International Methods Colloquium

Modeling Audio Data with Speaker Heterogeneity

Ludovic Rheault and Sophie Borwein
University of Toronto

UNIVERSITY OF TORONTO

Objectives of the Project

- Quantifying sentiment, activation and specific emotional states (anxiety) in political videos, using three modalities.
- In this talk: **TEXT** vs **AUDIO**.
- Methods: deep neural networks; transfer learning.
- Training data: annotated political videos with transcripts.
- Issues: heterogeneity across speakers; coder reliability.

Reference

A preliminary study introducing this project:
Rheault and Borwein (2019).

Looking for political transcripts? Check out lipad.ca.

Previous Work: Audio Data

- Political Science:
  - Dietrich et al. (2019a; 2019b): Pitch as measure of activation.
  - Knox and Lucas (2019): HMM model; skepticism in voice.
  - Neumann (2019): Phonetics; style-shifting.
  - Hwang et al. (2019): Audio and video; political ads.
  - ...
- Engineering/Computer Science:
  - Schuller (2018).
  - Tzirakis et al. (2017).
  - ...

Representing Emotions

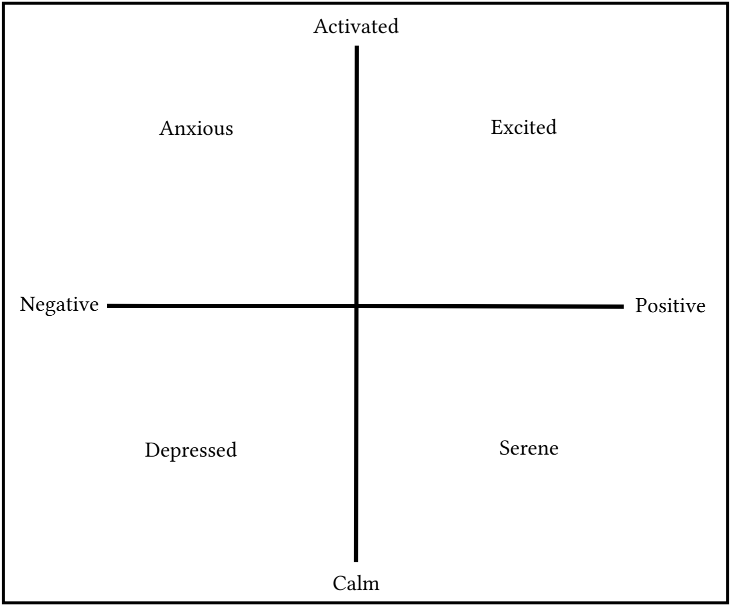
Categorical Approaches

- e.g. Ekman's six basic emotions (fear, anger, sadness, surprise, happiness, disgust).
- Problem: many of them not commonly observed in elites' speeches.
- Fear vs. anxiety.

Dimensional Approaches

- e.g. Russel's circumplex model of affect.
- Sentiment (or valence) and activation (arousal).

Circumplex Model of Affect



Our Data

Sources

3,635 videos: Canadian House of Commons, US Congress, Debates.

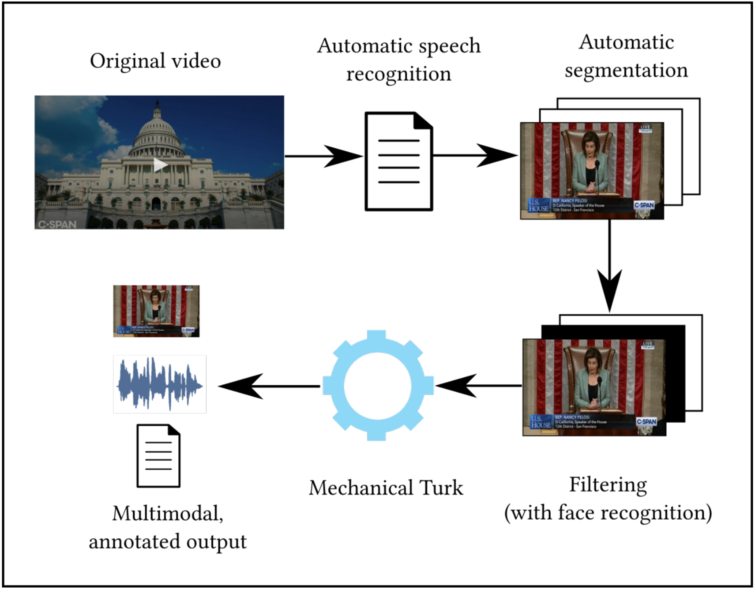
Annotations (Labels)

Three binary annotations (graduate students; MTurk workers).

