\boldsymbol{S}^{\mathcal{T}}$).
2. *Fit auditory parameters (Θ) by maximizing partial likelihood on training set ($\mathcal{T}$).*

```
for speech mode m in 1,...,M do
    Subset to training utterances labeled as tone m.
    while not converged do
        for utterance u in T and moment t in {1,...,T_u} do
            for sound k in 1,...,K do
                Compute emission probability of sound (m,k)
                generating audio (X_{u,t}).
            end
        end
        Predict sound being pronounced at each moment (R_{u,t}).
        Update cadence (usage patterns of constituent sounds, Γ^m).
        for sound k in 1,...,K do
            Update audio profile of sound k (μ^{m,k}, Σ^{m,k}).
        end
    end
end
```

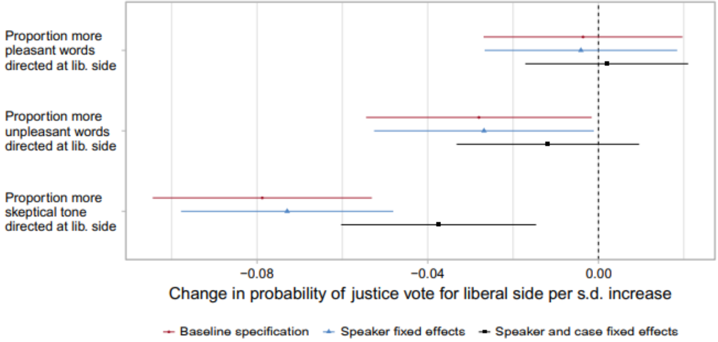
3. *Fit conversational flow parameters (ζ) using primary corpus ($\mathcal{C}$), conditional on Θ.*

```
for utterance u in C do
    for speech mode m in 1,...,M do
        Compute corrected emission probability of speech mode m
        generating utterance audio data (X_u), ignoring context.
    end
end
while not converged do
    Predict expected mode of speech for each utterance (S_u).
    Compute expected conversation context for each utterance (W_u).
    Update flow-of-speech parameters (ζ).
end
```

HMM Classifier of Skepticism

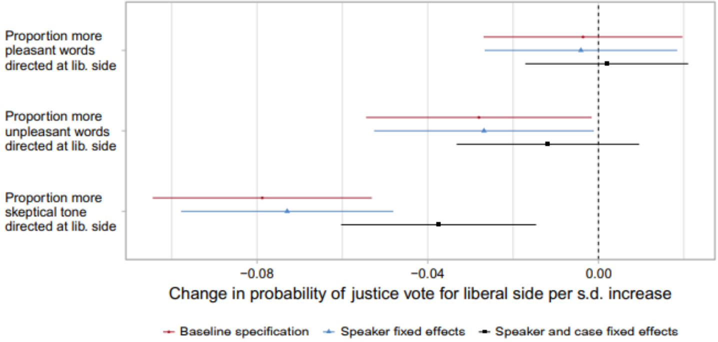
HMM Classifier of Skepticism

FIGURE 4. Predicting Justice Votes with Directed Skepticism and Directed Affective Language



HMM Classifier of Skepticism

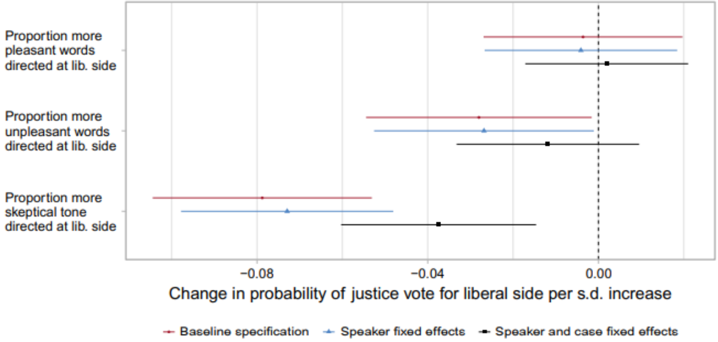
FIGURE 4. Predicting Justice Votes with Directed Skepticism and Directed Affective Language



What does the figure indicate?