- Sentiment
- Activation
- Anxiety

Current work: Improving coder reliability.

Pipeline for processing videos



Final Video Collection

Upcoming Steps

- Public release of video dataset with improved coder reliability.
- Lab subjects with biometric measurements as ground truth.

(with Jonathan Rose and Bazen Teferra, UofT Engineering)

Trade-Off

- Crowdsourced annotations of public domain videos:
  - Easier to make data public;
  - Usually low intercoder reliability;
- Human subjects with biometric ground truth:
  - Higher reliability;
  - Very difficult to anonymize audio and video signals.

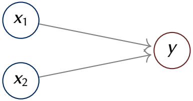
Methods, Concepts and Definitions

Machine learning in one slide

Social science (inference)	Machine learning (prediction)
GLM inverse link function	Activation function
$\mathbb{E}(y) = f(\boldsymbol{x}'\boldsymbol{\beta})$	$\mathbb{E}(y) = f(\boldsymbol{x}'\boldsymbol{\beta})$
Preferred objective function	
Log-likelihood	Cross-entropy
$\log \mathcal{L} = \sum_{i=1}^{n} \log P(y_i \vert \boldsymbol{x}_i, \boldsymbol{\beta})$	$-\log \mathcal{L} = -\sum_{i=1}^{n} \log P(y_i \vert \boldsymbol{x}_i, \boldsymbol{\beta})$
Solving algorithm	
Newton-Raphson	Gradient descent
$\boldsymbol{\beta}_t := \boldsymbol{\beta}_{t-1} - [\boldsymbol{H} \log \mathcal{L}]^{-1} \nabla \log \mathcal{L}$	$\boldsymbol{\beta}_t := \boldsymbol{\beta}_{t-1} - \eta \nabla(-\log \mathcal{L})$
Quantities of interest	
$\widehat{\boldsymbol{\beta}}; Var(\widehat{\boldsymbol{\beta}})$	$\widehat{\boldsymbol{y}}; \sum \mathbf{1}(\widehat{y} = y)/n$

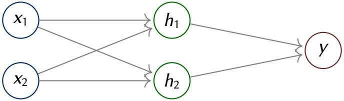
Preliminaries: Neural Networks

Logistic Regression as Neural Network



$$\mathbb{E}(y) = f(\alpha + \boldsymbol{x}'\boldsymbol{\beta})$$

Neural Network with Hidden Layer



$$\mathbb{E}(y) = f^{(2)}(\alpha^{(2)} + \boldsymbol{h}'\boldsymbol{\beta}^{(2)}); \quad h_k = f^{(1)}(\alpha_k^{(1)} + \boldsymbol{x}'\boldsymbol{\beta}_k^{(1)})$$

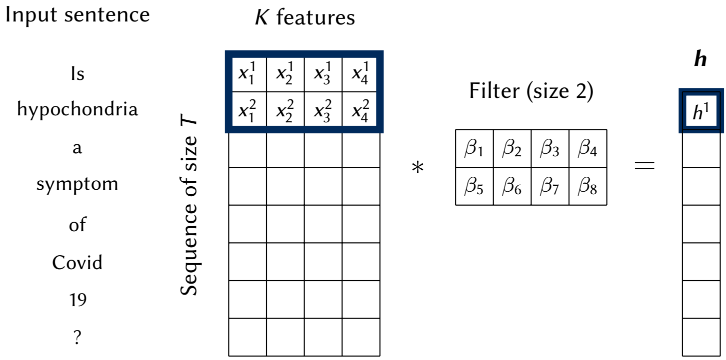
Deep Neural Network (DNN)

Multiple hidden layers: $\mathbb{E}(y) = f^{(4)}(f^{(3)}(f^{(2)}(f^{(1)}(\boldsymbol{x}))))$

Preliminaries: Convolutional Neural Networks

1D ConvNet (or CNN)

$$\boldsymbol{x} * \boldsymbol{\beta} = \boldsymbol{h}$$

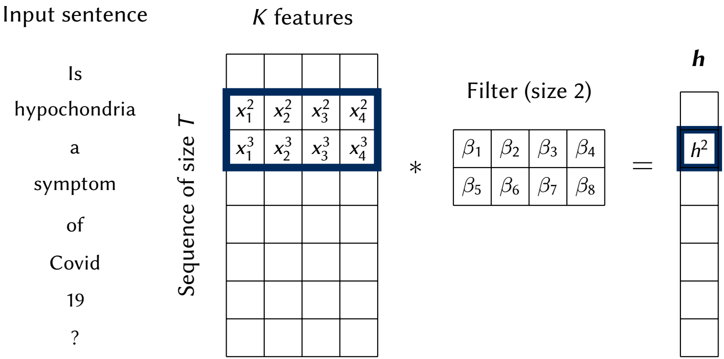


$$h^1 = \beta_1 x_1^1 + \beta_2 x_2^1 + \ldots + \beta_8 x_4^2$$

Preliminaries: Convolutional Neural Networks

1D ConvNet

$$\boldsymbol{x} * \boldsymbol{\beta} = \boldsymbol{h}$$

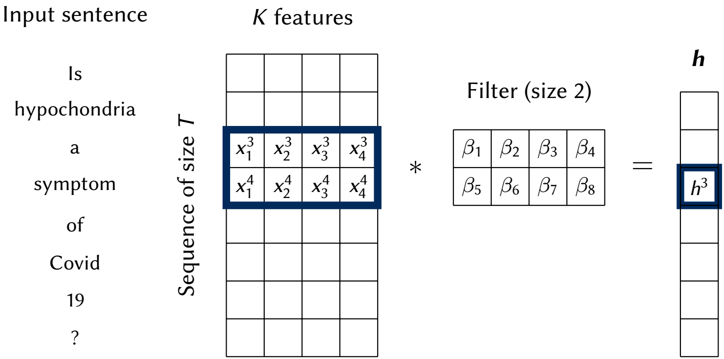


$$h^2 = \beta_1 x_1^2 + \beta_2 x_2^2 + \ldots + \beta_8 x_4^3$$

Preliminaries: Convolutional Neural Networks

1D ConvNet

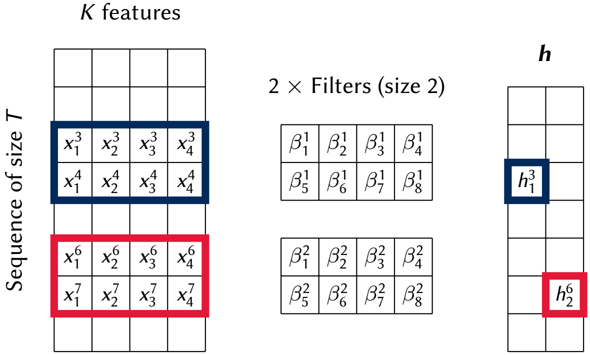
$$\boldsymbol{x} * \boldsymbol{\beta} = \boldsymbol{h}$$



$$h^3 = \beta_1 x_1^3 + \beta_2 x_2^3 + \ldots + \beta_8 x_4^4$$

Preliminaries: Convolutional Neural Networks

There will be $K \times L \times M$ trainable β parameters, where L is the chosen number of filters and M the filter size (or kernel size), plus a filter specific intercept.



Methods: Current Trends

Audio

- Two main approaches: HMMs (e.g. Knox and Lucas 2019) and deep neural networks (Hinton et al. 2012).
- Trends:
  - No features: use raw audio signal as input in ConvNets.
  - Transfer learning (e.g. Audioset, wav2vec, autoencoders).

Text

- Transfer learning everywhere:
  - Previous years: Word embeddings + DNN as default.
  - Now: transfer learning using more sophisticated language models (e.g. ULMFiT, BERT, DistilBERT).