HMM Classifier of Skepticism

FIGURE 4. Predicting Justice Votes with Directed Skepticism and Directed Affective Language

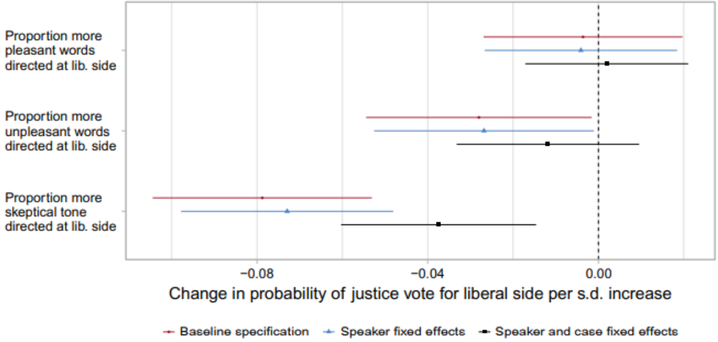


What does the figure indicate?

1. That skepticism is more accurately conveyed in the vocal tone

HMM Classifier of Skepticism

FIGURE 4. Predicting Justice Votes with Directed Skepticism and Directed Affective Language



What does the figure indicate?

1. That skepticism is more accurately conveyed in the vocal tone
2. That the text and audio-based measures tap into different underlying concepts

NN Classifier of Anxiety and Arousal

NN Classifier of Anxiety and Arousal

Text Modality

Emotion	Accuracy(%)	Modal Category(%)	PRE (%)
Valence	77.2	62.5	39.3
Activation	64.5	67.3	−8.7
Anxiety	66.1	68.4	−7.7

NN Classifier of Anxiety and Arousal

Text Modality

Emotion	Accuracy(%)	Modal Category(%)	PRE (%)
Valence	77.2	62.5	39.3
Activation	64.5	67.3	−8.7
Anxiety	66.1	68.4	−7.7

Audio Modality

Emotion	Model	Accuracy (%)	Modal Category (%)	PRE (%)
Valence	Pooled	61.5	60.1	3.6
	Speaker embeddings	73.4	60.1	34.3
Activation	Pooled	69.4	63.9	15.1
	Speaker embeddings	77.8	64.5	37.3
Anxiety	Pooled	71.1	67.6	10.7
	Speaker embeddings	80.3	62.7	47.2

NN Classifier of Anxiety and Arousal

Text Modality

Emotion	Accuracy(%)	Modal Category(%)	PRE (%)
Valence	77.2	62.5	39.3
Activation	64.5	67.3	−8.7
Anxiety	66.1	68.4	−7.7

Audio Modality

Emotion	Model	Accuracy (%)	Modal Category (%)	PRE (%)
Valence	Pooled	61.5	60.1	3.6
	Speaker embeddings	73.4	60.1	34.3
Activation	Pooled	69.4	63.9	15.1
	Speaker embeddings	77.8	64.5	37.3
Anxiety	Pooled	71.1	67.6	10.7
	Speaker embeddings	80.3	62.7	47.2

Audio Features

Table 2: Audio Features by Emotional Category

	Feature	Activated	Calm	t	p-value
Activation	Energy	0.020	0.018	2.918	0.004
	Pitch (Reaper)	189.527	154.971	21.038	<0.001
	Pitch Std. Dev. (Reaper)	54.229	41.889	15.461	<0.001
	Pitch (Praat)	202.117	165.16	19.277	<0.001
	Pitch Std. Dev. (Praat)	43.018	34.102	15.385	<0.001
	Speech Rate	3.935	3.833	3.587	<0.001

	Feature	Anxious	Non-Anxious	t	p-value
Anxiety	Energy	0.017	0.018	−2.182	0.029
	Pitch (Reaper)	174.705	154.854	9.290	<0.001
	Pitch Std. Dev. (Reaper)	44.492	35.099	11.986	<0.001
	Pitch (Praat)	186.088	160.628	11.823	<0.001
	Pitch Std. Dev. (Praat)	42.342	33.501	12.401	<0.001
	Speech Rate	3.784	3.750	1.070	0.285

Summary statistics and mean difference tests for a subset of audio features computed using the Parselmouth and pyAudioAnalysis libraries, as well as pitch estimates obtained with the Reaper algorithm. The dataset comprises 2,982 speeches annotated for activation, and 2,057 for anxiety.