Transfer Learning

- Defining the concept of transfer learning is controversial, but in a nutshell:

The Problem

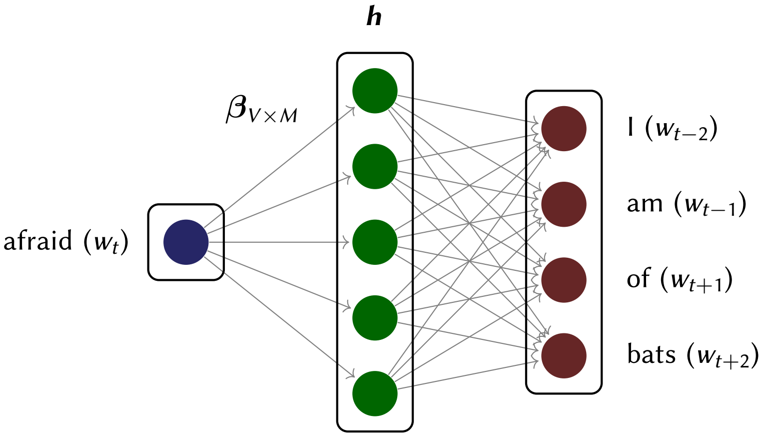
Specific applications usually have limited training data, resulting in poor predictive accuracy.

The Solution

Pre-train a model using a very large dataset, for a different task (e.g. an **autoencoder**). Use the parameters of this larger model as **feature representations** for the target task, or **fine-tune** the model for the target task using local data.

Text as Data

Transfer Learning (Word Embeddings, Skip-Gram)



The learned feature representation matrix $\beta_{V\times M}$ (V size of vocabulary, M size of hidden layer) contains information about semantics not available from a small sample. (Mikolov et al. 2013)

Transfer Learning (Word Embeddings)

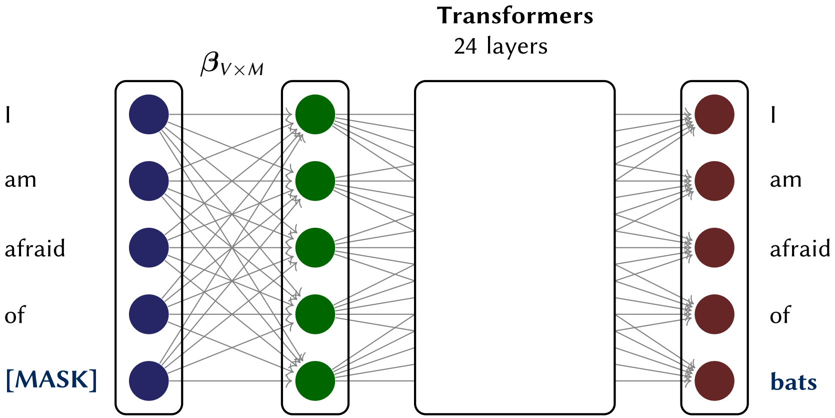
- Map words from a new dataset onto pre-trained embeddings:

$$
\begin{aligned}
\text{“I”} &\rightarrow [-0.51, 1.29, ..., 1.34] \\
\text{“am”} &\rightarrow [0.76, -2.44, ..., -1.06] \\
\text{“afraid”} &\rightarrow [-0.83, -3.09, ..., 0.86] \\
\text{“of”} &\rightarrow [2.25, -2.16, ..., -0.98] \\
\text{“Covid-19”} &\rightarrow [0, 0, ..., 0]
\end{aligned}
$$

- Each document is a matrix: sequence of T words with feature length M.
- Use ConvNets or recurrent neural network (RNNs) to predict target annotation (e.g. sentiment) from the sequences.

$$\text{RNNs:} \quad \boldsymbol{h}^t = f(\alpha + (\boldsymbol{x}^t)'\boldsymbol{\beta} + (\boldsymbol{h}^{t-1})'\boldsymbol{\theta})$$

Transfer Learning (BERT)



Bidirectional Encoder Representations from Transformers (BERT) trained on Wikipedia + BooksCorpus, using two tasks (predicting masked word shown above) (Devlin et al. 2019).