NN Classifier of Anxiety and Arousal

Text Modality

Emotion	Accuracy(%)	Modal Category(%)	PRE (%)
Valence	77.2	62.5	39.3
Activation	64.5	67.3	−8.7
Anxiety	66.1	68.4	−7.7

Audio Modality

Emotion	Model	Accuracy (%)	Modal Category (%)	PRE (%)
Valence	Pooled	61.5	60.1	3.6
	Speaker embeddings	73.4	60.1	34.3
Activation	Pooled	69.4	63.9	15.1
	Speaker embeddings	77.8	64.5	37.3
Anxiety	Pooled	71.1	67.6	10.7
	Speaker embeddings	80.3	62.7	47.2

Audio Features

Table 2: Audio Features by Emotional Category

	Feature	Activated	Calm	t	p-value
Activation	Energy	0.020	0.018	2.918	0.004
	Pitch (Reaper)	189.527	154.971	21.038	<0.001
	Pitch Std. Dev. (Reaper)	54.229	41.889	15.461	<0.001
	Pitch (Praat)	202.117	165.16	19.277	<0.001
	Pitch Std. Dev. (Praat)	43.018	34.102	15.385	<0.001
	Speech Rate	3.935	3.833	3.587	<0.001
	Feature	**Anxious**	**Non-Anxious**	t	p-value
Anxiety	Energy	0.017	0.018	−2.182	0.029
	Pitch (Reaper)	174.705	154.854	9.290	<0.001
	Pitch Std. Dev. (Reaper)	44.492	35.099	11.986	<0.001
	Pitch (Praat)	186.088	160.628	11.823	<0.001
	Pitch Std. Dev. (Praat)	42.342	33.501	12.401	<0.001
	Speech Rate	3.784	3.750	1.070	0.285

Summary statistics and mean difference tests for a subset of audio features computed using the Parselmouth and pyAudioAnalysis libraries, as well as pitch estimates obtained with the Reaper algorithm. The dataset comprises 2,982 speeches annotated for activation, and 2,057 for anxiety.

Why do the speaker embeddings improve the results?

Table of Contents

Sources of Bias and Measurement Error

Sources of Bias and Measurement Error

Audio signals are susceptible to factors outside our variation of interest:

Sources of Bias and Measurement Error

Audio signals are susceptible to factors outside our variation of interest:

- Recording heterogeneity: microphone quality and distance, room acoustics, and A/D quality

Sources of Bias and Measurement Error

Audio signals are susceptible to factors outside our variation of interest:

- Recording heterogeneity: microphone quality and distance, room acoustics, and A/D quality
- Speaker heterogeneity: vocal features are speaker-dependent

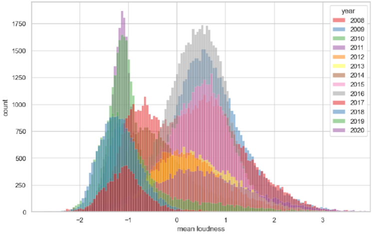
Exercise

1. What are the implications of recording and speaker heterogeneity when for instance training a classifier?
2. What are potential solutions, if any?

Recording Heterogeneity

Recording Heterogeneity

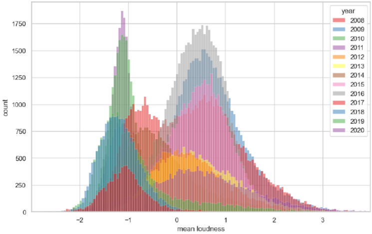
Speaker-normalized loudness:



⇝ Year-specific distributions

Recording Heterogeneity

Speaker-normalized loudness:



⇝ Year-specific distributions

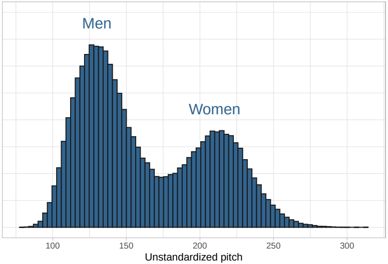
Temporal dependence:

feature	adj. r-square
mean f_0	0.006
voiced per sec	0.015
std f_0	0.064
mean MFCC1	0.120
mean loudness	0.404

Speaker Heterogeneity

Speaker Heterogeneity

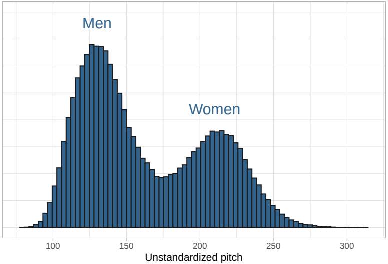
Raw $F0$:



–

Speaker Heterogeneity

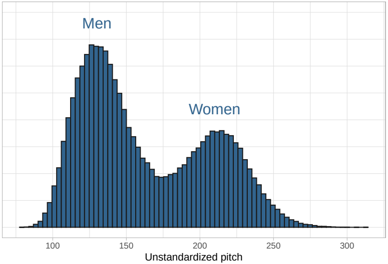
Raw $F0$:



- $\rightsquigarrow$ Bimodal distribution

Speaker Heterogeneity

Raw $F0$:

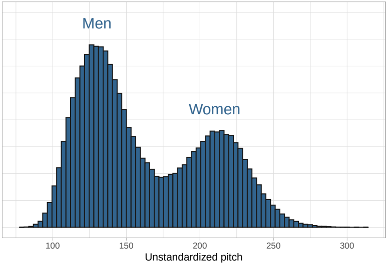


- $\rightsquigarrow$ Bimodal distribution

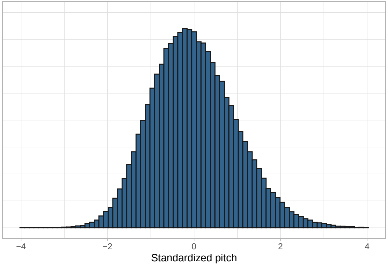
Speaker-standardized $F0$

Speaker Heterogeneity

Raw $F0$:



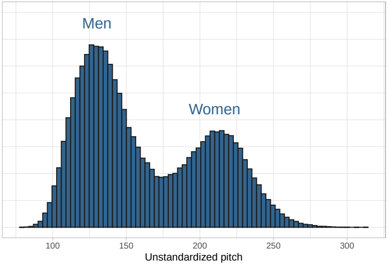
Speaker-standardized $F0$



- ⇝ Bimodal distribution

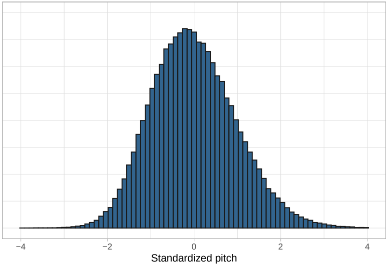
Speaker Heterogeneity

Raw $F0$:



- $\rightsquigarrow$ Bimodal distribution

Speaker-standardized $F0$



$\rightsquigarrow$ Normal distribution

Drawbacks

Drawbacks

Normalization and standardization are helpful but also have their drawbacks.

Drawbacks

Normalization and standardization are helpful but also have their drawbacks.

- Normalization: artificially deflates variance

Drawbacks

Normalization and standardization are helpful but also have their drawbacks.

- Normalization: artificially deflates variance
- Standardization: eliminates any between-speaker variation

Table of Contents

See you next week!

Theme: Audio

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

[1] C. Cochrane, L. Rheault, J.-F. Godbout, T. Whyte, M. W.-C. Wong, and S. Borwein, “The automatic analysis of emotion in political speech based on transcripts,” *Political Communication*, vol. 39, no. 1, pp. 98–121, 2022.

[2] M. Rask and F. Hjorth, “Nonverbal-based measures of elite conflict and polarization,” *Working Paper*, pp. 1–25, 2023.

[3] O. Rittmann, “Legislators’ emotional engagement with women’s issues: Gendered patterns of vocal pitch in the german bundestag,”, 2023.

[4] L. Rheault and S. Borwein, “Multimodal techniques for the study of a ect in political videos,” Working Paper, Tech. Rep., 2019.

[5] D. Knox and C. Lucas, "A dynamic model of speech for the social sciences," *American Political Science Review*, vol. 115, no. 2, pp. 649–666, 2021.