Transfer Learning (BERT)

- Like word embeddings, BERT can provide pretrained embeddings (first hidden layer, or encoder).
- Deep learning architecture (transformers with attention weights) with state-of-the art results on many NLP tasks.
- Two straightforward methods to adapt BERT for a new task:
  - "Freeze" the parameters, add an output layer on top of BERT (e.g. logistic or softmax), and fit with local data.
  - Add the output layer and continue training all parameters with local data (fine-tuning).
- BERT Large model: 24 transformer layers, hidden layer size of 1024.

Results, Text-as-Data Benchmark

BERT model (high-quality annotations only)

Emotion	Accuracy (%)	Modal (%)	PRE (%)	AUROC
Sentiment	88.2	56.2	73.0	0.89
Activation	74.6	71.5	11.0	0.65
Anxiety	62.5	52.0	21.9	0.63

- Text classification works well with sentiment, less so for activation and a specific emotion like anxiety.
- Substantive conclusion the same with other classifiers (e.g. word embeddings + RNN) and annotation quality.
- "The sentiment is in the transcript, but the arousal is not" (Cochrane et al. 2019).

(Accuracy calculated on a held out sample; we use the same with audio models for comparison.)

Audio as Data

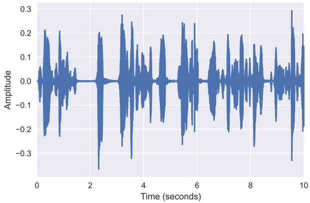
UNIVERSITY OF TORONTO

Audio Data: Raw Signals (Waveform)

A vector of signed integers with a specified bit depth (e.g. 16 bits ranges from -32767 to 32767), usually converted to float:

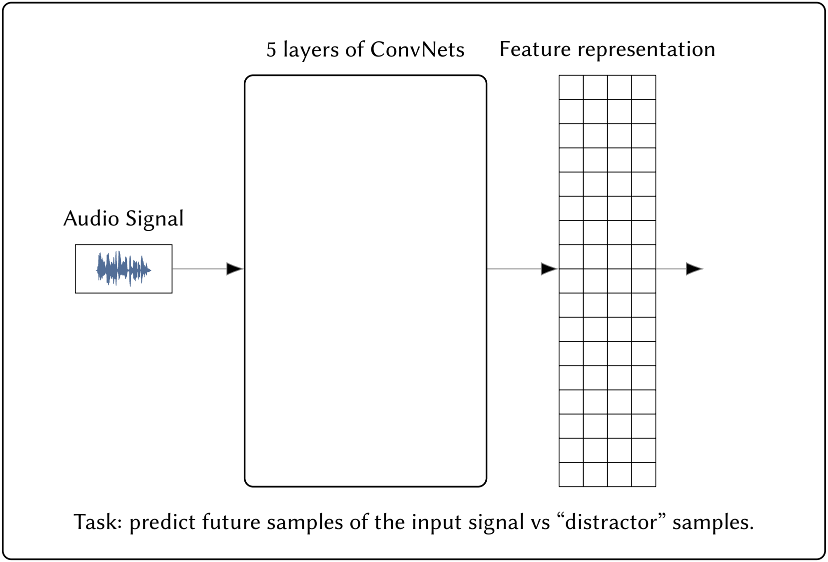
$$[0, 0, 0.15, 0.21, ..., 0]$$

with sampling rate in Hz (integers per second, e.g. 16KHz).



For a great intro on sound, check past IMC presentation from Christopher Lucas.

Transfer Learning (wav2vec)



Schneider et al. (2019)

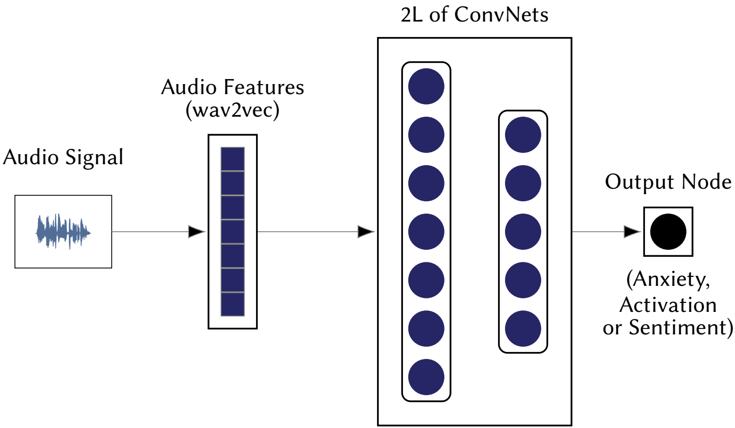
Transfer Learning (wav2vec)

- Use input audio samples to predict likelihood of future sample.
- Trained on 1,000 hours of spoken language (LibriSpeech).
- Two different outputs of wav2vec ConvNet blocks can be used as feature representation of wave inputs (10ms $\times$ 512):

$$
\begin{aligned}
\text{0ms–10ms} &\rightarrow [0.0, 0.03, ..., 0.05] \\
\text{10ms–20ms} &\rightarrow [0.04, 0.02, ..., 0.14] \\
\text{20ms–30ms} &\rightarrow [0.01, 0.0, ..., 0.16] \\
\text{30ms–40ms} &\rightarrow [0.0, 0.11, ..., 0.06] \\
\cdots &\rightarrow \cdots
\end{aligned}
$$

- Intuition similar to word embeddings/BERT.

Audio Data: Model I (Schematic Depiction)



Two 1D ConvNet layers with 16 and 8 filters (size 3), dropout and pooling.

Results Part I, Audio Data (ConvNet)

ConvNets with wav2vec

Emotion	Accuracy (%)	Modal (%)	PRE (%)	AUROC
Activation	80.1	71.5	30.0	0.75
Anxiety	71.7	52.0	41.1	0.72
Sentiment	56.2	56.2	0.0	0.54

- Better than text for activation and anxiety, but not impressive.

Issue: Speaker Heterogeneity

Speaker Heterogeneity

Each voice is unique. As with heterogeneity bias in panel data analysis, we would like to account for a speaker j's attributes:

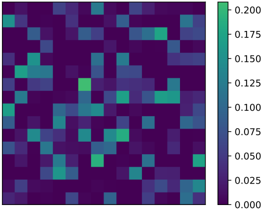
$$\mathbb{E}(y) = f(\alpha_j + \boldsymbol{x}'\boldsymbol{\beta})$$

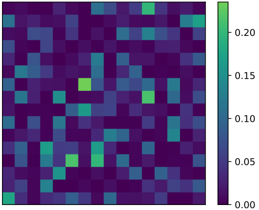
- Deep neural networks can learn to distinguish emotional states from speaker-specific attributes, but this would require a lot of training data.
- Speaker-specific intercepts wouldn't help for new speakers, unobserved during training stage.

Speaker Voice Recognition

Voice Encoder for Speaker Verification

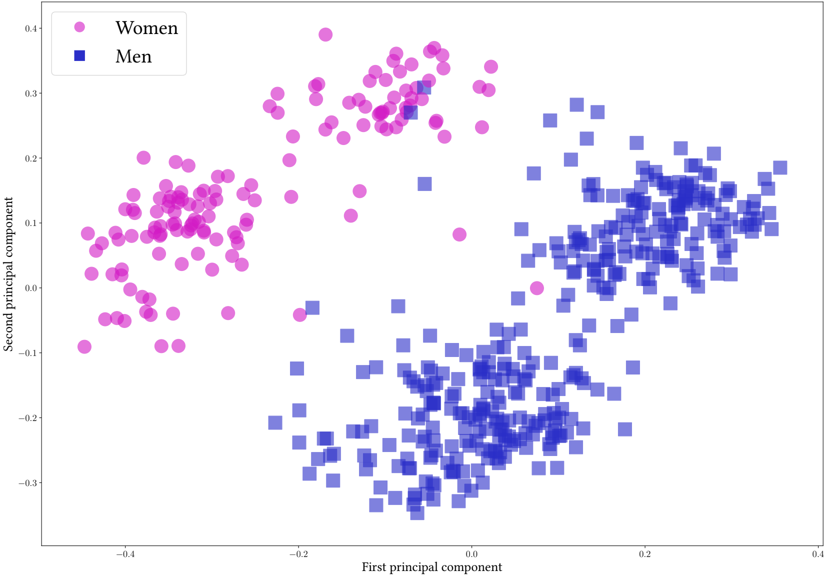
A voice encoder to represent each speaker's individual voice characteristics.



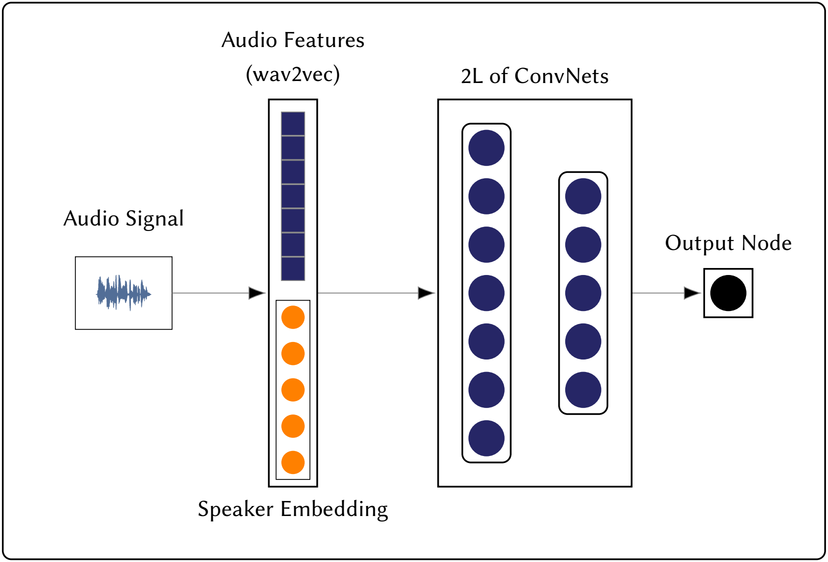


Used for voice synthesis and voice cloning, e.g. Google's Tacotron (Wan et al. 2018, Jia et al. 2019).

Audio Data: Speaker Embeddings (Voice Encoders)



Audio Data: Model II (Schematic Depiction)



Two 1D ConvNet layers with 16 and 8 filters (size 3), dropout and pooling.

Results Part II, Accounting for Heterogeneity

ConvNet with wav2vec AND speaker embeddings

Emotion	Accuracy (%)	Modal (%)	PRE (%)	AUROC
Activation	88.9	71.5	61.0	0.85
Anxiety	78.9	52.0	56.2	0.79
Sentiment	59.7	56.2	8.0	0.58

ConvNet with wav2vec, no speaker embeddings

Emotion	Accuracy (%)	Modal (%)	PRE (%)	AUROC
Activation	80.1	71.5	30.0	0.75
Anxiety	71.7	52.0	41.1	0.72
Sentiment	56.2	56.2	0.0	0.54

Results Part III, Impact of Annotation Quality

ConvNet with wav2vec and speaker embeddings

(High quality annotations only)

Emotion	Accuracy (%)	Modal (%)	PRE (%)	AUROC
Activation	88.9	71.5	61.0	0.85
Anxiety	78.9	52.0	56.2	0.79
Sentiment	59.7	56.2	8.0	0.58

ConvNet with wav2vec and speaker embeddings

(Including low quality annotations)

Emotion	Accuracy (%)	Modal (%)	PRE (%)	AUROC
Activation	76.6	62.6	37.4	0.75
Anxiety	75.7	52.0	49.3	0.76
Sentiment	62.5	50.8	23.7	0.63

Summary: Audio (ConvNet) vs Text (BERT)

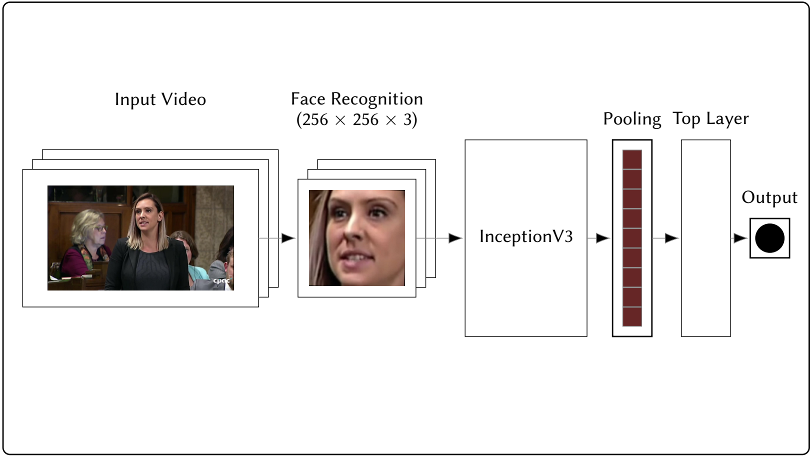
Audio Modality

Emotion	Accuracy (%)	Modal (%)	PRE (%)	AUROC
Activation	**88.9**	71.5	**61.0**	**0.85**
Anxiety	**78.9**	52.0	**56.2**	**0.79**
Sentiment	59.7	56.2	8.0	0.58

Text Modality

Emotion	Accuracy (%)	Modal (%)	PRE (%)	AUROC
Activation	74.6	71.5	11.0	0.65
Anxiety	62.5	52.0	21.9	0.63
Sentiment	**88.2**	56.2	**73.0**	**0.89**

Full Project: Visual Modality



Conclusion

- Text transcripts and audio signals of political speeches offer complementarity:
  - Audio better at capturing aroused/anxious speakers.
  - Transcript better at capturing sentiment (valence).
- Accounting for speaker heterogeneity matters in small samples.
- Quality of human coding a major issue in speech emotion recognition.
- Future step: Completion and public release of video collection.
- Future step: Audio vs Text vs Visual.

Feedback welcome!

References

- Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. “Multimodal Machine Learning: A Survey and Taxonomy.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2): 423–443.
- Cochrane, Christopher, et al. 2019. “The Automated Detection of Emotion in Transcripts of Parliamentary Speech.” APSA 2019.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of NAACL-HLT* 2019.
- Dietrich, Bryce J, Ryan D Enos, and Maya Sen. 2019a. “Emotional Arousal Predicts Voting on the U.S. Supreme Court.” *Political Analysis* 27(2): 237–243.
- Dietrich, Bryce J, Matthew Hayes, and Diana Z O’Brien. 2019b. “Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech.” *American Political Science Review* 113(4): 941–962.
- El Ayadi, Moataz, Mohamed S Kamel, and Fakhri Karray. 2011. “Survey on Speech Emotion Recognition” *Pattern Recognition* 44(3): 572–587.
- Hinton, Geoffrey, et al. 2012. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups.” *IEEE Signal Processing Magazine* 29(6): 82–97.
- Hwang, June, Kosuke Imai, and Alex Tarr. 2019. “Automated Coding of Political Campaign Advertisement Videos: An Empirical Validation Study.” PolMeth XXXVI.

References (Continued)

- Jia, Ye, et al. 2018. "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis." In Advances in Neural Information Processing Systems. pp. 4480–4490.
- Knox, Dean and Christopher Lucas. 2019. "A Dynamic Model of Speech for the Social Sciences." Available at SSRN: https://ssrn.com/abstract=3490753.
- Mikolov, Tomas, et al. 2013. "Efficient Estimation of Word Representations in Vector Space." ICLR 2013.
- Neumann, Markus. 2019. "Hooked With Phonetics: The Strategic Use of Style-Shifting in Political Rhetoric." PolMeth XXXVI.
- Russel, James. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* 39(6): 1161-1178.
- Schneider, Steffen, et al. 2019. "wav2vec: Unsupervised Pre-Training for Speech Recognition." Available at arXiv:1904.05862
- Schuller, Björn W. 2018. "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends." *Communications of the ACM* 61(5): 90–99.
- Tzirakis, Panagiotis, et al. 2017. "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks." *IEEE Journal of Selected Topics in Signal Processing* 11(8): 1301–1309.
- Wan, Li, et al. 2018. "Generalized End-to-End Loss for Speaker Verification." ICASSP 2018.